

AG
T

*Algebraic & Geometric
Topology*

Volume 23 (2023)

**A connection between cut locus, Thom space
and Morse–Bott functions**

SOMNATH BASU

SACHCHIDANAND PRASAD



A connection between cut locus, Thom space and Morse–Bott functions

SOMNATH BASU
SACHCHIDANAND PRASAD

Associated to every closed, embedded submanifold N in a connected Riemannian manifold M , there is the distance function d_N which measures the distance of a point in M from N . We analyze the square of this function and show that it is Morse–Bott on the complement of the cut locus $\text{Cu}(N)$ of N provided M is complete. Moreover, the gradient flow lines provide a deformation retraction of $M - \text{Cu}(N)$ to N . If M is a closed manifold, then we prove that the Thom space of the normal bundle of N is homeomorphic to $M/\text{Cu}(N)$. We also discuss several interesting results which are either applications of these or related observations regarding the theory of cut locus. These results include, but are not limited to, a computation of the local homology of singular matrices, a classification of the homotopy type of the cut locus of a homology sphere inside a sphere, a deformation of the indefinite unitary group $U(p, q)$ to $U(p) \times U(q)$ and a geometric deformation of $\text{GL}(n, \mathbb{R})$ to $O(n, \mathbb{R})$ which is different from the Gram–Schmidt retraction.

53B21, 53C22, 55P10; 32B20, 57R19, 58C05

1. Introduction	4186
2. Preliminaries	4189
3. Main results	4198
4. Applications to Lie groups	4218
Appendix A. The continuity of the map (3-2)	4227
Appendix B. Derivative of the square root map	4229
References	4232

1 Introduction

On a Riemannian manifold M , the distance function $d_N(\cdot) := d(N, \cdot)$ from a closed subset N is fundamental in the study of variational problems. For instance, the viscosity solution of the Hamilton–Jacobi equation is given by the flow of the gradient vector of the distance function d_N when N is the smooth boundary of a relatively compact domain in manifolds; see Li and Nirenberg [17] and Mantegazza and Menzucci [18]. Although the distance function d_N is not differentiable at N , squaring the function removes this issue. Associated to N and the distance function d_N is a set $\text{Cu}(N)$, the cut locus of N in M . The cut locus of a point, a notion initiated by Poincaré [23], has been extensively studied (see Kobayashi [16] for a survey as well as Buchner [4], Myers [20], Sakai [26] and Wolter [29]). There has been work on the structure of the cut locus of submanifolds. One may refer to the works of Hebda [9; 10], Sabau and Tanaka [25] and Singh [28]. Suitable simple examples indicate that $M - \text{Cu}(N)$ topologically deforms to N . One of our main results is the following:

Theorem A (Theorem 3.32) *Let N be a closed embedded submanifold of a complete Riemannian manifold M and $d_N : M \rightarrow \mathbb{R}$ denote the distance function with respect to N . If $f = d_N^2$, then its restriction to $M - \text{Cu}(N)$ is a Morse–Bott function, with N as the critical submanifold. Moreover, $M - \text{Cu}(N)$ deforms to N via the gradient flow of f .*

It is observed that this deformation takes infinite time. To obtain a strong deformation retract, one reparametrizes the flow lines to be defined over $[0, 1]$. It can be shown (Lemma 3.18) that the cut locus $\text{Cu}(N)$ is a strong deformation retract of $M - N$. A primary motivation for Theorem A came from understanding the cut locus of $N = O(n, \mathbb{R})$ inside $M = M(n, \mathbb{R})$, equipped with the Euclidean metric. We show in Section 2.2 that the cut locus is the set Sing of singular matrices and the deformation of its complement is not the Gram–Schmidt deformation but rather the deformation obtained from the polar decomposition, ie $A \in \text{GL}(n, \mathbb{R})$ deforms to $A\sqrt{A^T A}^{-1}$. Combining this with a result of Hebda [9, Theorem 1.4], we are able to compute the local homology of Sing (see Lemma 2.15 and Corollary 2.16).

