ELSEVIER

Topology and its Applications 126 (2002) 83–118

TOPOLOGY AND ITS APPLICATIONS

www.elsevier.com/locate/topol

Generalized billiard paths and Morse theory for manifolds with corners

David G.C. Handron

Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA (412)268-2545, USA

Received 30 March 2001; received in revised form 3 December 2001

Abstract

A billiard path on a manifold M embedded in Euclidean space is a series of line segments connecting reflection points on M. In a generalized billiard path we also allow the path to pass through M. The two segments at a 'reflection' point either form a straight angle, or an angle whose bisector is normal to M. Our goal is to estimate the number of generalized billiard paths connecting fixed points with a given number of reflections.

We begin by broadening our point of view and allowing line segments that connect any sequence of points on M. Since this sequence is determined by its 'reflection' points, the length of such a sequence with k reflections may be thought of as a function on M^k. Generalized billiard paths correspond to critical points of this length function. The length function is not smooth on M^k, having singularities along some of its diagonals. Following the procedure of Fulton and MacPherson we may blow up M^k to obtain a compact manifold with corners to which the length function extends smoothly.

We develop a version of Morse theory for manifolds with corners and use it to study this length function. There are already versions of Morse theory that may be used in this case, but ours is a generalization of the work of Braess, retaining both a global 'gradient' flow and the intrinsic stratification of a manifold with corners.

We find that the number of generalized billiard paths with k reflections connecting two points in $\mathbb{R}^N$ can be estimated in terms of the homology of the manifold M. In part, we show the number of these paths is at least

$$\sum_{j=0}^{n-1} \sum_{i_1+\cdots+i_k=j} b_{i_1}(M)\cdots b_{i_k}(M)$$

MSC: 57R70; 57R25

Keywords: Morse theory; Stratified spaces; Billiard path problem

E-mail address: handron@andrew.cmu.edu (D.G.C. Handron).

0166-8641/02/$ – see front matter
PII: S0166-8641(02)00036-6

1. Introduction

1.1. A motivational example

Imagine that we have a glass surface which has been half-silvered. Any time a beam of light struck the surface, half of the light would reflect off the surface and half would pass through.

One of the questions this article seeks to answer is: given two points in the vicinity of such a model, how many paths may a beam of light travel connecting one point to the other, with a given number of reflections?

Since the beam of light travels in a straight line between reflections, such a path can be described by listing the sequence of reflection points. Moreover, all of these paths have the property that wherever a reflection occurs, the angle of incidence is equal to the angle of reflection. This can be stated equivalently by saying that the bisector of the angle is perpendicular to the surface.

1.2. The general problem

This same question can be posed in more general terms. Let $M \hookrightarrow \mathbb{R}^N$ be a smooth n-manifold embedded in Euclidean space of dimension N. We can choose points $p, q \in \mathbb{R}^N$ and consider ordered sequences of points $\alpha_1, \ldots, \alpha_k \in M$.

Definition (*Definition* 9). A sequence $P = \{\alpha_1, \ldots, \alpha_k\}$ connecting $p = \alpha_0$ to $q = \alpha_{k+1}$ is a *generalized billiard path* with k-reflections if for each i one of the following is true:

(1) The bisector of $\measuredangle\alpha_{i-1}\alpha_i\alpha_{i+1}$ is normal to $T_{\alpha_i}M$.
(2) $\measuredangle\alpha_{i-1}\alpha_i\alpha_{i+1}$ is a straight angle.

Note that this definition allows the line segment $\overline{\alpha_i\alpha_{i+1}}$ to intersect the manifold. If M happens to be a convex hypersurface, however, the definition reduces to the usual notion of a billiard path. This situation is addressed in [3]. The task at hand now may be thought of as counting generalized billiard paths.

The space of all sequences can be thought of as the product $M^k = M \times \cdots \times M$. We can define the length of a sequence $P = \{\alpha_1, \ldots, \alpha_k\}$ to be the sum of the Euclidean distances between consecutive reflections:

$$L_k(P) = \sum_{i=0}^{k} d_{\mathrm{Euc}}(\alpha_i, \alpha_{i+1}).$$

This function will be central to our arguments, because of

Lemma (Lemma 10). *A sequence* $P = \{\alpha_1, \ldots, \alpha_k\}$ *with* $\alpha_i \neq \alpha_{i+1}$ *for* $0 \leqslant i \leqslant k$ *satisfies*

$$\nabla L_k(P) = 0$$

if and only if it is a generalized billiard path.

Unfortunately, the function L_k has a serious drawback. Wherever consecutive reflections coincide, L_k has a singularity that looks like $|x - y|$.

In Section 3.3 we describe how to 'blow up' M^k. The blow up we use was developed by Fulton and MacPherson [4], and allows us to remove from M^k the diagonals $\{\alpha_i = \alpha_{i+1}\}$ that are causing difficulty and replace them with something that is easier for us to deal with. The result is a manifold with corners, X_k.

We will use the versions of Morse theory developed in Sections 2 and 4 to study the critical points of L_k on X_k. In doing so, we define a modified gradient flow. An essential critical point is defined to be a stationary point of that flow. In Section 2.7, we prove the Morse theorems in this setting:

Theorem (Theorem 7). *Let $f : M \to \mathbb{R}$ be a Morse function on a manifold with corners M. If $a < b$ and $f^{-1}([a, b])$ contains no essential critical points, then M_a is a deformation retract of M_b, so the inclusion map $M_a \hookrightarrow M_b$ is a homotopy equivalence.*

Theorem (Theorem 8). *Let $f : M \to \mathbb{R}$ be a Morse function on a manifold with corners M. Let p be an essential critical point with index λ. Set $f(p) = c$. Suppose that, for some $\varepsilon > 0$, $f^{-1}([c - \varepsilon, c + \varepsilon])$ contains no essential critical points other than p. Then $M_{c+\varepsilon}$ is homotopy equivalent to $M_{c-\varepsilon}$ with a λ-cell attached.*

These theorems imply the Morse Inequalities, which we will use to deduce lower bounds for the number of generalized billiard paths.

In Section 3.4, we show that for a smooth embedding $M \hookrightarrow \mathbb{R}^N$, most choices of endpoints p and q result in a length function that satisfies the definition of a Morse function (Definitions 4 and 6):

Lemma (Lemma 19). *For any embedded manifold $M \hookrightarrow \mathbb{R}^N$, points $p, q \in \mathbb{R}^N$ and $\varepsilon > 0$, there are points $p' \in B_\varepsilon(p)$ and $q' \in B_\varepsilon(q)$ such that $-L_k^{(p',q')}$ is a Morse function.*

Section 3.5 applies the results of Section 2 to the function $-L_k$ on X_k. The result is given by

Theorem (Theorem 21). *The number of generalized billiard paths with k reflections is at least*

$$\sum_{i=0}^{kn} b_i(X_k),$$

where $b_i(X_k)$ denotes the ith Betti number of X_k.

In Section 4 we show that a stratified space structure can be imposed on M^k and that $-L_k$ is a then Morse function on M^k. A comparison of the critical points of $L_k : X_k \to \mathbb{R}$ with those of $-L_k : M^k \to \mathbb{R}$ allows us, in Section 5, to conclude

Theorem (Theorem 30). *The number of generalized billiard paths connecting p to q with k reflection is at least*

$$\sum_{j=0}^{n-1} \sum_{i_1+\cdots+i_k=j} b_{i_1}(M)\cdots b_{i_k}(M).$$

1.3. A brief history of Morse theory

The foundations of Morse theory were laid in the 1920s by Marston Morse [8]. His original work relates information about the critical points of a smooth function on a smooth manifold to information about the topology of the manifold. This relationship was presented at that time as a collection of inequalities, known as the Morse Inequalities.

By the late 1940s, the gradient flow of the function was coming into the picture more forcefully. Once a Riemannian metric has been chosen, each point in the manifold lies in exactly one gradient flow line, and each such flow line begins and ends at a critical point. Thom noticed that by bundling together all the flow lines having the same initial point, the manifold can be decomposed into a collection of 'descending cells'—one for each critical point [9]. The dimension of the cell associated to a critical point is equal to the index of that critical point.

In 1959, Smale showed that if the 'ascending cell' of each critical point intersect transversely with each descending cell it meets, then the descending cells form a CW-complex.

Morse theory has been generalized to deal with a large number of situations which are not addressed by the classical theory. The direction with the most direct relevance to this work, though, is treating functions on spaces other than smooth manifolds. Braess presented a version for manifolds with boundary in 1974 [2]. The most remarkable achievement in this area, though, is Goresky and MacPherson's stratified Morse theory. This version applies to a class called Whitney stratified spaces, which include manifolds with boundary and manifolds with corners [5]. Some of the proofs of Goresky and MacPherson's theorems have recently been simplified by Hamm in [6]. Vakhremeev has also proven the Morse theorems for the case of Manifolds with corners [10].

In Vakhrameev's work and the stratified Morse theory of Goresky and MacPherson, however, the gradient flow does not appear as prominently as it does in other versions. Indeed, the functions on these spaces may not even allow a gradient flow to be defined globally. My intention in the first part of this work is to produce a Morse theory for manifolds with corners, a type of stratified space, that retains the point of view developed by Thom and Smale. A more thorough history of Morse theory may be found in [1].

2. Morse theory for manifolds with corners

2.1. Manifolds with corners

Let $\{\boldsymbol{e}_1,\ldots,\boldsymbol{e}_j\}$ denote the standard basis vectors in $\mathbb{R}^n$. Define $\mathbb{H}^n_j$ to be the set

$$\mathbb{H}^n_j = \left\{\boldsymbol{w}\in\mathbb{R}^n\colon\ \boldsymbol{w}\cdot\boldsymbol{e}_i \geqslant 0 \text{ for all } 1\leqslant i\leqslant j\right\},$$

where $\cdot$ denotes the standard inner product on $\mathbb{R}^n$.

Definition 1. An *n-dimensional manifold with corners, M*, is a topological space together with an atlas, $\mathcal{A}$, of charts $\boldsymbol{x}_a : U_a \to \mathbb{H}^n_{j_a}$ such that $\bigcup_{a\in\mathcal{A}} U_a = M$.

If $p \in M$, we will say a *coordinate chart at p* is a chart $\boldsymbol{x}_p \in \mathcal{A}$ such that $\boldsymbol{x}_p(p) = 0 \in \mathbb{H}^n_j$. In this case, the number j is uniquely determined by the point p. Thus we can write $j = j(p)$.

The tangent space of a manifold with corners can be defined as equivalence classes of

$$C_p(M) = \big\{(\boldsymbol{x}, \boldsymbol{v})\colon\ \boldsymbol{x} \text{ is a coordinate chart at } p \in M \text{ and } \boldsymbol{v} \in \mathbb{R}^n\big\},$$

where $(\boldsymbol{x}, \boldsymbol{v}) \sim (\boldsymbol{y}, \boldsymbol{w})$ if $D(\boldsymbol{x} \circ \boldsymbol{y}^{-1})(\boldsymbol{w}) = \boldsymbol{v}$. If $p \in \partial M$ then some of the vectors in T_pM point away from the manifold with corners.

Definition 2. A tangent vector in T_pM points *outward* if some representative $(\boldsymbol{x}, \boldsymbol{v})$ has $\boldsymbol{v} \notin \mathbb{H}^n_{j(p)}$. A tangent vector is T_pM points *inward* (or *into M*) if some representative $(\boldsymbol{x}, \boldsymbol{v})$ has $\boldsymbol{v} \in \mathbb{H}^n_{j(p)}$.

Note that the definition of an inward pointing vector includes those vectors which are tangent to the boundary of M. These terms are well defined, since for any two coordinate charts at p, the transition functions preserve $\mathbb{H}^n_{j(p)}$.

2.2. Stratified spaces

There are a number of different notions of what constitutes a stratified space. We will not be using any results pertaining any particular theory of stratified spaces, but we will find the language to be convenient. Consequently, we will use a fairly general definition of 'stratified space'.

Definition 3. A *stratified space* consists of a topological space X, a partially ordered set $\mathcal{S}$ and a collection $\{H_i\}_{i\in\mathcal{S}}$ of subspaces of X satisfying

(1) Each H_i is a manifold.
(2) $X = \bigcup_{i\in\mathcal{S}} H_i$.
(3) $H_i \cap \overline{H_j} \neq \emptyset \Leftrightarrow H_i \subseteq \overline{H_j} \Leftrightarrow i \preceq j$. In this case we also write $H_i \preceq H_j$.

Each of the manifolds H_i is a *stratum* of X.

For us, the most important example of a stratified space is a manifold with corners. For a manifold with corners M, let $\mathcal{E}_j(M) = \{p \in M\colon\ j(p) = j\}$. It is not difficult to see that $\mathcal{E}_j(M)$ is a manifold of dimension $n - j$. We may think of each connected component of $\mathcal{E}_j(M)$ as a stratum with dimension $n - j$.

2.3. Morse functions on manifolds with corners

Let $f : M \to \mathbb{R}$ be a smooth function. If H is a stratum of M, and $p \in H$, we say p is a *critical point* of f whenever p is a critical point of $f|_H$.

If $p \in H$ is in the closure of another stratum, K, we can define the generalized tangent space

$$T_pK = \left\{ \boldsymbol{w} \in T_pM\colon\ \boldsymbol{w} = \lim_{i\to\infty} \boldsymbol{v}_i \in T_{q_i}K \text{ for some sequence } \{q_i\} \to p \right\}.$$

We may also write this as $T_pK = \lim_{q\to p} T_qK$.

Definition 4. For a manifold with corners M, we say a smooth function $f : M \to \mathbb{R}$ is a Morse function if it has the following properties:

(1) If H is a stratum of M, and $p \in H$ is a critical point of $f|_H : H \to \mathbb{R}$, then either
   (a) p is a non-degenerate critical point of $f|_H : H \to \mathbb{R}$, i.e., the Hessian has non-zero determinant, or
   (b) the vector $-\nabla f(p)$ points into M.
(2) If $p \in H$ is a critical point, then for any stratum $K \neq H$ with p in the closure of K, df_p is not identically zero on T_pK.

Notice that this definition involves only the first and second derivatives of f. In fact a Morse function need only be C^2 in a neighborhood of each critical point in the interior of M and each critical point such that $-\nabla f(p)$ points outward. It need only be C^1 elsewhere.

2.4. Modifying the gradient vector field

In classical Morse theory, a Morse function $f : M \to \mathbb{R}$ is studied by choosing a Riemannian metric on M and examining the flow induced by the vector field $-\nabla f$. When we allow the manifold M to have corners (or even just a boundary) a difficulty arises. If $-\nabla f$ points outward from any point in ∂M, the vector field cannot produce a flow that carries M to M. As a result, we must modify the gradient vector field to produce a new vector field that does induce such a flow. As we do this, we must keep in mind the two properties this flow must have. First, it must be continuous, and second, the value of the function f must decrease along the flow lines.

The point of view we will take is that we want to follow the gradient vector field as closely as possible. What we must do is project the vector $-\nabla f(p)$ onto the maximal stratum such that the resulting vector does not point outward from M.