Theorem B *For $A \in M(n, \mathbb{R})$,*

$$H_{n^2-1-i}(\text{Sing}, \text{Sing} - A) \cong \tilde{H}^i(O(n-k, \mathbb{R})),$$

where $A \in \text{Sing}$ has rank $k < n$.

When the cut locus is empty, we deduce that M is diffeomorphic to the normal bundle ν of N in M . In particular, M deforms to N . Among applications, we discuss two families of examples. We reprove the known fact that $GL(n, R)$ deforms to $O(n, \mathbb{R})$ for any choice of left-invariant metric on $GL(n, \mathbb{R})$ which is right- $O(n, \mathbb{R})$ -invariant. However, this deformation is not obtained topologically but by Morse–Bott flows. For a natural choice of such a metric, this deformation (4-2) is not the Gram–Schmidt deformation but one obtained from the polar decomposition. We also consider $U(p, q)$, the group preserving the indefinite form of signature (p, q) on \mathbb{C}^n . We show (Theorem 4.6) that $U(p, q)$ deforms to $U(p) \times U(q)$ for the left-invariant metric given by $\langle X, Y \rangle := \text{tr}(X^*Y)$. In particular, we show that the exponential map is surjective for $U(p, q)$ (Corollary 4.8). To our knowledge, this method is different from the standard proof.

For a Riemannian manifold we have the exponential map at $p \in M$, $\exp_p: T_p M \rightarrow M$. Let ν denote the normal bundle of N in M . We will modify the exponential map (see Section 3.2) to define the *rescaled exponential* $\widetilde{\exp}: D(\nu) \rightarrow M$, the domain of which is the unit disk bundle of ν . The main result (Theorem 3.16) here is the observation that there is a connection between the cut locus $\text{Cu}(N)$ and Thom space $\text{Th}(\nu) := D(\nu)/S(\nu)$ of ν .

Theorem C *Let N be an embedded submanifold inside a closed, connected Riemannian manifold M . If ν denotes the normal bundle of N in M , then there is a homeomorphism*

$$\widetilde{\exp}: D(\nu)/S(\nu) \xrightarrow{\cong} M/\text{Cu}(N).$$

This immediately leads to a long exact sequence in homology (see (3-6))

$$\dots \rightarrow H_j(\text{Cu}(N)) \xrightarrow{i_*} H_j(M) \xrightarrow{q} \widetilde{H}_j(\text{Th}(\nu)) \xrightarrow{\partial} H_{j-1}(\text{Cu}(N)) \rightarrow \dots$$

This is a useful tool in characterizing the homotopy type of the cut locus. We list a few applications and related results.

Theorem D *Let N be a homology k -sphere embedded in a Riemannian manifold M^d homeomorphic to S^d .*

- (1) *If $d \geq k + 3$, then $\text{Cu}(N)$ is homotopy equivalent to S^{d-k-1} . Moreover, if M and N are real analytic and the embedding is real analytic, then $\text{Cu}(N)$ is a simplicial complex of dimension at most $d - 1$.*
- (2) *If $d = k + 2$, then $\text{Cu}(N)$ has the homology of S^1 . There exist homology 3-spheres in S^5 for which $\text{Cu}(N) \simeq S^1$. However, for nontrivial knots K in S^3 , the cut locus is not homotopy equivalent to S^1 .*

The above results are a combination of Theorems 3.24 and 3.9 and Example 3.29. In general, the structure of the cut locus may be wild (see Gluck and Singer [7], Itoh and Sabau [13] and Itoh and Vîlcu [15]). Myers [20] had shown that, if M is a real analytic sphere, then $\text{Cu}(p)$ is a finite tree each of whose edges is an analytic curve with finite length. Buchner [4] later generalized this result to the cut locus of a point in higher-dimensional manifolds. Theorem 3.9, which states that the cut locus of an analytic submanifold (in an analytic manifold) is a simplicial complex, is a natural generalization of Buchner's result (and its proof). We attribute it to Buchner although it is not present in the original paper. This analyticity assumption also helps us to compute the homotopy type of the cut locus of a finite set of points in any closed, orientable, real analytic surface of genus g (Theorem 3.27). In Example 3.29 we make some observations about the cut locus of embedded homology spheres of codimension 2. This includes the case of real analytic knots in the round sphere \mathbb{S}^3 .