At first sight, is makes no sense to talk about $-\nabla(f|_H)(p)$, when $p \notin H$. When $p \in \overline{H}$, however, this can be reasonably defined. The approach requires us to remember that $f|U_p$ can be thought of as $f \circ \boldsymbol{x}_p^{-1} : \boldsymbol{x}_p(U_p) \to \mathbb{R}$ and extended to a function $\tilde{f} : \mathbb{R}^n \to \mathbb{R}$. Thus we can extend the stratum H to a manifold $\widetilde{H} \supset H$ that contains p. Then we can define

$$-\nabla(f|_H)(p) = -\nabla(f|_{\widetilde{H}})(p).$$

Since $-\nabla\tilde{f}$ is continuous on $\widetilde{H}$, this procedure provides a continuous extension of $-\nabla(f|_H)$ to $\widetilde{H}$.

How do we know there must be a maximal stratum K such that $-\nabla(f|_K)(p)$ does not point outward from M? Suppose we have two strata, H_1 and H_2, such that $H_1 \neq H_2$ and $-\nabla(f|_{H_i})(p)$ points inward toward H_i for $i = 1, 2$. Choose a standard coordinate chart $\boldsymbol{x}: U_p \to \mathbb{H}^n_{j(p)}$ at p. Then the coordinate x_i will be non-negative whenever $i > n - j(p)$. Let $A_i = \{\ell \in \{1, \ldots, n\}\colon\ \boldsymbol{e}_\ell \text{ points into } H_i\}$. Then $\boldsymbol{x}(H_i \cap U_p)$ is an open subset of $\operatorname{span}\{\boldsymbol{e}_\ell\}_{\ell \in A_i}$. Since $-\nabla(f|_{H_i})(p)$ points into M, we must have $\boldsymbol{e}_\ell[f] \geqslant 0$ when $\ell \geqslant n - j(p)$ and $\ell \in A_i$. (Recall that according to Definition 2 vectors tangent to a stratum in ∂M are considered inward pointing.)

Let $B = A_1 \cup A_2$. Then $\operatorname{span}\{\boldsymbol{e}_\ell\}_{\ell \in B} \cap U_p = K \cap U_p$ for some stratum K, and $-\nabla(f|_K)(p)$ points into M.

Then either $\dim(K) > \dim(H_i)$ for $i = 1, 2$, or one of the two strata is contained in the boundary of the other. Consequently for each point $p \in M$, there is a unique maximal stratum K_p such that $-\nabla(f|_{K_p})(p)$ points into M. This allows us to make

Definition 5. At each point $p \in M$, let K_p be the unique maximal stratum such that $p \in \overline{K_p}$ and $-\nabla(f|_{K_p})(p)$ does not point outward from M. Set $G(p) = -\nabla(f|_{K_p})(p)$.

Then G is a well defined vector field on M. From the above construction we see that the directional derivative $G(p)[f] \leqslant 0$, so the value of f will decrease along the flow lines of any flow induced by G. What we must show is that such a flow exists and is continuous.

2.5. The modified gradient vector field induces a continuous flow

First we will show that even though the modified gradient vector field G is not continuous, it does induce a flow. The G-flow will follow the $-\nabla f$-flow until it hits a stratum H in the boundary. It then follows the $-\nabla(f|_H)$-flow until it either hits a lower-dimensional stratum, or flows back into the interior. To ensure uniqueness, we must impose another condition on our Morse functions.

Definition 6. We say that a Morse function $f: M \to \mathbb{R}$ satisfies Property (3) if for any standard coordinate chart $\boldsymbol{x}$, whenever $-\nabla f(p)$ is tangent to a stratum $H \subseteq \partial M$ with $\boldsymbol{e}_i \perp H$ and $\frac{\partial f}{\partial x_i}(p) = 0$, the directional derivative of $\frac{\partial f}{\partial x_i}$ in the direction $-\nabla f(p)$ is not zero, i.e.,

$$\big(-\nabla f(p)\big)\left[\frac{\partial f}{\partial x_i}\right] = \sum_{j=1}^{n} -\frac{\partial f}{\partial x_j}\frac{\partial^2 f}{\partial x_j \partial x_i} \neq 0.$$

This is equivalent to the statement $-\nabla f(p)$ is not tangent to the set $\{q \in M\colon\ \frac{\partial^2 f}{\partial x_i^2} = 0\}$.

Lemma 1. *If $f: M \to \mathbb{R}$ is a function satisfying Property* (3)*, then the modified gradient vector field G induces a flow $\varphi: M \times [0, \infty) \to M$ satisfying*

$$\varphi(\cdot, 0) = \textit{identity}$$
$$\frac{\partial}{\partial t}\varphi(p, t)|_{t=t_0} = G\big(\varphi(p, t_0)\big).$$

Proof. Recall that in Definition 5 we selected at each point p in a stratum $S \subset M$ a stratum K_p and defined $G(p) = -\nabla(f|_{K_p})(p)$. Since $f|_S$ is a smooth function, the vector field $-\nabla f|_S$ induces a continuous flow φ_S on S. We can use this flow to define a stratum $H_p = \lim_{t\to 0} K_{\varphi_S(t,p)}$. In general we will find that $H_p = K_p$, but if $\frac{\partial f}{\partial x_i}(p) = 0$ for some i, this may not be the case.

We can solve the initial value problem $\sigma_p'(t) = -\nabla(f|_{H_p})(\sigma_p(t))$, $\sigma_p(0) = p$. The solution σ_p lies in H_p. Let t_1 be given by

$$t_1 = \sup\bigl\{t \in \mathbb{R}_+ \colon \sigma_p'(t) = G\bigl(\sigma_p(t)\bigr)\bigr\}.$$

Then for $t \leqslant t_1$, we set $\varphi(t, p) = \sigma_p(t)$. For $t > t_1$, we must repeat this procedure starting from $\sigma_p(t_1)$, and flowing for time $t - t_1$. □

This shows that the vector field G induces some flow on M. Our goal now is to show that this flow is continuous.

Lemma 2. *If $f : M \to \mathbb{R}$ is a function satisfying Property* (3)*, then the flow induced by the modified gradient G is continuous.*

Near a point $p \in M$, there is a coordinate system $\boldsymbol{x} : U_p \to V_p \subseteq \mathbb{H}^n_j$ such that $\boldsymbol{x}(q) = (x_1(q), \ldots, x_n(q))$. As usual, the coordinates are chosen so that $x_i \in (-\infty, \infty)$ for $i \leqslant n - j$, $x_i \in [0, \infty)$ for $i > n - j$, and $\boldsymbol{x}(p) = (0, \ldots, 0)$.

Choose an $R > 0$ such that $B_R(p) \subset U_p$. Let $\mu = \sup_{q\in M} \|G(q)\|$. Then we can choose r and τ_0 such that $\mu\tau_0 < R - r$. Then $\varphi(t, B_r(p)) \subset B_R(p)$ for every $t < \tau_0$. Consequently, it is sufficient to view the situation in terms of the coordinate system $\boldsymbol{x}$.

Define a projection $\pi : \mathbb{R}^n \to \mathbb{H}^n_j$ by $\pi(v) = (\pi_1(v_1), \ldots, \pi_n(v_n))$, where

$$\pi_i(v_i) = \begin{cases} v_i, & \text{if } 1 \leqslant i \leqslant n - j \text{ or } v_i \geqslant 0 \\ 0, & \text{else.} \end{cases}$$

Note that π is a continuous map and $d_{\mathrm{Euc}}(\pi(x), y) \leqslant d_{\mathrm{Euc}}(x, y)$.

We have a vector field $\widetilde{G} = \boldsymbol{x}_*(G)$ on V_p. There is also another vector field $F = \boldsymbol{x}_*(-\nabla f)$. Extend F to all of $\pi^{-1}(V_p)$ by setting $\widetilde{F}$ to be "constant" (i.e., parallel in the Euclidean metric) along each preimage, $\pi^{-1}(q)$ for $q \in V_p$. The extended vector field $\widetilde{F}$ is Lipschitz, and so induces a continuous flow, denoted by ψ.

Next, we want to define maps $T_i : \pi_{-1}(V) \to \{0, 1\}$ for $i = n - j + 1, \ldots, n$. The idea is that T_i will be zero where the flow φ stays within a stratum where $x_i = 0$. T_i changes to 1 when the flow enters a higher-dimensional stratum where $x_i > 0$.

$$T_i(q) = \begin{cases} 0, & \text{if } y_i(q) \leqslant 0 \text{ and } \bigl\langle -\widetilde{G}(q), \frac{\partial}{\partial y_i} \bigr\rangle \leqslant 0 \\ 1, & \text{else.} \end{cases}$$

Definition 7. Say that $\psi(\cdot, q)$ has an *uptick* at time t if for some $i \in \{n - j + 1, \ldots, n\}$,

$$\lim_{s\to t^-} T_i\bigl(\psi(s, q)\bigr) < \lim_{s\to t^+} T_i\bigl(\psi(s, q)\bigr).$$

Lemma 3.

(1) *If $q \in V_p$ and $\psi(\cdot, q)$ has no upticks in $(0, \tau)$, then*

$$\varphi(\tau, q) = \pi \circ \psi(\tau, q).$$

(2) *If $\psi(\cdot, q)$ has an uptick in $(0, \tau)$, then*

$$d_{\mathrm{Euc}}\big(\varphi(\tau, q), \pi \circ \psi(\tau, q)\big) < 2\mu\tau.$$

Proof. Since μ is the maximal speed for both flows, the farthest they can diverge in time τ is $2\mu\tau$. That proves the second part of the Lemma.

Now suppose that $\psi(\cdot, q)$ has no upticks in $(0, \tau)$. Let $\varphi(t, q) = (x_1(t), \ldots, x_n(t))$ and $\psi(t, q) = (y_1(t), \ldots, y_n(t))$. It is sufficient to consider the case where the flow φ remains in a single stratum, say $H \subset \mathcal{E}_j(M)$.

The x_i's satisfy the system of differential equations

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = g_i(x_1, \ldots, x_n) = g_i(x_1, \ldots, x_j, 0, \ldots, 0).$$

We are able to set $x_{j+1} = \cdots = x_n = 0$, since this flow remains in H.

The y_i's, on the other hand, are determined by the system

$$\frac{\mathrm{d}y_i}{\mathrm{d}t} = f_i(y_1, \ldots, y_n) = f_i(y_1, \ldots, y_j, 0, \ldots, 0).$$

Here, we replace $y_{j+1}, \ldots, y_n$ with 0, because the f_i are constant on

$$\pi^{-1}(y_1, \ldots, y_j, 0, \ldots, 0).$$

Moreover, for $1 \leqslant i \leqslant j$,

$$f_i(y_1, \ldots, y_j, 0, \ldots, 0) = g_i(y_1, \ldots, y_j, 0, \ldots, 0).$$

Consequently for $1 \leqslant i \leqslant j$, $y_i(t) = x_i(t)$.

For $i > j$, $\pi_i(y_i) = 0$. Since $x_{j+1} = \cdots = x_n = 0$, it follows that

$$\pi\big(y_1(t), \ldots, y_n(t)\big) = \big(x_1(t), \ldots, x_n(t)\big).$$

It follows then, that $\pi \circ \psi(\tau, q) = \varphi(\tau, q)$. □

Lemma 4. *For $q \in U$ and a suitably chosen τ there is a finite upper limit, N, to the number of upticks along $\psi(\cdot, q) : [0, \tau] \to M$.*

Proof. The set $f^{-1}((-\infty, f(q))$ is compact and contains the image of the curve $\psi(\cdot, q) : [0, \tau] \to M$. Suppose that the set $\{p_i\}$ of points where $\psi(\cdot, q) : [0, \tau] \to M$ has an uptick is infinite. Then some subsequence of $\{p_i\}$ has a limit point p_0.

Property (3), however, ensures that there is a neighborhood of p_0 that contains no other upticks, deriving a contradiction. This shows that each such curve has a finite number N_q of upticks. We need to show that there is a finite upper bound for $\{N_q \colon q \in M\}$.

Suppose there is no such upper bound. Then choose a sequence $\{q_i\}$ so that $N_{q_i} > i$. Finally, choose a pair $\{a_i = \psi(\tau_{a_i}, q_i), b_i = \psi(\tau_{b_i}, q_i)\}$ so that $|\tau_{a_i} - \tau_{b_i}|$ is minimized along the curve $\psi(\cdot, q) : [0, \tau] \to M$. Then $|\tau_{a_i} - \tau_{b_i}| \to 0$ as $i \to \infty$.

Using compactness again, we can find a subsequence of pairs $\{a_i, b_i\}$ so that $a_i \to p_0$ and $b_i \to p_0$ as $i \to \infty$. It follows that in a standard coordinate system, for some $n - j(p_0) < i \leqslant n$,

$$\frac{\partial f}{\partial x_i}(p_0) = 0$$

since p_0 is a limit of upticks, and $\frac{\partial f}{\partial x_i}$ is continuous. Moreover, because p_0 is the limit of two consecutive upticks, the directional derivative of $\frac{\partial f}{\partial x_i}$ in the $-\nabla f(p_0)$ direction satisfies

$$\big(-\nabla f(p_0)\big)\left[\frac{\partial f}{\partial x_i}\right] = 0.$$

But this contradicts the fact that f satisfies Property (3). □

Proof of Lemma 2. We can define a family of maps $\psi_k : [0, \tau_0) \times M \to M$ by

$$\psi_k(\tau, q) = \left[\pi \circ \psi\left(\frac{\tau}{k}, \cdot\right)\right]^k (q).$$

Fig. 1 shows an example where $k = 3$. If $\psi(\tau, q)$ has upticks at times $t_1, \ldots, t_n$, we may write this as

$$\begin{aligned}
\psi_k(\tau, q) = {} & \left[\pi \circ \psi\left(\frac{\tau}{k}, \cdot\right)\right]^{N_n} \circ \left[\pi \circ \psi\left(\frac{\tau}{k}, \cdot\right)\right] \circ \left[\pi \circ \psi\left(\frac{\tau}{k}, \cdot\right)\right]^{N_{n-1}} \circ \cdots \\
& \circ \left[\pi \circ \psi\left(\frac{\tau}{k}, \cdot\right)\right]^{N_1} \circ \left[\pi \circ \psi\left(\frac{\tau}{k}, \cdot\right)\right] \circ \left[\pi \circ \psi\left(\frac{\tau}{k}, \cdot\right)\right]^{N_0} (q) \\
= {} & \left[\pi \circ \psi\left(N_n \frac{\tau}{k}, \cdot\right)\right] \circ \left[\pi \circ \psi\left(\frac{\tau}{k}, \cdot\right)\right] \circ \left[\pi \circ \psi\left(N_{n-1} \frac{\tau}{k}, \cdot\right)\right] \circ \cdots \\
& \circ \left[\pi \circ \psi\left(N_1 \frac{\tau}{k}, \cdot\right)\right] \circ \left[\pi \circ \psi\left(\frac{\tau}{k}, \cdot\right)\right] \circ \left[\pi \circ \psi\left(N_0 \frac{\tau}{k}, \cdot\right)\right](q),
\end{aligned}$$

where $t_i \in \left(\frac{\tau M_{i-1}}{k}, \frac{\tau (M_{i-1}+1)}{k}\right)$. Using the first part of the lemma, we can write

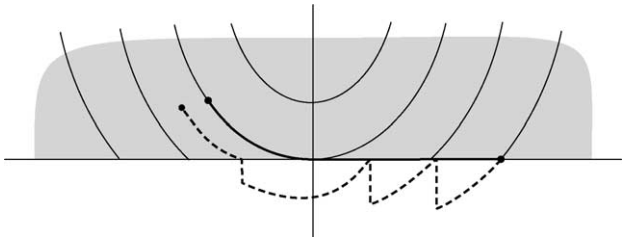

Fig. 1. The solid line shows the path from q to $\varphi(\tau, q)$. The dotted line shows the path from q to $\psi_3(q) = [\pi \circ \psi(\frac{\tau}{3}, \cdot)]^3(q)$.

$$\psi_k(\tau,q) = \left[\varphi\left(N_n\frac{\tau}{k},\cdot\right)\right]\circ\left[\pi\circ\psi\left(\frac{\tau}{k},\cdot\right)\right]\circ\left[\varphi\left(N_{n-1}\frac{\tau}{k},\cdot\right)\right]\circ\cdots$$
$$\circ\left[\varphi\left(N_1\frac{\tau}{k},\cdot\right)\right]\circ\left[\pi\circ\psi\left(\frac{\tau}{k},\cdot\right)\right]\circ\left[\varphi\left(N_0\frac{\tau}{k},\cdot\right)\right](q).$$

Now, $[0,\tau]\times M$ is compact, so there is a constant $K>0$ such that

$$d_{\mathrm{Euc}}\big(\psi(t,p),\psi(s,q)\big)\leqslant K\big(d_{\mathrm{Euc}}(p,q)+\|t-s\|\big).$$

Combining this estimate with the fact that π does not increase distances and the lemma, we get the estimate

$$d_{\mathrm{Euc}}\big(\psi_k(\tau,q),\varphi(\tau,q)\big)\leqslant 2\mu\frac{\tau}{k}\sum_{i=1}^{n}K^i\leqslant\frac{1}{k}\left(2\mu\tau_0\sum_{i=1}^{N}K^i\right).$$

This bound is independent of τ and q, so as $k\to\infty$, ψ_k converges uniformly to φ. Since the maps ψ_k are all continuous it follows that $\varphi:[0,\infty)\times M\to M$ is a continuous flow. □

2.6. Essential critical points

In classical Morse theory, critical points of a Morse function f appear as stationary points of the $-\nabla f$-flow. Analyzing the behavior of the flow near these points allows one to prove the Morse theorems. In Section 1.2 we defined a critical point to be any point p such that $-\nabla(f|_S)(p)=0$, where S is the stratum containing p. Which of these critical points are stationary points of the modified gradient flow?

A point p in the stratum $S\subseteq M$ will be a stationary point if $G(p)=0$. This means that the projection of $-\nabla f(p)$ onto any stratum other than S must point outward from M. This is equivalent to saying that if $\boldsymbol{x}:U_p\to\mathbb{H}^n_j$ is a standard coordinate chart near p, then $\boldsymbol{e}_i[f](p)>0$ for $i>n-j$.

Definition 8. An *essential critical point* is a point $p\in M$ satisfying $G(p)=0$.

In classical Morse theory, a critical point of f is labeled with a number called its index. The index λ of a critical point p is the number of negative eigenvalues of the Hessian matrix $H(p)$ of second partial derivatives of f at p. The lemma of Morse tells us that near such a critical point, there is a system of local coordinates $\boldsymbol{x}_p$ such that $f=f(p)-x_1^2-\cdots-x_\lambda^2+x_{\lambda+1}^2+\cdots+x_n^2$. Our situation requires a slight modification of this lemma.

Lemma 5. *Let p be an essential critical point of a Morse function f which satisfies Property* (3)*, and suppose that p is contained in a stratum S having dimension $n-j$. Then there is a local coordinate system $\boldsymbol{x}_p:U_p\to\mathbb{H}^n_j$ such that the identity*

$$f=f(p)-x_1^2-\cdots-x_\lambda^2+x_{\lambda+1}^2+\cdots+x_{n-j}^2+x_{n-j+1}^1+\cdots+x_n^1$$

holds throughout U_p.

Proof. We may assume without loss of generality that $f(p) = 0$. Since p is a non-degenerate critical point of $f|_S$, we can choose coordinates $(u_1, \ldots, u_{n-j})$ at p such that

$$f|_S = -u_1^2 - \cdots - u_\lambda^2 + u_{\lambda+1}^2 + \cdots + u_{n-j}^2,$$

and extend this to a standard coordinate chart $\boldsymbol{u}$ on M.

We can express f as a Taylor series in these coordinates:

$$f = -u_1^2 - \cdots - u_\lambda^2 + u_{\lambda+1}^2 + \cdots + u_{n-j}^2 + \sum_{i=n-j+1}^{n} u_i \left[f_{u_i} + \frac{1}{2} \sum_{j=1}^{n} f_{u_i u_j} u_j + \cdots \right],$$

then set $x_i = u_i$ for $1 \leqslant i \leqslant n-j$ and for $i > n-j$ define

$$x_i = u_i \left[f_{u_i} + \frac{1}{2} \sum_{j=1}^{n} f_{u_i u_j} u_j + \cdots \right].$$

Then for each i, $x_i(u_1, \ldots, u_n)$ is a smooth function. Let h be the map that carries $(u_1, \ldots, u_n)$ to $\boldsymbol{x}(u_1, \ldots, u_n)$. Then h is smooth and

$$\det\big(Dh(0)\big) = \det \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & 0 & \\ & & 1 & & & & \\ & & & f_{u_{n-j+1}} & & & \\ & & & & & \ddots & \\ & 0 & & & & & \\ & & & & & & f_{u_n} \end{bmatrix}.$$

Since p is an essential critical point, $f_{u_i} > 0$ for all $i > n-j$. Thus $\det(Dh(0)) \neq 0$. It follows from the Implicit Function Theorem that on some neighborhood of p, $\boldsymbol{x}$ is a coordinate system, compatible with $\boldsymbol{u}$.

Moreover, it is clear from the definition of x_i that in the domain of the coordinate chart $\boldsymbol{x}$, $x_i = 0$ if and only if $u_i = 0$. Further, $x_i > 0$ if and only if $u_i > 0$. Consequently, $\boldsymbol{x}$ is a standard coordinate system. In this coordinate system, f is given by

$$-x_1^2 - \cdots - x_\lambda^2 + x_{\lambda+1}^2 + \cdots + x_{n-j}^2 + x_{n-j+1}^1 + \cdots + x_n^1$$

as required. □

We will call the number λ the *index* of p, and we will take this lemma to be the definition of the index of an essential critical point. The coordinate system $\boldsymbol{x}$ in the lemma induces a coordinate system $\tilde{\boldsymbol{x}} = (x_1, \ldots, x_{n-j})$ on S. From this it is clear that λ is the index of p as a critical point of $f|_S$.

2.7. The Morse theorems

In this section we see how the number and type of essential critical points a function on a manifold with corners may have is governed by the topology of the domain. We will use the following notation: $M_a = f^{-1}((-\infty, a])$. We will assume that M_a is compact for each $a \in \mathbb{R}$.

Lemma 6. *If $a < b$, and $f^{-1}([a,b])$ contains no essential critical points, then there is a time $\tau > 0$ such that $\varphi(\tau, M_b) \subset M_a$.*

Proof. Suppose that there is a point $q \in f^{-1}([a,b])$ such that $\varphi(t,q) \notin f^{-1}((-\infty,a])$ for all $t > 0$. Let $\{q_i\}_{i>0}$ be the sequence $\varphi(i,q)$. Then $\{q_i\}$ is contained in the compact set $f^{-1}([a,b])$. Consequently, there is a subsequence of $\{q_i\}$ that converges to a limit $q_0 \in f^{-1}([a,b])$. We must have $f(\varphi(t,q)) > f(q_0)$ for all t and $\lim_{t\to\infty} f(\varphi(t,q)) = f(q_0)$.

Since q_0 is not an essential critical point, $G(q_0)$ is non-zero. We can choose some time t_0 such that $f(\varphi(t_0,q_0)) < f(q_0)$. Let U be a neighborhood of $\varphi(t_0,q_0)$ such that $f(U) < f(q_0)$. Since $\varphi(t_0,\cdot)$ is continuous, $\varphi(t_0,\cdot)^{-1}(U)$ is an open set containing q_0. It follows that there is some i such that $f(\varphi(i+t_0,q)) < f(q_0)$ which contradicts the assumption that $\{q_i\} \subset f^{-1}([a,b])$. □

We are now in a position to prove three of the central theorems of Morse theory.

Theorem 7. *Let $f : M \to \mathbb{R}$ be a Morse function satisfying Property* (3) *on a manifold with corners M. If $a < b$ and $f^{-1}([a,b])$ contains no essential critical points, then M_a is a deformation retract of M_b, so the inclusion map $M_a \hookrightarrow M_b$ is a homotopy equivalence.*

Proof. Since there are no essential critical points in $f^{-1}([a,b])$ and the value of f decreases along the flow lines of φ, for each point $p \in M_b$ there is a time t such that $\varphi(t,p) \in M_a$. Let $t_p = \inf\{t \in \mathbb{R}_+ \colon \varphi(t,p) \in M_a\}$.

Now we can define a homotopy $H : M_b \times [0,1] \to M_a$ by

$$H(p,s) = \begin{cases} \varphi(p, \frac{s}{1-s}), & \text{if } \frac{s}{1-s} \leqslant t_p, \\ \varphi(p,t_p), & \text{if } \frac{s}{1-s} \geqslant t_p. \end{cases} \quad \square$$

Theorem 8. *Let $f : M \to \mathbb{R}$ be a Morse function satisfying Property* (3) *on a manifold with corners M. Let p be an essential critical point with index λ. Set $f(p) = c$. Suppose that, for some $\varepsilon > 0$, $f^{-1}([c-\varepsilon, c+\varepsilon])$ contains no essential critical points other than p. Then $M_{c+\varepsilon}$ is homotopy equivalent to $M_{c-\varepsilon}$ with a λ-cell attached.*

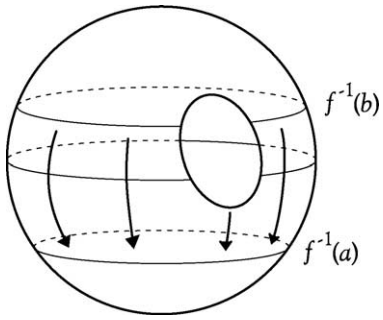

Fig. 2. The arrows illustrate the homotopy from M_b to M_a.

Proof. Choose a coordinate system $\boldsymbol{x}: U_p \to \mathbb{R}^{n-j} \times [0,\infty)^j$ in which we can write

$$f = f(p) - x_1^2 - \cdots - x_\lambda^2 + x_{\lambda+1}^2 + \cdots + x_{n-j}^2 + x_{n-j+1}^1 + \cdots + x_n^1.$$

Then choose $\varepsilon > 0$ sufficiently small so that $f^{-1}[c-\varepsilon, c+\varepsilon]$ contains no essential critical points other than p, and the image $\boldsymbol{x}(U_p)$ contains the closed 'ball'

$$\left\{(x_1,\ldots,x_n)\colon \sum_{i=1}^{n-j} x_i^2 + \sum_{i=n-j+1}^{n} x_i^1 \leqslant 2\varepsilon\right\}.$$

The proof from here will consist of the following three steps:

(1) Define a region H, as shown in Fig. 3.
(2) Show $M_{c-\varepsilon} \cup H \simeq M_{c+\varepsilon}$.
(3) Show $M_{c-\varepsilon} \cup e^\lambda \simeq M_{c-\varepsilon} \cup H$.

We begin by tweaking the function f a bit. Choose a C^∞ function $\mu: \mathbb{R} \to \mathbb{R}$ that satisfies

$$\mu(0) > \varepsilon,$$
$$\mu(r) = 0, \quad \text{for } r > 2\varepsilon,$$
$$-1 < \mu' \leqslant 0.$$

If we write

$$\xi = x_1^2 + \cdots + x_\lambda^2,$$
$$\eta = x_{\lambda+1}^2 + \cdots + x_{n-j}^2,$$
$$\zeta = x_{n-j+1}^1 + \cdots + x_n^1,$$

then we can write $f = c - \xi + \eta + \zeta$.

Define a new function F by

$$F = f - \mu(\xi + 2\eta + 2\zeta) = c - \xi + \eta + \zeta - \mu(\xi + 2\eta + 2\zeta).$$

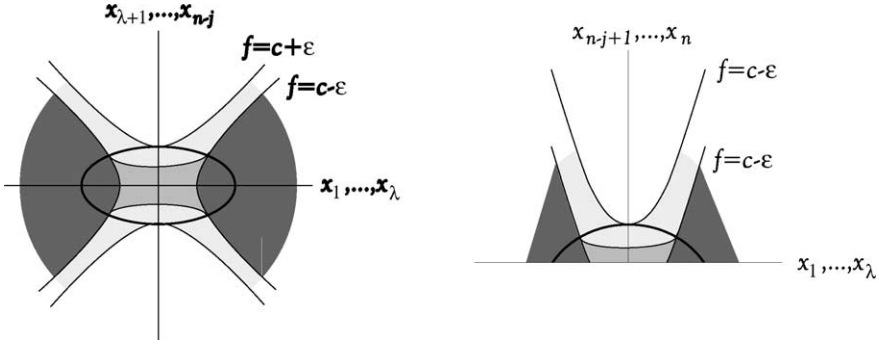

Fig. 3. M_b is the shaded region. M_a is the darkly shaded region. The heavy outline shows the set $\{\xi + 2\eta + 2\zeta = \varepsilon\}$, and the medium shaded region is H.

We will use this function (and its level sets) to define the region H.

Claim 1. *The essential critical points of F and f are identical.*

Outside our 'ball' of 'radius' 2ε, $F = f$ and so any critical points there must coincide. Inside, the function f has a single essential critical point at p. To find the essential critical points of F we must compute $\mathrm{d}F$.

$$\mathrm{d}F = (-1-\mu')\,\mathrm{d}\xi + (1-2\mu')\,\mathrm{d}\eta + (1-2\mu')\,\mathrm{d}\zeta.$$

The coefficients $(-1-\mu')$ and $(1-2\mu')$ are nowhere zero and $\mathrm{d}\xi$ and $\mathrm{d}\eta$ are simultaneously zero only at p. Thus p is an essential critical point provided that $\boldsymbol{e}_i[F](p) > 0$ for $i > n-j$. A computation shows that

$$\boldsymbol{e}_i[F](p) = \mathrm{d}F(\boldsymbol{e}_i)(p) = (1-2\mu')\,\mathrm{d}\zeta(\boldsymbol{e}_i)(p) = (1-2\mu')(1) > 0,$$

so p is indeed an essential critical point of F.

Claim 2. $F^{-1}(-\infty, c+\varepsilon) = f^{-1}(-\infty, c+\varepsilon)$.

Outside the set $\{\xi + 2\eta + 2\zeta \leqslant 2\varepsilon\}$ we know that $\mu = 0$, so $F = f$. Inside this set, we see that

$$F \leqslant f = c - \xi + \eta + \zeta.$$

Equality holds on the boundary of the 'ball'. Also,

$$c - \xi + \eta + \zeta \leqslant c + \left(\frac{1}{2}\xi + \eta + \zeta\right).$$

Here equality holds when $\xi = 0$. Finally, we note that

$$c + \left(\frac{1}{2}\xi + \eta + \zeta\right) \leqslant c + \varepsilon.$$

Here, again, equality holds on the boundary of the 'ball'. So we see that within this set, $F \leqslant c+\varepsilon$ and $f \leqslant c+\varepsilon$ unless $\xi = 0$ and $\eta + \zeta = \varepsilon$, in which case $F = f = c+\varepsilon$.