We apply our study of gradient of distance-squared function to two families of Lie groups: $\text{GL}(n, \mathbb{R})$ and $U(p, q)$. With a particular choice of left-invariant Riemannian metric which is right-invariant with respect to a maximally compact subgroup K , we analyze the geodesics and the cut locus of K . In both cases, we obtain that G deforms to K via Morse–Bott flow (Lemma 4.1 and Theorem 4.6). Although these results are deducible from classical results of Cartan and Iwasawa, our method is geometric and specific to suitable choices of Riemannian metrics. It also makes very little use of structure theory of Lie algebras.

Organization of the paper In Section 2 we first recall basic definitions of Morse–Bott functions and cut locus of a subset (see Section 2.1). In Section 2.2 we analyze the distance function from $O(n, \mathbb{R})$ in $M(n, \mathbb{R})$. This highlights and motivates Theorem A as well as allows for computation of local homology of singular matrices (Theorem B). In Section 3 we first recall some relevant basic definitions from geometry (see Section 3.1). We make some observations about the differentiability of the distance function (following Wolter [29]) and show that the cut locus is a simplicial complex for an analytic pair (following Buchner [4]). In Section 3.2 we prove Theorem C and discuss some applications, including Theorem D. In Section 3.3 we prove Theorem A. In Section 4 we discuss two specific examples: we analyze the cut locus of $O(n, \mathbb{R})$ inside $\text{GL}(n, \mathbb{R})$ in Section 4.1 and the cut locus of $U(p) \times U(q)$ inside $U(p, q)$ in Section 4.2. In Appendix A we prove Proposition 3.14, the continuity of the map s (see (3-2)). This result is crucial for Section 3.2. In Appendix B we compute the

derivative of the square root map for positive-definite matrices (Lemma B.1). We also analyze the differentiability of the map $A \mapsto \text{tr}(\sqrt{A^T A})$ in Lemma B.2.

Acknowledgements Basu acknowledges the support of the SERB MATRICS grant MTR/2017/000807. Prasad was supported by a UGC (NET)-JRF fellowship.

2 Preliminaries

We recall the notion of Morse function and Morse–Bott function in Section 2.1, keeping in mind the square of the distance function from a submanifold being a potential Morse–Bott function, which we will analyze in Section 3.3. We also recall the definition of cut locus of a subset in a Riemannian manifold. In Example 2.7 we observe that the join of spheres being a sphere can be observed geometrically via cut locus. In Section 2.2 we analyze the cut locus of orthogonal matrices and compute the relative homology of the cut locus (2-8). Along the way, we note that the geometric deformation of $\text{GL}(n, \mathbb{R})$ to $O(n, \mathbb{R})$, obtained via the distance-squared function, is *not* the Gram–Schmidt deformation.

2.1 Background

Given a smooth n -dimensional manifold M , we say that a point $p \in M$ is a *critical point* of a smooth function $f : M \rightarrow \mathbb{R}$ if

$$df_p : T_p M \rightarrow T_{f(p)} \mathbb{R}$$

vanishes. In a coordinate neighborhood $(\phi = (x_1, x_2, \dots, x_n), U)$ around p , for all $j = 1, 2, \dots, n$ we have

$$\frac{\partial(f \circ \phi^{-1})}{\partial x_j}(\phi(p)) = 0.$$

A critical point p is called *nondegenerate* if the determinant of the Hessian matrix

$$\text{Hess}_p(f) := \left(\frac{\partial^2(f \circ \phi^{-1})}{\partial x_i \partial x_j}(\phi(p)) \right)$$

is nonzero. Let us denote the set of all critical points of f by $\text{Cr}(f)$. If all the critical points are nondegenerate, then f is said to be a *Morse function*. Morse–Bott functions are generalizations of Morse functions, where we are allowed to have nondegenerate critical submanifolds.

Definition 2.1 (Morse–Bott functions) Let M be a Riemannian manifold. A smooth submanifold $N \subset M$ is said to be a *nondegenerate critical submanifold* of f if $N \subseteq \text{Cr}(f)$ and, for any $p \in N$, $\text{Hess}_p(f)$ is nondegenerate in the direction normal to N at p . The function f is said to be *Morse–Bott* if the connected components of $\text{Cr}(f)$ are nondegenerate critical submanifolds.

Note that “ $\text{Hess}_p(f)$ is nondegenerate in the direction normal to N at p ” means for any $V \in (T_p N)^\perp$ there exists $W \in (T_p N)^\perp$ such that $\text{Hess}_p(f)(V, W) \neq 0$.

Example 2.2 Let $M = \mathbb{R}^{n+1}$ equipped with the Euclidean metric d . If $N = \mathbb{S}^n$ is the unit sphere, then the distance between a point $p \in \mathbb{R}^{n+1}$ and N is given by

$$d(N, p) := \inf_{q \in N} d(q, p).$$

We shall denote by d^2 the square of the distance. Now consider the function

$$f: M \rightarrow \mathbb{R}, \quad x \mapsto d^2(N, x) = (\|x\| - 1)^2.$$

The function $f: M - \{0\}$ is a Morse–Bott function with $N = \mathbb{S}^n$ as the critical submanifold.

The trace function on $\text{SO}(n, \mathbb{R})$, $U(n, \mathbb{C})$ and $\text{Sp}(n, \mathbb{C})$ is a Morse–Bott function (see Banyaga and Hurtubise [1, Exercise 22, page 90]). We refer the interested reader to [1] for basic results on Morse–Bott theory.

We shall now define the cut locus for a point. The notion of cut locus was first introduced for convex surfaces by Poincaré [23] in 1905 under the name *la ligne de partage*, meaning *the dividing line*.

Definition 2.3 (cut locus) Let M be a complete Riemannian manifold and $p \in M$. If $\text{Cu}(p)$ denotes the *cut locus* of p , then a point q is in $\text{Cu}(p)$ if there exists a minimal geodesic joining p to q any extension of which beyond q is not minimal.

Recall that a minimal geodesic joining p and q is a geodesic that realizes the distance between p and q . The existence of minimal geodesics joining two given points is implied by completeness of the Riemannian manifold. Therefore, in almost all of the examples, the manifolds under consideration will be complete Riemannian manifolds. When $M = \mathbb{S}^n$, ie an n -sphere with the round metric induced from \mathbb{R}^{n+1} , for any $p \in \mathbb{S}^n$, the cut locus $\text{Cu}(p)$ will be the corresponding antipodal point. Later, in

Definition 3.11, we give a slightly different but equivalent definition of cut locus, following Sakai’s book [26, Section 4.1].

In order to have a definition of the cut locus for a submanifold (or a subset), we need to generalize the notion of a minimal geodesic.

Definition 2.4 A geodesic γ is called a *distance-minimal geodesic* joining N to p if there exists $q \in N$ such that γ is a minimal geodesic joining q to p and $\ell(\gamma) = d(N, p) =: d_N(p)$. We will refer to such geodesics as *N -geodesics*.

If N is an embedded submanifold, then an N -geodesic is necessarily orthogonal to N . This follows from the first variational principle. We are ready to define the cut locus for $N \subset M$.

Definition 2.5 (cut locus) Let M be a Riemannian manifold and N be any nonempty subset of M . If $\text{Cu}(N)$ denotes the *cut locus of N* , then we say that $q \in \text{Cu}(N)$ if and only if there exists a distance-minimal geodesic joining N to q such that any extension of it beyond q is not a distance-minimal geodesic.

The cut locus of a sphere (see Example 2.2) is its center. The set $\text{Cu}(p)$ is closed; see Postnikov [24, Exercise 28.4, page 363]. In general, the cut locus of a subset need not be closed, as the following example, due to Sabau and Tanaka [25], illustrates.