Claim 3. *$F^{-1}(-\infty, c-\varepsilon]$ is a deformation retract of $M_{c+\varepsilon}$.*

Consider the region $F^{-1}[c-\varepsilon, c+\varepsilon]$. It is compact, but does it contain any critical points? The only possibility is p, but

$$F(p) = c - \mu(0) < c - \varepsilon,$$

so $p \notin F^{-1}[c-\varepsilon, c+\varepsilon]$ and Theorem 7 applies: $F^{-1}(-\infty, c-\varepsilon]$ is a deformation retract of

$$F^{-1}(-\infty, c+\varepsilon] = f^{-1}(-\infty, c+\varepsilon] = M_{c+\varepsilon}.$$

Define the region H by

$$H = \overline{F^{-1}(-\infty, c-\varepsilon] - M_{c-\varepsilon}}.$$

Recall that we have defined the λ-cell e^λ such that $e^\lambda = \{q\colon\ \xi(q) < \varepsilon,\ \eta(q) = \zeta(q) = 0\}$. Note that $e^\lambda \subseteq H$, since $\frac{\partial F}{\partial \xi} = -1 - \mu' < 0$ implies for $q \in e^\lambda$,

$$F(q) < F(p) < c - \varepsilon.$$

Note also that $e^\lambda \cap M_{c-\varepsilon} = \partial e^\lambda$.

Claim 4. *$M_{c-\varepsilon} \cup e^\lambda$ is a deformation retract of $M_{c-\varepsilon} \cup H$.*

For each $t \in [0, 1]$ we define a map $r_t : M_{c-\varepsilon} \cup H \to M_{c-\varepsilon} \cup H$ as follows:

Case 1. If $q \in M_{c-\varepsilon}$, set $r_t(q) = q$ for all t.

Case 2. If $q \in H$ and $\xi(q) < \varepsilon$, then set

$$r_t(x_1, \ldots, x_n) = \big(x_1, \ldots, x_\lambda, (1-t)x_{\lambda+1}, \ldots, (1-t)x_n\big).$$

Case 3. If $\varepsilon \leqslant \xi(q) \leqslant \eta(q) + \zeta(q) + \varepsilon$, then define r_t by

$$r_t(x_1, \ldots, x_n) = (x_1, \ldots, x_\lambda, s_t x_{\lambda+1}, \ldots, s_t x_n),$$

where

$$s_t = (1-t) + t\left[\frac{\xi - \varepsilon}{\eta + \zeta}\right]^{1/2}.$$

Then r_0 is the identity map, and $r_1 : M_{c-\varepsilon} \cup H \to M_{c-\varepsilon} \cup e^\lambda$. Moreover, $r_t(q) \in e^\lambda$ for each t, because $\frac{\partial F}{\partial \eta} > 0$ and $\frac{\partial F}{\partial \zeta} > 0$. (Moving toward e^λ decreases F.)

We must show that the functions $s_t x_i$ are continuous as $\xi \to \varepsilon$, $\eta \to 0$, $\zeta \to 0$. Since $x_i \to 0$ as $\eta + \zeta \to 0$,

$$\lim_{\eta+\zeta\to 0}\left[\frac{\xi - \varepsilon}{\eta + \zeta}\right]^{1/2} x_i = \left[\lim_{\eta+\zeta\to 0}\frac{\xi - \varepsilon}{\eta + \zeta}\right]^{1/2}(0).$$

Since

$$0 = \frac{(\varepsilon) - \varepsilon}{\eta + \zeta} \leqslant \frac{\xi - \varepsilon}{\eta + \zeta} \leqslant \frac{(\eta + \zeta + \varepsilon) - \varepsilon}{\eta + \zeta} = 1,$$

the limit is zero, and it follows that each $s_t x_i$ is continuous.

Note that this definition agrees with Case 1 when $\xi = \varepsilon$ and with Case 2 when $\xi - \eta - \zeta = \varepsilon$. Thus r provides a deformation retraction of $M_{c-\varepsilon} \cup H$ to $M_{c-\varepsilon} \cup e^\lambda$. This concludes the proof of Theorem 8. □

Theorems 7 and 8 together imply

Theorem 9 (Main Theorem). *If M is a manifold with corners, $f : M \to \mathbb{R}$ a Morse function on M which satisfies Property* (3) *and $f^{-1}(-\infty, c]$ is compact for each c, then M has the homotopy type of a CW complex with one cell of dimension λ for each essential critical point with index λ.*

Finally, because the homology groups are a homotopy invariant, the Morse inequalities hold for a Morse function on a manifold with corners, $f : M \to \mathbb{R}$. If $b_i(M)$ is the ith Betti number of M, and $m_i(f)$ is the number of essential critical points of f with index i, then

$$\sum_{i=0}^{k}(-1)^{k+i}b_i(M) \leqslant \sum_{i=0}^{k}(-1)^{k+i}m_i(f).$$

These are the Strong Morse Inequalities. It is a simple matter to deduce from these the Weak Morse Inequalities:

$$m_k(f) \geqslant b_k(M) \quad \text{for each } k \geqslant 0.$$

3. Generalized billiard paths

3.1. Statement of the problem

We now return to the problem posed in Section 1.1. We have a compact n-manifold embedded in some Euclidean space, $M \hookrightarrow \mathbb{R}^N$. Given $p, q \in \mathbb{R}^N$ we wish to count the number of generalized billiard paths from p to q:

Definition 9. A sequence $P = \{\alpha_1, \ldots, \alpha_k\}$ connecting $p = \alpha_0$ to $q = \alpha_{k+1}$ is a *generalized billiard path* with k-reflections if for each i one of the following is true:

(1) The bisector of $\measuredangle\alpha_{i-1}\alpha_i\alpha_{i+1}$ is normal to $T_{\alpha_i}M$.
(2) $\measuredangle\alpha_{i-1}\alpha_i\alpha_{i+1}$ is a straight angle.

3.2. The length of a sequence

We can define the length of a sequence $P = \{\alpha_1, \ldots, \alpha_k\}$ connecting p and q to be

$$L_k^{(p,q)}(P) = \sum_{i=0}^{k} d_{\text{Euc}}(\alpha_i, \alpha_{i+1}),$$

and think of $L_k^{(p,q)}$ as a function

$$L_k^{(p,q)} : \underbrace{M \times \cdots \times M}_{k \text{ copies}} \to \mathbb{R}.$$

When there is no confusion regarding the endpoints, we will write L_k for $L_k^{(p,q)}$.

This length function has one bad property. Wherever consecutive points of a sequence coincide, L_k has a singularity that looks like $|x - y|$. It has another property, though, that makes us willing to put up with this. Away from this bad set, we can compute ∇L_k. Paths for which $\nabla L_k = 0$ will be of special interest, as shown by the following

Lemma 10. *A sequence $P = \{\alpha_1, \ldots, \alpha_k\}$ with $\alpha_i \neq \alpha_{i+1}$ for $0 \leqslant i \leqslant k$ satisfies*

$$\nabla L_k(\alpha_1, \ldots, \alpha_k) = 0$$

if and only if it is a generalized billiard path.

Proof. $\nabla L_k = 0$ if an only if the directional derivative $\boldsymbol{v}[L_k] = 0$ for all $\boldsymbol{v} \in T(M^k)$. Since we can identify $T(M^k)$ with $TM \oplus \cdots \oplus TM$, we can write $\boldsymbol{v} = \boldsymbol{v}_1 \oplus \cdots \oplus \boldsymbol{v}_k$. Then, since $(\boldsymbol{v}_1 \oplus \cdots \oplus \boldsymbol{v}_k)[L_k] = \boldsymbol{v}_1[L_k] + \cdots + \boldsymbol{v}_k[L_k]$, it is sufficient to show that $\nabla L_k = 0$ if and only if $\boldsymbol{v}_i[L_k] = 0$ for any choice of $\boldsymbol{v}_i$.

Now, $\boldsymbol{v}_i[L_k] = \boldsymbol{v}_i[\sum_{j=0}^{k} d_{\mathrm{Euc}}(\alpha_j, \alpha_{j+1})]$. Only two of the terms on the right are non-zero:

$$\boldsymbol{v}_i[L_k] = \boldsymbol{v}_i\big[d_{\mathrm{Euc}}(\alpha_{i-1}, \alpha_i)\big] + \boldsymbol{v}_i\big[d_{\mathrm{Euc}}(\alpha_i, \alpha_{i+1})\big].$$

In order to compute $\boldsymbol{v}_i[d_{\mathrm{Euc}}(\alpha_{i-1}, \alpha_i)]$ we can choose a curve in M (which we will also call α_i) satisfying

$$\alpha_i(0) = \alpha_i \quad \text{and} \quad \alpha_i'(0) = \boldsymbol{v}_i.$$

In this way, we think of α_i as varying along the curve, rather than as a fixed point. Then

$$\boldsymbol{v}_i\big[d_{\mathrm{Euc}}(\alpha_{i-1}, \alpha_i)\big] = \frac{\mathrm{d}}{\mathrm{d}t} L_k\big(\alpha_1, \ldots, \alpha_i(t), \ldots, \alpha_k\big)|_{t=0}.$$

Then we can compute

$$\begin{aligned}
&\boldsymbol{v}_i\big[d_{\mathrm{Euc}}(\alpha_{i-1}, \alpha_i) + d_{\mathrm{Euc}}(\alpha_i, \alpha_{i+1})\big] \\
&\quad = \frac{\partial}{\partial t}\big[\big(\alpha_{i-1} - \alpha_i(t)\big) \cdot \big(\alpha_{i-1} - \alpha_i(t)\big)\big]^{1/2} + \big[\big(\alpha_i(t) - \alpha_{i+1}\big) \cdot \big(\alpha_i(t) - \alpha_{i+1}\big)\big]^{1/2} \\
&\quad = \frac{-\alpha_i'(t) \cdot (\alpha_{i-1} - \alpha_i(t))}{[(\alpha_{i-1} - \alpha_i(t)) \cdot (\alpha_{i-1} - \alpha_i(t))]^{1/2}} + \frac{\alpha_i'(t) \cdot (\alpha_i(t) - \alpha_{i+1})}{[(\alpha_i(t) - \alpha_{i+1}) \cdot (\alpha_i(t) - \alpha_{i+1})]^{1/2}} \\
&\quad = \alpha_i'(t) \cdot \left(\frac{\alpha_{i-1} - \alpha_i(t)}{\|\alpha_{i-1} - \alpha_i(t)\|} + \frac{\alpha_i(t) - \alpha_{i+1}}{\|\alpha_i(t) - \alpha_{i+1}\|} \right).
\end{aligned}$$

Evaluating at $t = 0$ we find

$$\boldsymbol{v}_i\big[d_{\mathrm{Euc}}(\alpha_{i-1}, \alpha_i) + d_{\mathrm{Euc}}(\alpha_i, \alpha_{i+1})\big] = \boldsymbol{v}_i \cdot \left(\frac{\alpha_{i-1} - \alpha_i}{\|\alpha_{i-1} - \alpha_i\|} + \frac{\alpha_i - \alpha_{i+1}}{\|\alpha_i - \alpha_{i+1}\|} \right).$$

This will be zero for all $\boldsymbol{v}_i$ provided that the vector

$$\left(\frac{\alpha_{i-1} - \alpha_i}{\|\alpha_{i-1} - \alpha_i\|} + \frac{\alpha_i - \alpha_{i+1}}{\|\alpha_i - \alpha_{i+1}\|} \right)$$

is either normal to $T_{\alpha_i} M$ or zero. When this vector is non-zero it is a bisector of the angle $\measuredangle \alpha_{i-1} \alpha_i \alpha_{i+1}$. When it is zero, $\measuredangle \alpha_{i-1} \alpha_i \alpha_{i+1}$ is a straight angle. Thus the gradient is zero exactly when the sequence is a generalized gradient path. □

Having established that L_k is a function worth considering, let's look more closely at its behavior near the diagonals $\{\alpha_i = \alpha_{i+1}\}$. Consider a sequence $\{\alpha, \beta, \gamma\}$ where $\alpha = \beta$. If β moves slightly to β', as in Fig. 4, the triangle inequality tells us that we have increased the length of the sequence.

We know that the vector field $-\nabla L_k$ points in the direction of decreasing length. Consequently, under any modified gradient flow, nearby consecutive reflections would tend to flow toward each other. We will define such a flow in Section 4.2, but for now we are interested in generalized billiard paths with k *distinct* reflections. Consequently we will

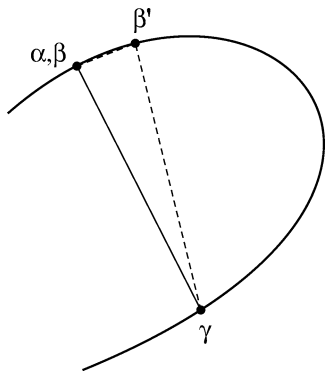

Fig. 4. When β is moved away from α the length of the path increases.

instead look at the function $-L_k$. It has the same critical points, but the ∇L_k-flow tends away from the diagonals. Then we will 'blow up' the product M^k along the appropriate diagonals.

3.3. The blow up space

This notion of blowing up was introduced by Fulton and MacPherson in [4]. To understand what is meant by blow up, let's think about a simple example. Take $M = S^1$ and $k = 2$. Assume that $p, q \notin M$. Then $-L_2$ is a function on the torus which is singular along the diagonal $\Delta \subset S^1 \times S^1$.

As $(\alpha, \beta) \to (\alpha, \alpha)$ the limit of $-\nabla L_2$ depends on the direction of approach. The gradient vector $-\nabla L_k(\alpha, \beta)$ consists of a vector in $T_\alpha M$ pointing away from β and a vector in $T_\beta M$ pointing away from α. If β is allowed to approach α from the opposite side, the gradient vector is reversed.

We need to produce a closure of $S^1 \times S^1 - \Delta$ on which we can extend ∇L_2 continuously. Consequently, as (α, β) approaches Δ, we keep track not only of the limiting point, but also of the relative positions of α and β. The result is shown in Fig. 5.

Now lets consider a path with k reflections on an n-manifold M. Fig. 6 shows the situation when two consecutive points coincide. This collision is described by the limiting point and an *infinitesimal tangent space diagram*. This diagram shows points $\boldsymbol{v}_\alpha$ and $\boldsymbol{v}_\beta$ in the tangent space of the limiting point. Two such diagrams are equivalent if they differ by translation and multiplication by a positive constant. We can translate the diagram so that $\boldsymbol{v}_\alpha$ is at the origin, and then scale it so $\boldsymbol{v}_\beta$ is on the unit circle. This shows that each such point will be blown up into a copy of S^{n-1}.

Fig. 7 shows what may happen when α, β, and γ coincide at a point $\theta \in M$. The situation is a bit more complicated now. Again, we can translate the diagram so that $\boldsymbol{v}_\alpha$ is at the origin, and then scale it so $\boldsymbol{v}_\gamma$ is on the unit circle. The point $\boldsymbol{v}_\beta$ now may lie anywhere in $T_\theta M \cup \infty$. It would seem that each such point θ is blown up to a copy of $S^{n-1} \times S^n$. In fact this is not the case. Whenever $\boldsymbol{v}_\beta = \boldsymbol{v}_\alpha$ or $\boldsymbol{v}_\beta = \boldsymbol{v}_\gamma$ the resulting double point must also be blown up. On the other hand, if scaling the diagram so that $\boldsymbol{v}_\gamma$ is on the unit circle pushes $\boldsymbol{v}_\beta$ off to infinity, we would do better to scale the diagram so that $\boldsymbol{v}_\beta$ is on

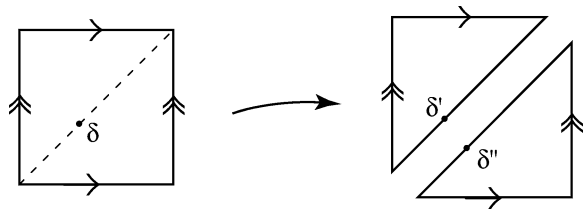

Fig. 5. The point $\delta \in \Delta$ is blown up to the two points δ' and δ''.

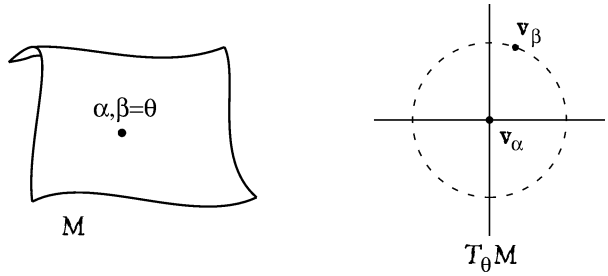

Fig. 6. The infinitesimal tangent space diagram for two consecutive reflections colliding in M.

the unit circle. (We are free to choose, since all these diagrams are equivalent.) When we rescale in this fashion, we will find that $\boldsymbol{v}_\gamma = \boldsymbol{v}_\alpha$. This point need not be blown up further, because α and γ are not consecutive reflections. All of these special situations correspond to a situation where two of the points approach each other much more quickly than they approach the third.

When more points collide, there will be more of these cascading diagrams. In addition, two collections of points may collide independently at different points in the manifold. In this case we have two separate collections of infinitesimal diagrams corresponding to the two collections of points. We will denote the space that results from blowing up M^k in this way by $X_k = X_k(M)$. The spaces that result are somewhat difficult to describe. There is one thing we can say about these spaces which is of particular importance to us.

Lemma 11. *For any smooth manifold M, the space $X_k = X_k(M)$ is a manifold with corners.*

Proof. It is shown in [4] that the result of blowing up all the diagonals is a manifold with corners. In our case, we are only concerned with the diagonals corresponding to the collision of consecutive reflections. Here we show that blowing up only these diagonals also leads to a manifold with corners.

First, we define some convenient notation for referring to a stratum of the blow up X_k. When we write

$$\big(\alpha_1, \ldots, \alpha_{i-1}, \{\alpha_i, \ldots, \alpha_{i+j}\}, \ldots, \alpha_k\big),$$

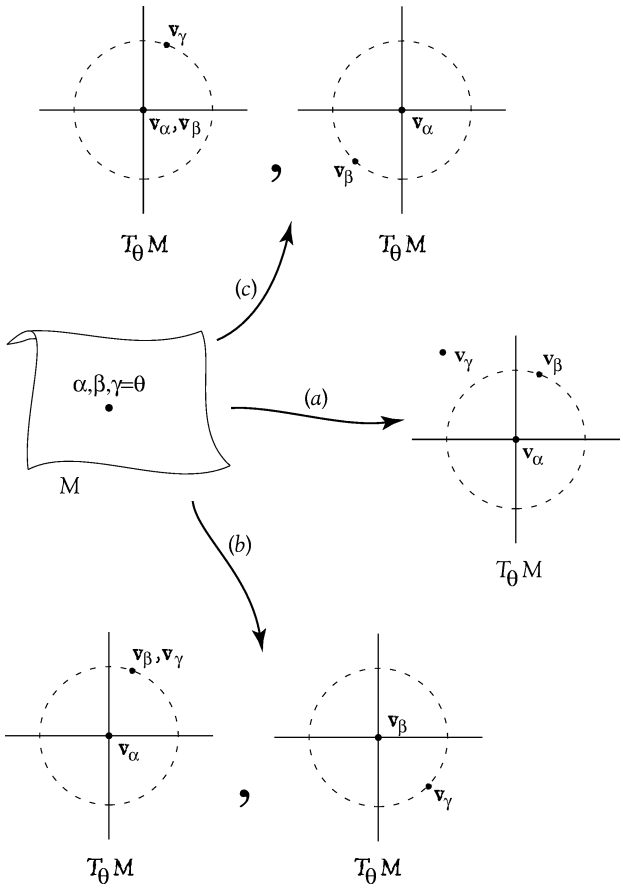

Fig. 7. Possible infinitesimal tangent space diagrams for three consecutive reflections colliding in M. In (a), α, β and γ all approach each other at approximately the same rate. In (b), β and γ approach each other much faster then they approach α. In (c), α and β approach each other much faster than they approach γ. The situation where α and γ approach each other faster than they approach β need not be considered separately, since these reflections are not consecutive.

we mean that $\alpha_i = \cdots = \alpha_{i+j}$, and all these points come together at commensurable rates. This stratum will be described by an infinitesimal diagram in $T_\theta M$ in which $\boldsymbol{v}_i \neq \cdots \neq \boldsymbol{v}_{i+j}$. Furthermore, when we write

$$\big(\ldots, \big\{\alpha_i, \ldots, \{\alpha_{i+\ell}, \ldots, \alpha_{i+\ell+m}\}, \ldots, \alpha_{i+j}\big\}, \ldots\big)$$

we mean that $\boldsymbol{v}_{i+\ell} = \cdots = \boldsymbol{v}_{i+\ell+m}$ in the first infinitesimal diagram, requiring a second diagram.

Each pair of braces must enclose a proper subset of the points in preceding set of braces. Each grouping designates a stratum with as many infinitesimal diagrams as there are pairs of braces. Moreover, for each pair of braces we add, the codimension of the stratum is increased by one. To see this, consider what happens when we add a single set of braces:

$$\big(\ldots, \{\alpha_i, \ldots, \alpha_{i+j}\}, \ldots\big).$$

Before the braces were added, these points represented $j+1$ distinct points in an infinitesimal diagram (or in M^k)—an $n(j+1)$-dimensional set. They now represent a single point in an n-dimensional space, and a new diagram. We can scale the diagram so that α_i is at the origin and α_{i+j} is on the unit circle. The remaining $j-1$ points lie elsewhere in the tangent plane.

Altogether, we note that with the braces in place, the points $\alpha_i, \ldots, \alpha_{i+j}$ account for $n+(n-1)+n(j-1)$ or $n(j+1)-1$ dimensions. Thus adding the braces decreases the dimension by one.

To see how a coordinate chart $\boldsymbol{x}$ may be defined near a point on this stratum, first choose a coordinate chart $\boldsymbol{u}_i$ at each of the distinct points α_i in M. The chart $\boldsymbol{u}_\theta$ at θ induces a coordinate chart $\boldsymbol{w}_1$ on $T_{\theta'}M$ for θ' near θ. Then we may choose coordinates $\boldsymbol{w}_2$ on the unit sphere in $T_{\theta'}M$ that vary smoothly with θ'.

When $\alpha_i, \ldots, \alpha_{i+j}$ are all sufficiently close together we can write uniquely

$$(\alpha_i, \ldots, \alpha_{i+j}) = \big(\exp_{\theta'}(t\boldsymbol{v}_i), \ldots, \exp_{\theta'}(t\boldsymbol{v}_{i+j})\big),$$

by requiring $\boldsymbol{v}_i = 0$ (so that $\theta' = \alpha_i$), $|\boldsymbol{v}_{i+1}| = 1$ and $t \geqslant 0$. Then the limit as $t \to 0$ is the infinitesimal diagram defined by $\{\boldsymbol{v}_i, \ldots, \boldsymbol{v}_{i+1}\}$. Set, for $\boldsymbol{v}_i = 0$ and $|\boldsymbol{v}_{i+1}| = 1$ and θ' in a small neighborhood of θ,

$$\begin{aligned}\boldsymbol{x}\big(&\alpha_1, \ldots, \alpha_{i-1}, \exp_{\theta'}(t\boldsymbol{v}_i), \ldots, \exp_{\theta'}(t\boldsymbol{v}_{i+j}), \alpha_{i+j+1}, \ldots, \alpha_k\big)\\ &= \big(\boldsymbol{u}_1(\alpha_1), \ldots, \boldsymbol{u}_{i-1}(\alpha_{i-1}), \boldsymbol{u}_\theta(\theta'), \boldsymbol{u}_{i+j+1}(\alpha_{i+j+1}), \ldots, \boldsymbol{u}_k(\alpha_k),\\ &\qquad \boldsymbol{w}_1(\boldsymbol{v}_i), \boldsymbol{w}_2(\boldsymbol{v}_{i+1}), \boldsymbol{w}_1(\boldsymbol{v}_{i+2}), \ldots, \boldsymbol{v}_{i+j}, t\big).\end{aligned}$$

Then on a neighborhood of $(\alpha_1, \ldots, \alpha_k)$ this map defines a coordinate chart.

The same procedure can be used for any grouping of the α_i's, using one parameter $0 \leqslant t_i \in \mathbb{R}$ for each pair of braces. □

There is a map $g : X_k \to M^k$ that assigns to each point in X_k the corresponding limiting point in M^k. We can define (abusing notation in the process)

$$-L_k : X_k \to \mathbb{R}$$

by

$$-L_k(q) = -L_k \circ g(q).$$

Now, we wish to study this function on X_k. There is just one more order of business to attend to first.

3.4. When is $-L_k$ a Morse function?

We want to show that $-L_k$ satisfies the properties in Definitions 4 and 6. First of all, we must show that ∇L_k extends continuously to X_k. Recall the definition of L_k:

$$L_k(P) = \sum_{i=0}^{k} d_{\mathrm{Euc}}(\alpha_i, \alpha_{i+1}).$$

It is sufficient to show that $\nabla d_{\mathrm{Euc}}(\alpha_i, \alpha_{i+1})$ extends continuously for each i. If α_i and α_{i+1} do not approach each other, then $d_{\mathrm{Euc}}(\alpha_i, \alpha_{i+1})$ is smooth as it approaches the boundary of X_k, and $\nabla d_{\mathrm{Euc}}(\alpha_i, \alpha_{i+1})$ can be extended continuously. If α_i and α_{i+1} do approach each other, we must show that $\nabla d_{\mathrm{Euc}}(\alpha_i, \alpha_{i+1})$ approaches a limit.

As the points α_i and α_{i+1} approach each other, we can write

$$\alpha_j = \exp_{\theta'}\bigl(t\boldsymbol{v}'_j\bigr),$$

where $\theta' \to \theta$ and $\boldsymbol{v}'_j \to \boldsymbol{v}_j \in T_\theta M$ as $t \to 0$. Then, since $D(\exp_{\theta'})_0$ is the identity on $T_{\theta'}M$, we can write,

$$\alpha_j = \theta' + t\boldsymbol{v}'_j + \mathrm{O}\bigl(t^2\bigr).$$

Here we are thinking of $T_{\theta'}M$ as a linear subspace of $\mathbb{R}^N$. Since M is compact, $\mathrm{O}(t^2)$ is a uniform bound for bounded $\boldsymbol{v}'_j$. Then the distance from α_i to α_{i+1} is given by

$$d_{\mathrm{Euc}}(\alpha_i, \alpha_{i+1}) = t\bigl|\boldsymbol{v}'_i - \boldsymbol{v}'_{i+1}\bigr| + \mathrm{O}\bigl(t^2\bigr).$$

So $\nabla d_M(\alpha_i, \alpha_{i+1})$ is given by

$$\left(\frac{\boldsymbol{v}'_i - \boldsymbol{v}'_{i+1}}{|\boldsymbol{v}'_i - \boldsymbol{v}'_{i+1}|}\right) + \mathrm{O}\bigl(t^2\bigr) \oplus \left(\frac{\boldsymbol{v}'_{i+1} - \boldsymbol{v}'_i}{|\boldsymbol{v}'_{i+1} - \boldsymbol{v}'_i|}\right) + \mathrm{O}\bigl(t^2\bigr).$$

The first vector in the sum is in $T_{\alpha_i}M$. The second lies in $T_{\alpha_{i+1}}M$. The limit as $t \to 0$ exists and is equal to

$$\left(\frac{\boldsymbol{v}_i - \boldsymbol{v}_{i+1}}{|\boldsymbol{v}_i - \boldsymbol{v}_{i+1}|}\right) \oplus \left(\frac{\boldsymbol{v}_{i+1} - \boldsymbol{v}_i}{|\boldsymbol{v}_{i+1} - \boldsymbol{v}_i|}\right) \in T_\theta M \oplus T_\theta M,$$

so ∇L_k extends continuously to all of X_k.

Since ∇L_k points inward at each point in ∂X_k, all the essential critical points of $-L_k$ are in the interior of X_k. To show that $-L_k$ is a Morse function, we have only to determine when the Hessian is non-singular.

This property requires that the critical points be non-degenerate, i.e., the determinant of the Hessian at a critical point must be non-zero. To begin with lets look at

$$-L_k : X_k \to \mathbb{R}.$$

Here there are k reflection points, $\alpha_1, \ldots, \alpha_k$. Choose an orthonormal coordinate system $\boldsymbol{x}^i$ satisfying $\boldsymbol{x}^i(\alpha_i) = 0$ for each i. The function then can be written as

$$-L_k = -\sum_{i=0}^{k} \|\alpha_i - \alpha_{i+1}\|.$$

Our first goal is to get an explicit representation for the Hessian.

The Hessian is given by

Lemma 12. *The Hessian of $-L_k$ is given by*

$$H(\alpha_1,\dots,\alpha_k)=\begin{bmatrix} A_1 & K_1 & & & & & & \\ K_1 & A_2 & K_2 & & & & 0 & \\ & K_2 & \ddots & & & & & \\ & & & \ddots & K_{\ell-1} & & & \\ & & & K_{\ell-1} & A_\ell & K_\ell & & \\ & & & & K_\ell & \ddots & & \\ & 0 & & & & & \ddots & K_{k-1} \\ & & & & & & K_{k-1} & A_k \end{bmatrix}$$

where

$$\begin{aligned}(A_\ell)_{i,j} &= \beta_{x_j x_i}\cdot(\boldsymbol{v}_{\alpha\beta}-\boldsymbol{v}_{\beta\gamma})\\ &\quad+\big[\cos(\psi_{x_j})\cos(\psi_{x_i})-\delta_{ij}\big]\left(\frac{1}{\|\alpha-\beta\|}+\frac{1}{\|\beta-\gamma\|}\right),\end{aligned}$$

and

$$(K_\ell)_{i,j}=\frac{1}{\|\beta-\gamma\|}\big(\beta_{x_i}\cdot\gamma_{y_j}-\cos(\psi_{x_i})\cos(\psi_{y_j})\big).$$

In the above expressions $\boldsymbol{v}_{\alpha\beta}=\frac{\alpha-\beta}{\|\alpha-\beta\|}$ and δ_{ij} is the Kronecker delta. Also $\alpha=\alpha_{\ell-1}$, $\beta=\alpha_\ell$, $\gamma=\alpha_{\ell+1}$, $\boldsymbol{x}$ is a coordinate system at β, and $\boldsymbol{y}$ is a coordinate system at γ.