Example 2.6 (Sabau and Tanaka 2016) Consider \mathbb{R}^2 with the Euclidean inner product. Let $\{\theta_n\}$, with $\theta_1 \in (0, \pi)$, be a decreasing sequence converging to 0. Let $\overline{B(\mathbf{0}, 1)}$ be the closed unit ball centered at $(0, 0)$. Let $B_n := B(q_n, 1)$ be the open ball with radius 1 and centered at q_n . We have chosen q_n so that it does not belong to $\overline{B(\mathbf{0}, 1)}$ and denotes the center of the circle passing through $p_n = (\cos \theta_n, \sin \theta_n)$ and $p_{n+1} = (\cos \theta_{n+1}, \sin \theta_{n+1})$. Define $N \subset \mathbb{R}^2$ by

$$N := \overline{B(\mathbf{0}, 1)} \setminus \bigcup_{n=1}^{\infty} B(q_n, 1).$$

See Figure 1. Note that N is a closed set and the sequence $\{q_n\}$ of cut points of N converges to the point $(2, 0)$. However, $(2, 0)$ is not a cut point of N .

In Theorem 3.30 we will prove that, for a submanifold N , the set $\text{Cu}(N)$ is closed, by showing that it is the closure of the set of all points in M which have at least two minimal geodesics joining N to $p \in M$.

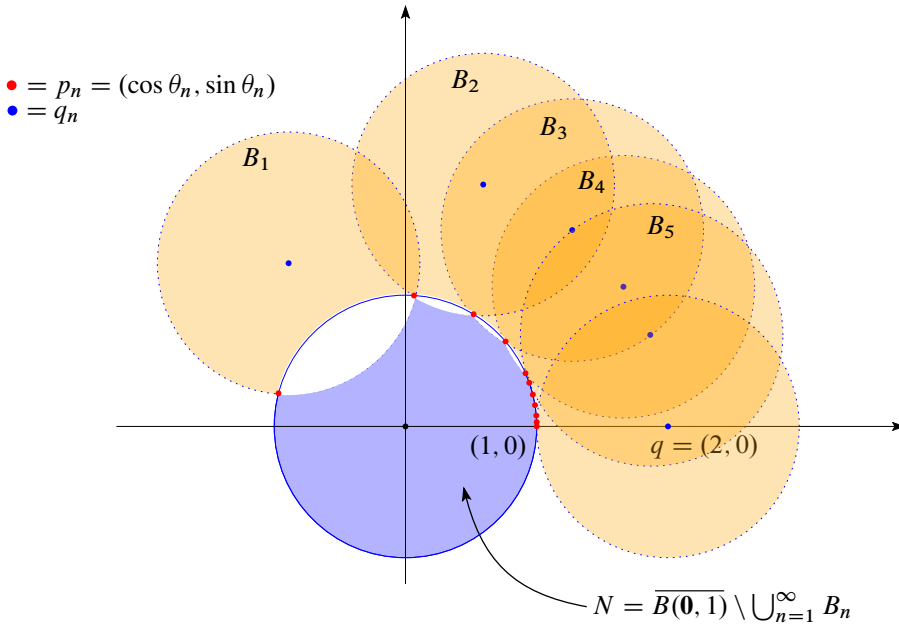


Figure 1: The cut locus need not be closed.

Example 2.7 (join induced by cut locus) Let $S_i^k \hookrightarrow S^n$ denote the embedding of the k -sphere in the first $k + 1$ coordinates and S_i^{n-k-1} denote the embedding of the $(n-k-1)$ -sphere in the last $n - k$ coordinates. It can be seen that $\text{Cu}(S_i^k) = S_i^{n-k-1}$. In fact, starting at a point $p \in S_i^k$ and traveling along a unit-speed geodesic in a direction normal to $T_p S_i^k$, we obtain a cut point at a distance $\frac{\pi}{2}$ from S_i^k . Moreover, in this case, $\text{Cu}(S_i^{n-k-1}) = S_i^k$ and the n -sphere S^n can be expressed as the union of geodesic

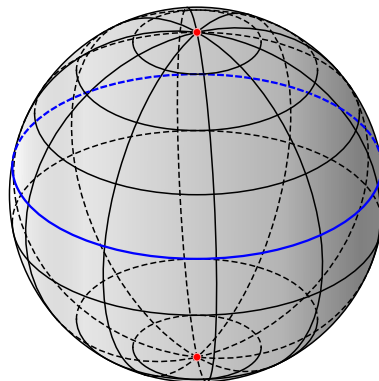


Figure 2: The cut locus of the equator in S^2 .