Proof. We need to compute A_ℓ and K_ℓ. First compute

$$\begin{aligned}\frac{\partial(-L_k)}{\partial x_i} &= \frac{\partial}{\partial x_i}\big(-\|\alpha-\beta\|-\|\beta-\gamma\|\big)=-\frac{-\boldsymbol{\beta}_{x_i}\cdot(\alpha-\beta)}{\|\alpha-\beta\|}-\frac{\boldsymbol{\beta}_{x_i}\cdot(\beta-\gamma)}{\|\beta-\gamma\|}\\ &= \boldsymbol{\beta}_{x_i}\cdot\left(\frac{\alpha-\beta}{\|\alpha-\beta\|}-\frac{\beta-\gamma}{\|\beta-\gamma\|}\right).\end{aligned}$$

In order to compute K we must differentiate in one $\boldsymbol{x}$ variable and one $\boldsymbol{y}$ variable. In this case, the result is

$$\begin{aligned}\frac{\partial^2(-L_k)}{\partial y_j\partial x_i} &= \beta_{x_i}\cdot\frac{\partial}{\partial y_j}\left(-\frac{\beta-\gamma}{\|\beta-\gamma\|}\right)\\ &= \beta_{x_i}\cdot\left(\frac{\|\beta-\gamma\|\gamma_{y_j}-(\beta-\gamma)\frac{\gamma_{y_j}(\beta-\gamma)}{\|\beta-\gamma\|}}{\|\beta-\gamma\|^2}\right).\end{aligned}$$

This expression can be simplified by setting $\boldsymbol{v}_{\beta\gamma}=(\beta-\gamma)/\|\beta-\gamma\|$ and letting ψ_{x_i} be the angle between $\boldsymbol{v}_{\beta\gamma}$ and β_{x_i}. Similarly, ψ_{y_j} will be the angle between $\boldsymbol{v}_{\beta\gamma}$ and γ_{y_j}. Then

$$\begin{aligned}\frac{\partial^2(-L_k)}{\partial y_j\partial x_i} &= \beta_{x_i}\cdot\left(\frac{\gamma_{y_j}}{\|\beta-\gamma\|}-\frac{\beta-\gamma}{\|\beta-\gamma\|^2}\cos(\psi_{y_j})\right)\\ &= \frac{1}{\|\beta-\gamma\|}\big(\beta_{x_i}\cdot\gamma_{y_j}-\cos(\psi_{x_i})\cos(\psi_{y_j})\big).\end{aligned}$$

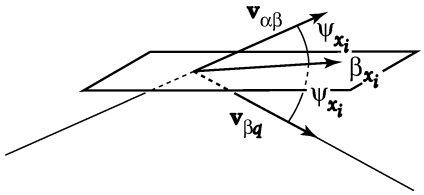

Fig. 8. The angle ψ_{x_i}.

In order to compute A, we must differentiate by two different $\boldsymbol{x}$-variables. In that case, we find

$$\frac{\partial^2(-L_k)}{\partial x_j \partial x_i} = \beta_{x_j x_i} \cdot \left(\frac{\alpha-\beta}{\|\alpha-\beta\|} - \frac{\beta-\gamma}{\|\beta-\gamma\|} \right) + \beta_{x_i} \cdot \frac{\partial}{\partial x_j} \left(\frac{\alpha-\beta}{\|\alpha-\beta\|} - \frac{\beta-\gamma}{\|\beta-\gamma\|} \right)$$

and

$$\begin{aligned} &\frac{\partial}{\partial x_j} \left(\frac{\alpha-\beta}{\|\alpha-\beta\|} - \frac{\beta-\gamma}{\|\beta-\gamma\|} \right) \\ &\quad = \frac{-\|\alpha-\beta\|\beta_{x_j} - (\alpha-\beta)\frac{-\beta_{x_j}(\alpha-\beta)}{\|\alpha-\beta\|}}{\|\alpha-\beta\|^2} - \frac{\|\beta-\gamma\|\beta_{x_j} - (\beta-\gamma)\frac{\beta_{x_j}(\beta-\gamma)}{\|\beta-\gamma\|}}{\|\beta-\gamma\|^2} \\ &\quad = \frac{-\beta_{x_j}}{\|\alpha-\beta\|} + \frac{\boldsymbol{v}_{\alpha\beta}\cos(\psi_{x_j})}{\|\alpha-\beta\|} - \frac{\beta_{x_j}}{\|\beta-\gamma\|} + \frac{\boldsymbol{v}_{\beta\gamma}\cos(\psi_{x_j})}{\|\beta-\gamma\|}. \end{aligned}$$

Since we are at a critical point, it is easily shown that $\beta_{x_i} \cdot \boldsymbol{v}_{\alpha\beta} = \beta_{x_i} \cdot \boldsymbol{v}_{\beta\gamma}$. From this it follows that the angle between β_{x_i} and $\boldsymbol{v}_{\beta\gamma}$ is also ψ_{x_i}. Using this and the orthonormality of the coordinate system $\boldsymbol{x}$, we can write

$$\frac{\partial^2(-L_k)}{\partial x_j \partial x_i} = \beta_{x_j x_i} \cdot (\boldsymbol{v}_{\alpha\beta} - \boldsymbol{v}_{\beta\gamma}) + \cos(\psi_{x_i})\cos(\psi_{x_j}) \left(\frac{1}{\|\alpha-\beta\|} + \frac{1}{\|\beta-\gamma\|} \right).$$

We also must compute $\frac{\partial^2(-L_2)}{\partial {x_i}^2}$. Here it is:

$$\begin{aligned} \frac{\partial^2(-L_2)}{\partial x_i \partial x_i} &= \beta_{x_i x_i} \cdot \left(\frac{\alpha-\beta}{\|\alpha-\beta\|} - \frac{\beta-\gamma}{\|\beta-\gamma\|} \right) + \beta_{x_i} \cdot \frac{\partial}{\partial x_i} \left(\frac{\alpha-\beta}{\|\alpha-\beta\|} - \frac{\beta-\gamma}{\|\beta-\gamma\|} \right) \\ &= \beta_{x_i x_i} \cdot (\boldsymbol{v}_{\alpha\beta} - \boldsymbol{v}_{\beta\gamma}) \\ &\quad + \beta_{x_i} \cdot \left(\frac{-\beta_{x_i}}{\|\alpha-\beta\|} + \frac{\boldsymbol{v}_{\alpha\beta}\cos(\psi_{x_i})}{\|\alpha-\beta\|} - \frac{\beta_{x_i}}{\|\beta-\gamma\|} + \frac{\boldsymbol{v}_{\beta\gamma}\cos(\psi_{x_i})}{\|\beta-\gamma\|} \right) \\ &= \beta_{x_i x_i} \cdot (\boldsymbol{v}_{\alpha\beta} - \boldsymbol{v}_{\beta\gamma}) + \left[\cos^2(\psi_{x_i}) - 1\right] \left(\frac{1}{\|\alpha-\beta\|} + \frac{1}{\|\beta-\gamma\|} \right). \end{aligned}$$

Finally, if z is a coordinate system at α_m, where $|m-\ell| > 1$, then

$$\frac{\partial^2(-L_k)}{\partial z_j \partial x_i} = \frac{\partial}{\partial z_j} \left(\frac{\partial}{\partial x_i}\|\alpha_{\ell-1} - \alpha_\ell\| + \frac{\partial}{\partial x_i}\|\alpha_\ell - \alpha_{\ell+1}\| \right) = 0. \qquad \square$$

Now we can use Lemma 12 to prove

Lemma 13. *For a given non-degenerate* (*i.e., eigenvalues of the corresponding Hessian at that point are all non-zero*) *generalized billiard path P, as the endpoints p and q are varied, the eigenvalues of the Hessian vary continuously.*

Proof. Let $\boldsymbol{x}$ be a coordinate system near α_k, the last reflection of P. For the other reflections, α_i, let $\boldsymbol{y}^i = (y_1^i, \ldots, y_n^i)$ be a coordinate system in a neighborhood. We may, without loss of generality, assume that the domain of $(\boldsymbol{y}^1; \ldots; \boldsymbol{y}^{n-1}; \boldsymbol{x})$ is contained in the interior of X_k.

Recall the formulae for the entries of the Hessian, given in Lemma 12. If the endpoint q is moved to q', the vector

$$\boldsymbol{v}_\ell = \frac{\alpha_k - q'}{\|\alpha_k - q'\|}$$

varies continuously with q'. The vectors $\boldsymbol{v}_{k-1} = \frac{\alpha_{k-1}-\alpha_k}{\|\alpha_{k-1}-\alpha_k\|}$, α_{k,x_j} and α_{i,y_j^i} are all constant. It follows that $\frac{\partial(-L_k^{(p,q')})}{\partial y_j^i}$ is zero and $\frac{\partial(-L_k^{(p,q')})}{\partial x_i}$ varies continuously with q'. So the gradient $\nabla(-L_k^{(p,q')})$ varies continuously with q'. It follows that there is a generalized billiard path P' whose reflections are close to the reflections of P. Moreover, P' varies continuously with q'.

In addition, the quantity $\|\alpha_k - q'\|$ and the angle ψ_{x_i} vary continuously with q'. It follows that the entries of the Hessian $H(P')$ vary continuously, and hence so do the eigenvalues. □

Lemma 13 shows that for a given embedding, $M \hookrightarrow \mathbb{R}^N$, the set of pairs $(p; q) \in \mathbb{R}^N \times \mathbb{R}^N$ such that $-L^{(p,q)}$ is a Morse function is open in $\mathbb{R}^{2N}$.

Notice that as the endpoints are moved, the eigenvalues of a critical point (i.e., a generalized billiard path) vary continuously, but the critical point itself varies as well. We say that a generalized billiard path P_0 from p_0 to q_0 is *related to* a generalized billiard path P_1 from p_1 to q_1 if there are paths $p : [0, 1] \to M$ and $q : [0, 1] \to M$, with $p(0) = p_0$, $p(1) = p_1$, $q(0) = q_0$ and $q(1) = q_1$, and for each $t \in [0, 1]$ a generalized billiard path $P(t)$ whose endpoints vary continuously from P_0 to P_1. Sometimes in moving the endpoints from (p_0, q_0) to (p_1, q_1), there will be no generalized billiard path from p_1 to q_1 related to a path P from p_0 to q_0. In this case, we say the movement *destroys* the path P.

There are two things that may prevent $-L_k$ from being a Morse function. One of these is that one or both of the endpoints may be located a focal point. Another problem occurs when there is a billiard path that has a *tangential reflection*, i.e., the angle of incidence and angle of reflection are both zero. We will find in Lemma 20 these are rare occurrences. The proof of Lemma 20 requires Lemma 15, which in turn requires

Lemma 14. *Let $M \subset \mathbb{R}^N$ be an embedded manifold, and let $\alpha, \beta \in M$ be such that the line $\overleftrightarrow{\alpha\beta}$ is tangent to M at β, but not at α. Then in any neighborhood U_α of α there is a point α' such that the line $\overleftrightarrow{\alpha'\beta}$ is tangent to M at neither α' nor β.*

Proof. Since the line $\overleftrightarrow{\alpha\beta}$ is tangent to M at β, $\overleftrightarrow{\alpha\beta}$ is contained in $T_\beta M$. It follows that $\alpha \in T_\beta M$. (We are thinking of $T_\beta M$ as a linear subspace of $\mathbb{R}^N$.)

Suppose there is a neighborhood U_α such that $\overleftrightarrow{\alpha'\beta}$ is tangent to M at β for all $\alpha' \in U_\alpha$. Then $U_\alpha \subset T_\beta M$.

If the entire neighborhood U_α is contained in $T_\beta M$, then $T_\alpha M$ coincides with $T_\beta M$. Consequently, $\overleftrightarrow{\alpha\beta} \subset T_\alpha M$, and so $\overleftrightarrow{\alpha\beta}$ is tangent to M at α, contrary to assumption. □

Lemma 15. *Let $M \subset \mathbb{R}^n$ and $p, q \in \mathbb{R}^n$. If P is a generalized billiard path from p to q that has a tangential reflection at $\alpha_{i+1} = \beta$, then for any $\varepsilon > 0$ there is a $p' \in B_\varepsilon(p)$ and $q' \in B_\varepsilon(q)$ such that either there is a non-tangential generalized billiard path from p' to q' related to P, or the movement from (p, q) to (p', q') destroys P.*

Proof. Assume β is the first tangential reflection along the path. Then $\overleftrightarrow{\alpha_i\beta}$ is not tangent to M at α_i. Fix β and move α_i along a curve $\alpha_i : [0, 1] \to M$ such that $\alpha_i(0) = \alpha_i$ and $\overleftrightarrow{\alpha_i(t)\beta}$ is not tangent to M at β for $t > 0$.

Since the generalized billiard path $(\alpha_1, \ldots, \alpha_i)$ from p to β is non-tangential, there is a $\delta > 0$ and curves $p : [0, \delta] \to M$ and $\alpha_j : [0, \delta] \to M$ for $j \in \{1, \ldots, i-1\}$ such that $(\alpha_1(t), \ldots, \alpha_i(t))$ is a (non-tangential) generalized billiard path from $p(t)$ to β.

If there are paths $\alpha_j : [0, \delta] \to M$ for $j \in \{i+2, \ldots, k\}$ and $q : [0, \delta] \to M$ such that $(\alpha_1(t), \ldots, \alpha_i(t), \beta, \alpha_{i+2}(t), \ldots, \alpha_n(t))$ is a non-tangential generalized billiard path from $p(t)$ to $q(t)$, then we can choose t_0 such that $p(t_0) \in B_\varepsilon(p)$ and $q(t_0) \in B_\varepsilon(p)$. Then set $p' = p(t_0)$ and $q' = q(t_0)$. (Note, this can be done in such a way that every non-degenerate generalized billiard path from p to q has a related path from p' to q'.)

If there are tangential reflections along the generalized billiard path $(\alpha_1(t), \ldots, \alpha_i(t), \beta, \alpha_{i+2}(t), \ldots, \alpha_n(t))$ from $p(t)$ to $q(t)$, then the above procedure must be repeated before choosing p' and q'.

Finally, if there is no such generalized billiard path from $p(t)$ to $q(t)$, then the path P has been destroyed. □

Now we turn our attention to the problem of non-tangential degenerate billiard paths. We begin by investigating the eigenvectors of the Hessian matrix $H(\alpha_1, \ldots, \alpha_k)$. We can think of an eigenvector $\boldsymbol{V}$ as a vector $\boldsymbol{v}_i$ in each of the tangent spaces $T_{\alpha_i} X_k$, $\boldsymbol{V} = (\boldsymbol{v}_1; \ldots; \boldsymbol{v}_k)$. In this case

$$\begin{bmatrix} A_1 & K_1 & & & & & & \\ K_1 & A_2 & K_2 & & & & 0 & \\ & K_2 & \ddots & & & & & \\ & & & \ddots & K_{i-1} & & & \\ & & & K_{i-1} & A_i & K_i & & \\ & & & & K_i & \ddots & & \\ & & & & & & \ddots & K_{k-1} \\ & 0 & & & & & K_{k-1} & A_k \end{bmatrix} \begin{pmatrix} \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_k \end{pmatrix} = \lambda \begin{pmatrix} A_1\boldsymbol{v}_1 + K_1\boldsymbol{v}_2 \\ \vdots \\ K_{k-1}\boldsymbol{v}_{k-1} + A_k\boldsymbol{v}_k \end{pmatrix},$$

where λ is the corresponding eigenvalue.

From Lemma 12, the matrix K_{k-1} can be written

$$K_{k-1} = \frac{1}{\|\beta - \gamma\|}\left(\beta_{x_i} \cdot \gamma_{y_j} - \cos(\psi_{x_i})\cos(\psi_{y_j})\right)_{i,j} = \frac{K' - K''}{\|\alpha - \beta\|},$$

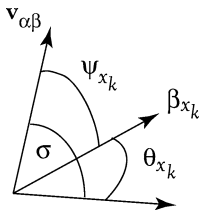

Fig. 9. The three angles σ, ψ_{x_ℓ}, and θ_{x_ℓ}.

where $\beta = \alpha_{k-1}$ and $\gamma = \alpha_k$

Lemma 16. *If $H(\alpha_1, \ldots, \alpha_k)$ has an eigenvector $\boldsymbol{V} = (\boldsymbol{v}_1; \ldots; \boldsymbol{v}_{k-1}; 0)$ with eigenvalue zero, then $(\alpha_1, \ldots, \alpha_k)$ is a sequence with a tangential reflection.*

Proof. In this case, $K_{k-1}\boldsymbol{v}_{k-1} + A_k\boldsymbol{v}_k = K_{k-1}\boldsymbol{v}_{k-1} = 0$. It is sufficient to show that if $K_{k-1}\boldsymbol{v} = 0$ for any $\boldsymbol{v} \in T_{\alpha_{k-1}}M$, then the sequence has a tangential reflection.

Let us investigate $K''\boldsymbol{v}$ first. The ℓth component is given by

$$\begin{aligned}\big(K''\boldsymbol{v}\big)_\ell &= \sum_j v_j \cos(\psi_{x_\ell})\cos(\psi_{y_j}) = \cos(\psi_{x_\ell})\sum_j v_j \cos(\psi_{y_j}) \\ &= (\beta_{x_\ell}\cdot \boldsymbol{v}_{\alpha\beta})\sum_j v_j(\alpha_{y_j}\cdot \boldsymbol{v}_{\alpha\beta}) = (\beta_{x_\ell}\cdot \boldsymbol{v}_{\beta\gamma})\left(\sum_j v_j\gamma_{y_j}\right)\cdot \boldsymbol{v}_{\beta\gamma} \\ &= (\beta_{x_\ell}\cdot \boldsymbol{v}_{\beta\gamma})(\boldsymbol{v}\cdot \boldsymbol{v}_{\beta\gamma}).\end{aligned}$$

We also have

$$\big(K'\boldsymbol{v}\big)_\ell = \sum_j v_j\beta_{x_\ell}\cdot\gamma_{y_j} = \beta_{x_\ell}\cdot\sum_j v_j\gamma_{y_j} = \beta_{x_\ell}\cdot\boldsymbol{v}.$$

When is $\beta_{x_\ell}\cdot\boldsymbol{v} = (\beta_{x_\ell}\cdot\boldsymbol{v}_{\beta\gamma})(\boldsymbol{v}\cdot\boldsymbol{v}_{\beta\gamma})$? Let θ_{x_ℓ} be the angle between β_{x_ℓ} and $\boldsymbol{v}$, and let σ be the angle between $\boldsymbol{v}_{\beta\gamma}$ and $\boldsymbol{v}$. The statement the reduces to

$$\cos(\theta_{x_\ell}) = \cos(\psi_{x_\ell})\cos(\sigma)$$

(since the vectors in question are all unit vectors).

Let π_β denote the projection onto $T_\beta M$. Then the following identities hold.

$$\pi_\beta(\boldsymbol{v}_{\beta\gamma}) = \sum_\ell \cos(\psi_{x_\ell})\beta_{x_\ell}, \qquad \pi_\beta(\boldsymbol{v}) = \sum_\ell \cos(\theta_{x_\ell})\beta_{x_\ell}.$$

From this we see

$$\pi_\beta(\boldsymbol{v}) = \sum_\ell \cos(\psi_{x_\ell})\cos(\sigma)\beta_{x_\ell} = \cos(\sigma)\sum_\ell \cos(\psi_{x_\ell})\beta_{x_\ell} = \cos(\sigma)\pi_\beta(\boldsymbol{v}_{\beta\gamma}).$$

We can write $\boldsymbol{v} = \boldsymbol{v}_{\beta\gamma}\cos(\sigma) + (\boldsymbol{v}_{\beta\gamma})^\perp\sin(\sigma)$ for some $(\boldsymbol{v}_{\beta\gamma})^\perp$ orthogonal to $\boldsymbol{v}$. It then follows that either $\pi_\beta(\boldsymbol{v}_{\beta\gamma}{}^\perp) = 0$ or $\sin(\sigma) = 0$. In the first case the conclusion is that $\boldsymbol{v}_{\beta\gamma} \in T_\beta M$, and so the path is tangent to M at the point β. The second condition implies that $\boldsymbol{v}_{\beta\gamma} = \pm\boldsymbol{v}$, and so $\boldsymbol{v}_{\beta\gamma} \in T_\gamma M$. Here the path is tangent to M at γ. □

Lemma 17. *For each non-tangential generalized billiard path P, as the length ℓ from the last reflection to q is increased, the eigenvalues of the Hessian $H(P_\ell)$ increase strictly monotonically.*

Proof. Recall from the proof of Lemma 12 that the Hessian can be written as

$$H(P_\ell) = N - \frac{1}{\ell}\begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & C \end{bmatrix},$$

where N is a constant matrix and

$$C = \left(\delta_{ij} - \cos(\psi_{x_i})\cos(\psi_{x_j})\right)_{i,j}.$$

A standard result of linear operator theory tells us that the ith eigenvalue (in increasing order) is given by

$$\begin{aligned}\lambda_i &= \sup_{\{\boldsymbol{X}_1,\dots,\boldsymbol{X}_{i-1}\}} \inf_{\boldsymbol{V}\in\{\boldsymbol{X}_1,\dots,\boldsymbol{X}_{i-1}\}^\perp} \left(\frac{\langle H\boldsymbol{V},\boldsymbol{V}\rangle}{\langle \boldsymbol{V},\boldsymbol{V}\rangle}\right) \\ &= \sup_{\{\boldsymbol{X}_1,\dots,\boldsymbol{X}_{i-1}\}} \inf_{\boldsymbol{V}\in\{\boldsymbol{X}_1,\dots,\boldsymbol{X}_{i-1}\}^\perp} \left(\frac{\langle N\boldsymbol{V},\boldsymbol{V}\rangle}{\langle \boldsymbol{V},\boldsymbol{V}\rangle} - \frac{1}{\ell}\frac{\langle C\boldsymbol{v}_k,\boldsymbol{v}_k\rangle}{\langle \boldsymbol{v}_k,\boldsymbol{v}_k\rangle}\right),\end{aligned}$$

where $\{\boldsymbol{X}_1,\dots,\boldsymbol{X}_{i-1}\}$ are taken to be linearly independent, and $V = (\boldsymbol{v}_1;\dots;\boldsymbol{v}_k)$.

The value λ_i can be realized by choosing $\boldsymbol{X}_j$ to be an eigenvector corresponding to λ_j and $\boldsymbol{V}$ to be an eigenvector corresponding to λ_i. Because of this, we may restrict the inf to those vectors $V \in \{\boldsymbol{X}_1,\dots,\boldsymbol{X}_{i-1}\}^\perp$ with $\boldsymbol{v}_k \neq 0$. (All eigenvectors are of this form.) We may also restrict our attention to those vectors with $\|\boldsymbol{v}_k\| = 1$.

If $\boldsymbol{v}$ is a unit vector, then a calculation shows

$$\langle C\boldsymbol{v},\boldsymbol{v}\rangle = \langle \boldsymbol{v},\boldsymbol{v}\rangle - \boldsymbol{v}^{\mathrm{T}}\cdot\left[\cos(\psi_{x_i})\cos(\psi_{x_j})\right]_{i,j}\cdot\boldsymbol{v} = 1 - (\boldsymbol{v}\cdot\boldsymbol{v}_{\alpha_k q})^2.$$

Thus $\langle C\boldsymbol{v},\boldsymbol{v}\rangle$ is positive unless $\boldsymbol{v} = \pm\boldsymbol{v}_{\alpha_k q}$. This cannot be the case, though, since P is a non-tangential reflection. It follows that when ℓ increases, the value of

$$\frac{\langle H\boldsymbol{V},\boldsymbol{V}\rangle}{\langle \boldsymbol{V},\boldsymbol{V}\rangle}$$

increases continuously for every vector $\boldsymbol{V}$ with $\boldsymbol{v}_k \neq 0$. As a consequence of this, we see that λ_i must increase continuously as ℓ increases. □

Lemma 18. *Given two points $p, q \in \mathbb{R}^N$, a non-tangential generalized billiard path P connecting them and an $\varepsilon > 0$, there exists a $q' \in B_\varepsilon(q)$ such that there is a non-degenerate generalized billiard path P' from p to q' related to P. Moreover, q' can be chosen in such a way that each non-degenerate generalized billiard path from p to q has a related non-degenerate generalized billiard path from p to q'.*

Proof. Let $\boldsymbol{v}_1$ be the unit vector pointing from the last reflection of P to q. As q' is moved in the direction of $\boldsymbol{v}_1$ from q, all the eigenvalues of P and all the non-degenerate paths vary continuously (Lemmas 13 and 17). Choose a $0 < \delta < \varepsilon$ so that all of the non-zero

eigenvalues (of P and the non-degenerate generalized billiard paths) are bounded away from zero between q and $q + \delta \boldsymbol{v}_1$.

By Lemma 17, the eigenvalues of the path P_1 will increase monotonically. Hence all of the zero eigenvalues will have increased to positive values. Since no new zero eigenvalues have been created, we can set $q' = q + \delta \boldsymbol{v}_1$. □

Lemma 19. *For any embedded manifold $M \hookrightarrow \mathbb{R}^N$, points $p, q \in \mathbb{R}^N$ and $\varepsilon > 0$, there are points $p' \in B_\varepsilon(p)$ and $q' \in B_\varepsilon(q)$ such that $-L_k^{(p',q')}$ is a Morse function.*

Proof. Choose a degenerate generalized billiard path P. If it has a tangential reflection, use Lemma 15 to find $p_1 \in B_{\varepsilon/4}(p)$ and $q_1 \in B_{\varepsilon/4}(q)$ such that there is either a related non-tangential generalized billiard path P_1 from p_1 to q_1, or no generalized billiard path related to P. This can be done without destroying any non-degenerate generalized billiard paths.

Next use Lemma 18 to choose $p_2 \in B_{\varepsilon/4}(p_1)$ and $q_2 \in B_{\varepsilon/4}(q_1)$ such that there is a non-degenerate generalized billiard path P_2 from p_2 to q_2 related to P_1. Again, this can be done without destroying any non-degenerate generalized billiard paths.

Repeat these two steps as often as needed, each time choosing p_{2j-1} and q_{2j-1} within $\varepsilon/2j + 1$ of p_{2j-2} and q_{2j-2}, then choosing p_{2j} and q_{2j} within $\varepsilon/2j + 1$ of p_{2j-1} and q_{2j-1}. This procedure must terminate after a finite number of steps, otherwise we have constructed an infinite sequence of non-degenerate (and hence isolated) critical points in the compact manifold M^k. □

Lemma 20. *Given an embedding of $M \hookrightarrow \mathbb{R}^N$, the set of points $(p; q) \in \mathbb{R}^N \times \mathbb{R}^N$ such that $-L_k^{(p,q)}$ is a Morse function is open and dense.*

Proof. Lemma 13 shows that the set is open. Lemma 19 shows the set is dense. □

3.5. Application of the Morse inequalities

In this section, we finally apply the results of Section 2 to the case of $-L_k : X_k \to \mathbb{R}$.

Theorem 21. *Suppose $M \hookrightarrow \mathbb{R}^N$ is a smooth embedding of an n-manifold, and $p, q \in \mathbb{R}^N$. Then for every $\varepsilon > 0$, there is a $p' \in B_\varepsilon(p)$ and a $q' \in B_\varepsilon(q)$ such that if N_k is the number of billiard paths with k reflections connecting p' to q'. Then*

$$N_k \geqslant \sum_{i=0}^{kn} b_i(X_k).$$

Proof. Choose p' and q' so that $-L_k$ is a Morse function. Then

$$N_k = \sum_{i=0}^{kn} m_i(-L_k).$$

Since $m_i(-L_k) \geqslant b_i(X_k)$, the result follows. □

This shows that the more complicated the topology of the path space X_k, the greater the number of generalized paths there must be. In the next two sections the Betti numbers of X_k will be related to the Betti numbers of M. This will allow an estimate of the number of billiard paths to be stated in terms of the topology of the underlying manifold M.

4. Morse Theory for the length function on sequences

4.1. $M \times \cdots \times M$ as a stratified space

The space M^k is naturally a smooth manifold. We impose on this smooth space a stratified structure, by treating all of the various diagonals as separate strata.

A point P in M^k is an ordered k-tuple of points in M, $P = (\alpha_1, \ldots, \alpha_k)$. Define $F_j = F_j(M^k) = \{(\alpha_1, \ldots, \alpha_k)\colon\ \alpha_i \neq \alpha_{i+1}$ for exactly $j+1$ choices of $i \in \{0, \ldots, k\}\}$. Here, as usual, $\alpha_0 = p$ and $\alpha_{k+1} = q$. F_j is the set of sequences with j distinct "reflections". The components of F_j will be the strata of M^k.

4.2. A continuous flow on M^k

In classical Morse theory, one considers the negative gradient flow of a function. The function $L_k : M^k \to \mathbb{R}$ is not differentiable everywhere, but we will show that its restriction to any stratum is, in general, a Morse function. Viewing M^k as a stratified space, we can define a flow that will substitute for a negative gradient flow.

Lemma 22. *For $j = 1, \ldots, k$, if $-L_j^{(p,q)} : X_j \to \mathbb{R}$ is a Morse function (in the sense of Section 2) then $L_k^{(p,q)}|_{F_j} : F_j \to \mathbb{R}$ is a (classical) Morse function.*

Proof. F_j is the set of all configurations of k points in M such that the maximal subset of distinct points has magnitude j. X_k is the (closure of the) set of configurations of j distinct points in M. For each connected component H of F_j, there is a diffeomorphism $F : H \to \mathrm{Int}(X_k)$. (Provided that X_k is connected, otherwise H maps onto the interior of a connected component of X_k.) If $(\alpha_1, \ldots, \alpha_k) \in H$ and $\alpha_{i_1}, \ldots, \alpha_{i_j}$ are distinct, then $f(\alpha_1, \ldots, \alpha_k) = (\alpha_{i_1}, \ldots, \alpha_{i_j}) \in X_k$.

Now, if $P \in H$, then $L_k(P) = L_k(f(P))$. Since $-L_k$ is a Morse function on the interior of X_k (a smooth manifold), L_k is a Morse function as well. □

Lemma 23. *The set of endpoints p and q such that $L_k^{(p,q)}|_{F_j} : F_j \to \mathbb{R}$ is a Morse function for every $j \in \{1, \ldots, k\}$ is open and dense in $\mathbb{R}^{2N}$.*

Proof. For each j, the set of endpoints such that $-L_j : X_j \to \mathbb{R}$ is open and dense. Consequently, the intersection of all k of these sets is open and dense. □

Now we will define a vector field G^+ on M^k as follows. For $P \in F_j$, set $G^+(P) = -\nabla(L_k|_{F_j(P)})$. That is, $G^+(P)$ is the negative gradient of L_k restricted to the stratum containing P.

Now, if H is a stratum of M^k, the vector field G^+ induces a flow $\vartheta_H : U \to H$, where U is an open neighborhood of $\{0\} \times H \subset \mathbb{R} \times H$. Assume that U is the largest neighborhood on which such a flow may be defined. For $P \in H$, let $t_P = \sup\{t \mid (t, P) \in U\}$.

We can build a flow $\vartheta : [0, \infty) \times M^k \to M^k$ on all of M^k from these individual flows in the following way: define for $P \in H$

$$\vartheta(t, P) = \begin{cases} \vartheta_H(t, P) & t < t_P, \\ \vartheta_K\big(t - t_P, \lim_{s \to t_P} \vartheta_H(s, P)\big) & t \geqslant t_P, \end{cases}$$

where K is the stratum containing $\lim_{s \to t_P} \vartheta_H(s, P)$. The idea is that the point flows until it reaches the boundary of the initial stratum, and hence reaches a lower-dimensional stratum, then continues in the lower-dimensional stratum. Note that this is a recursive definition, but since when the flow moves from one stratum to another the dimension of the stratum always decreases, only finitely many iterations are needed.

Lemma 24. *The flow $\vartheta : [0, \infty) \times M^k \to M^k$ is continuous.*

Proof. Certainly $\vartheta|_{F_1}$ is continuous, since it is just the negative gradient flow of a smooth function. Now, suppose that ϑ restricted to the union $\bigcup_{i=1}^{j-1} F_i$ is continuous. We will show that

$$\vartheta|_{\bigcup_{i=1}^{j} F_i}$$

is continuous.

Let $U = \{(t, P) \in [0, \infty) \times F_j \mid t < t_P\}$. Then $\vartheta|_U$ is continuous. We must show that ϑ is still continuous when $t \geqslant t_P$. It follows from the continuity of ϑ on U that the map $P \mapsto t_P$ is continuous.

Let $E_j = \{P \in F_j \mid t_P < \infty\}$. Since moving two nearby reflections closer decreases the length of the sequence, E_j contains a neighborhood of ∂F_j. We can define the function $f : E_j \to \bigcup_{i=1}^{j-1} F_i(M^k)$ by $f(P) = \lim_{t \to t_P} \vartheta(t, P)$. Then f is a continuous function, again due to the continuity of $\vartheta|_U$.

Finally, define $\tau : \mathbb{R} \times E_j \to \mathbb{R}$ by $\tau(t, P) = t - t_P$. Then τ is continuous as well. Now we can write $\vartheta(t, P)$ as a composition of continuous functions:

$$\vartheta(t, P) = \vartheta\big(\tau(t, P), f(P)\big).$$

By induction, it follows that ϑ is continuous on M^k. □

4.3. The Morse theorems

Before we may proceed to prove the Morse theorems, we need to make a

Definition 10. An essential critical point of $L_k : M^k \to \mathbb{R}$ is a point $P \in M^k$ such that $G^+(P) = 0$.

By the definition of G^+, this condition is equivalent to the requirement that P is a critical point of one of the functions $L_k|_{F_j} : F_j \to \mathbb{R}$. If $P \in F_j$ is an essential critical point, then there is a coordinate system on M^k such that

$$L_k(x_1,\dots,x_{kn}) = L_k(P) - x_1^2 - \cdots - x_\lambda^2 + x_{\lambda+1}^2 + \cdots + x_{jn} + |x_{jn+1}| + \cdots + |x_{kn}|.$$

Here $(x_1,\dots,x_{jn})$ represents a coordinate system on a neighborhood of P in F_j. We say that λ is the *index* of P.

We will use the following notation: $(M^k)_a = L_k^{-1}((-\infty,a])$.

Lemma 25. *If there are no essential critical points in $L_k^{-1}([a,b])$, then for each point P in $(M^k)_b$ there is a time t such that $\vartheta(t,P)\in(M^k)_a$.*

Proof. The set $L_k^{-1}([a,b])$ is compact, and contains no essential critical points. It follows that

$$\mu = \min_i \inf_{P\in L_k^{-1}([a,b])} \left\|\nabla(L_k|_{F_i})\right\| > 0.$$

Since the directional derivative $(-\nabla(L_k|_{F_i}))[L_k]$ is given by

$$\begin{aligned}\big(-\nabla(L_k|_{F_i})\big)[L_k] &= -\left(\sum_{\ell=1}^{jn}\frac{\partial L_k}{\partial x_\ell}\boldsymbol{e}_\ell\right)[L_k] = -\sum_\ell \frac{\partial L_k}{\partial x_\ell}(\boldsymbol{e}_\ell)[L_k]\\ &= -\sum_\ell \frac{\partial L_k}{\partial x_\ell}\frac{\partial L_k}{\partial x_\ell} = -\left\|\nabla(L_k|_{F_i})\right\|,\end{aligned}$$

and $\frac{\partial}{\partial t}\vartheta = (-\nabla(L_k|_{F_i}))[L_k]$, it follows that along flow lines, the value of L_k is decreasing at a rate bounded away from zero. So for $t > \frac{b-a}{\mu}$, $\vartheta(P,t)\in(M^k)_a$ for all $P\in(M^k)_b$. □

Theorem 26. *If $a<b$ and $L_k^{-1}([a,b])$ contains no essential critical points, then $(M^k)_a$ is a deformation retract of $(M^k)_b$, so the inclusion map $(M^k)_a \hookrightarrow (M^k)_b$ is a homotopy equivalence.*

Proof. Since there are no essential critical points in $L_k^{-1}([a,b])$, and the value of L_k decreases along the flow lines of ϑ, for each point $P\in(M^k)_b$, there is a time t such that $\vartheta(t,p)\in(M^k)_a$. Let $s_P = \inf\{t\in\mathbb{R}_+\colon \vartheta(t,P)\in(M^k)_a\}$.

Now we can define a homotopy $H:(M^k)_b\times[0,1]\to(M^k)_a$ by

$$H(P,s) = \begin{cases}\vartheta\left(\frac{s}{1-s},P\right), & \frac{s}{1-s}\leqslant s_P,\\ \vartheta(s_P,P), & \frac{s}{1-s}\geqslant s_P.\end{cases}$$ □

Theorem 27. *Let P be an essential critical point of L_k with index λ. Set $L_k(P)=c$. Suppose that for some $\varepsilon>0$, $L_k^{-1}([c-\varepsilon,c+\varepsilon])$ contains no essential critical points other than P. Then $(M^k)_{c+\varepsilon}$ is homotopy equivalent to $(M^k)_{c-\varepsilon}$ with a λ-cell attached.*

Proof. Choose a coordinate system $\boldsymbol{x}:U_P\to\mathbb{R}^{kn}$ in which we can write

$$f = f(P) - x_1^2 - \cdots - x_\lambda^2 + x_{\lambda+1}^2 + \cdots + x_{jn}^2 + |x_{jn+1}| + \cdots + |x_{kn}|.$$

Then choose $\varepsilon > 0$ sufficiently small so that $L_k^{-1}[c-\varepsilon, c+\varepsilon]$ contains no essential critical points other than P, and the image $\boldsymbol{x}(U_P)$ contains the closed 'ball'

$$\left\{(x_1,\dots,x_n)\colon \sum_{i=1}^{n(k-j)} x_i^2 + \sum_{i=n(k-j)+1}^{kn} |x_i| \leqslant 2\varepsilon\right\}.$$

We proceed as in Section 2.7, defining μ, ξ and η as we do there. Now, though, we must define

$$\zeta = |x_{n-j+1}| + \cdots + |x_n|.$$

Then we can write $L_k = c - \xi + \eta + \zeta$ and $\Gamma = L_k - \mu(\xi + 2\eta + 2\zeta)$. These functions play the roles of f and F respectively in Section 2.7. As before, it follows that the essential critical points of Γ and L_k are identical. Likewise, $\Gamma^{-1}(-\infty, c+\varepsilon) = L_k^{-1}(-\infty, c+\varepsilon)$ and $\Gamma^{-1}(-\infty, c-\varepsilon]$ is a deformation retract of $(M^k)_{c+\varepsilon}$.

It is now necessary only to show that $(M^k)_{c-\varepsilon} \cup e^\lambda$ is a deformation retract of $\Gamma^{-1}(-\infty, c-\varepsilon]$. For each $t \in [0,1]$ we must define a map $r_t : \Gamma^{-1}(-\infty, c-\varepsilon] \to (M^k)_{c-\varepsilon} \cup H$ as follows:

Case 1. If $Q \in (M^k)_{c-\varepsilon}$, set $r_t(Q) = Q$ for all t.

Case 2. If $Q \in \Gamma^{-1}(-\infty, c-\varepsilon]$ but $Q \notin (M^k)_{c-\varepsilon}$ and $\xi(Q) < \varepsilon$, then set

$$r_t(x_1,\dots,x_n) = \big(x_1,\dots,x_\lambda,(1-t)x_{\lambda+1},\dots,(1-t)x_n\big).$$

Case 3. If $\varepsilon \leqslant \xi(Q) \leqslant \eta(Q) + \zeta(Q) + \varepsilon$, then define r_t by

$$r_t(x_1,\dots,x_n) = (x_1,\dots,x_\lambda, s_t x_{\lambda+1},\dots,s_t x_n),$$

where

$$s_t = (1-t) + t\left[\frac{\xi-\varepsilon}{\eta+\zeta}\right]^{1/2}.$$

Then r_0 is the identity map, and $r_1 : \Gamma^{-1}(-\infty, c-\varepsilon] \to (M^k)_{c-\varepsilon} \cup e^\lambda$. Note that this Case 3 agrees with Case 1 when $\xi = \varepsilon$ and with Case 2 when $\xi - \eta - \zeta = \varepsilon$. Continuity follows from the proof of Theorem 8. Thus r provides a deformation retraction of $(M^k)_{c-\varepsilon} \cup H$ to $(M^k)_{c-\varepsilon} \cup e^\lambda$. This concludes the proof of Theorem 27. □

Theorems 26 and 27 may be used to show that M^k is homotopy equivalent to a CW-complex having one λ cell for each essential critical point of $L_k : M^k \to \mathbb{R}$ with index λ. (See [7].) This in turn allows us to deduce the Morse Inequalities, both strong:

$$\sum_{i=0}^{\ell} (-1)^{\ell+i} b_i\big(M^k\big) \leqslant \sum_{i=0}^{\ell} (-1)^{\ell+i} m_i(L_k),$$

and weak:

$$m_\ell(L_k) \geqslant b_\ell\big(M^k\big)$$

or each $0 \leqslant \ell \leqslant kn$.

5. Generalized billiard paths revisited

5.1. A comparison of essential critical points in X_k and M^k

We can now investigate the relationship between critical points of $-L_k : X_k \to \mathbb{R}$ and $L_k : M^k \to \mathbb{R}$. If a stratum H is in F_j, then its dimension is nj. Consequently, every essential critical point in H has index at most nj. Consequently, we get the following

Lemma 28. *Any essential critical point of $L_k : M^k \to \mathbb{R}$ with index $\lambda > n(k-1)$ is in F_k.*

For any path $P \in F_k$, the preimage $g^{-1}(P)$ consists of a single point, P'. It is easy to see that P is an essential critical point of L_k if and only if P' is an essential critical point of $-L_k$, since

$$-\nabla(L_k) = 0 \iff -\nabla(-L_k) = 0.$$

Since the function $-L_k$ decreases along paths moving away from ∂X_k, it follows that all of the essential critical points of $-L_k : X_k \to \mathbb{R}$ lie in the interior of X_k. As a result of this we have the following

Lemma 29. *The essential critical points of $-L_k : X_k \to \mathbb{R}$ are in one-to-one correspondence with the essential critical points of $L_k : M^k \to \mathbb{R}$ that lie in F_k.*

5.2. Application of the Morse inequalities

So we see that counting the number of generalized billiard paths with exactly k reflections is equivalent to counting the number of essential critical points in F_k. There are at least as many of these essential critical points as there are essential critical points with index greater than nk.

Theorem 30. *The number of generalized billiard paths connecting p to q in the vicinity of a manifold M satisfies*

$$N_k^{(p,q)} \geqslant \sum_{j=0}^{n-1} \sum_{i_1+\cdots+i_k=j} b_{i_1}(M) \cdots b_{i_k}(M).$$

Proof. If $m_j(L_k)$ denotes the number of essential critical points of $L_k : M^k \to \mathbb{R}$, then

$$N_k^{(p,q)} \geqslant \sum_{j=0}^{nk} m_j(L_k) \geqslant \sum_{j=n(k-1)+1}^{nk} m_j(L_k) \geqslant \sum_{j=n(k-1)+1}^{nk} b_j\big(M^k\big).$$

Since M^k is a manifold, we can use Poincaré duality to say $b_i(M^k) = b_{nk-i}(M^k)$. Then we see

$$N_k^{(p,q)} \geqslant \sum_{j=0}^{n-1} b_j\big(M^k\big).$$

Using the Künneth Theorem, it can be shown that

$$b_j\left(M^k\right) = \sum_{i_1+\cdots+i_k=j} b_{i_1}(M)\cdots b_{i_k}(M).$$

Now we can finally write

$$N_k^{(p,q)} \geqslant \sum_{j=0}^{n-1} \sum_{i_1+\cdots+i_k=j} b_{i_1}(M)\cdots b_{i_k}(M)$$

proving the theorem. □

Two things are evident from this expression. First, the more complicated the topology of M, the more generalized billiard paths there will be. The second is that as the number of reflections k increases, the number of generalized billiard paths with k reflections increases, and rather quickly.

Acknowledgements

I would like to thank Robin Forman and my reviewers and editors for their generous advice and assistance in giving this article its final form.

This article commemorates the birth of my son, David Christian Handron, born December 27, 2001.

References

[1] R. Bott, Morse theory indomitable, in: Proceedings from a Conference in Honor of R. Thom, Paris, 1988, pp. 99–114.
[2] D. Braess, Morse-Theorie für berandete Mannigfaltigkeiten, Math. Ann. 208 (1974) 133–148.
[3] M. Farber, S. Tabachnikov, Topology of cyclic configuration spaces and periodic trajectories of multi-dimensionsl billiards, http://xxx.lanl.gov/abs/math/9911226.
[4] W. Fulton, R. MacPherson, A compactification of configuration spaces, Ann. of Math. 139 (1994) 183–225.
[5] M. Goresky, R. MacPherson, Stratified Morse Theory, in: Ergeb. Math. Grenzgeb., Vol. 14, Springer-Verlag, New York, 1988.
[6] H.A. Hamm, On stratified Morse theory, Topology 38 (2) (1999) 427–438.
[7] J. Milnor, Morse Theory, in: Ann. Math. Stud., Vol. 51, Princeton University Press, Princeton, NJ, 1969.
[8] M. Morse, Relations between the critical points of a real function of n independent variables, Trans. Amer. Math. Soc. 27 (1925) 345–396.
[9] R. Thom, Sur une partition en cellules assodiée à une fonction sur une variété, C. R. Acad. Sci. Paris 228 (1949) 661–692.
[10] S.A. Vakhrameev, Morse lemmas for smooth functions on manifolds with corners, Dynamical Systems, Vol. 8, J. Math. Sci. 100 (2000) 2428–2445.