segments joining \mathbb{S}_i^k to \mathbb{S}_l^{n-k-1} . This is a geometric variant of the fact that the n -sphere is the (topological) join of S^k and S^{n-k-1} . We also observe that $\mathbb{S}^n - \mathbb{S}_l^{n-k-1}$ deforms to \mathbb{S}_i^k while $\mathbb{S}^n - \mathbb{S}_i^k$ deforms to \mathbb{S}_l^{n-k-1} .

In our example, let ν_i^{n-k} and ν_l^{k+1} denote the normal bundles of \mathbb{S}_i^k and \mathbb{S}_l^{n-k-1} , respectively. We may express \mathbb{S}^n as the union of normal disk bundles $D(\nu_i)$ and $D(\nu_l)$. These disk bundles are trivial and are glued along their common boundary $\mathbb{S}_i^k \times \mathbb{S}_l^{n-k-1}$ to produce \mathbb{S}^n . Moreover, \mathbb{S}_i^k is an analytic submanifold of the real analytic Riemannian manifold \mathbb{S}^n with the round metric. There is a generalization of this phenomenon due to Omori [21, Lemmas 1.3–1.5 and Theorem 3.1].

Theorem 2.8 (Omori 1968) *Let M be a compact, connected, real analytic Riemannian manifold which has an analytic submanifold N such that the cut point of N with respect to every geodesic which starts from N and whose initial direction is orthogonal to N has a constant distance π from N . Then $N' = \text{Cu}(N)$ is an analytic submanifold and M has a decomposition $M = DN \cup_\varphi DN'$, where DN and DN' are the normal disk bundles of N and N' , respectively, and φ is the gluing map.*

2.2 An illuminating example

Let $M = M(n, \mathbb{R})$, the set of $n \times n$ matrices, and $N = O(n, \mathbb{R})$, the set of all orthogonal $n \times n$ matrices. Let $A, B \in M(n, \mathbb{R})$. We fix the standard flat Euclidean metric on $M(n, \mathbb{R})$ by identifying it with \mathbb{R}^{n^2} . This induces a distance function given by

$$d(A, B) := \sqrt{\text{tr}((A - B)^T (A - B))}.$$

Consider the distance-squared function

$$f : \text{GL}(n, \mathbb{R}) \rightarrow \mathbb{R}, \quad A \mapsto d_{O(n, \mathbb{R})}^2(A).$$

Lemma 2.9 *The function f can be explicitly expressed as*

$$(2-1) \quad f(A) = n + \text{tr}(A^T A) - 2 \text{tr}(\sqrt{A^T A}).$$

Proof Let $A \in \text{GL}(n, \mathbb{R})$ be any invertible matrix. Then

$$\begin{aligned} (2-2) \quad f(A) &= \inf_{B \in O(n, \mathbb{R})} \text{tr}((A - B)^T (A - B)) \\ &= \inf_{B \in O(n, \mathbb{R})} [\text{tr}(A^T A) - \text{tr}(A^T B) - \text{tr}(B^T A) + \text{tr}(B^T B)] \\ &= \text{tr}(A^T A) + \inf_{B \in O(n, \mathbb{R})} [-2 \text{tr}(A^T B)] + n \\ &= \text{tr}(A^T A) + n - 2 \sup_{B \in O(n, \mathbb{R})} \text{tr}(A^T B). \end{aligned}$$

In order to maximize the function

$$h_A: O(n, \mathbb{R}) \rightarrow \mathbb{R}, \quad B \mapsto \text{tr}(A^T B),$$

for any invertible matrix A , we may first assume that A is a diagonal matrix with positive entries. Then

$$|h_A(B)| = |\text{tr}(A^T B)| = \left| \sum_{i=1}^n a_{ii} b_{ii} \right| \leq \sum_{i=1}^n |a_{ii} b_{ii}| \leq \sum_{i=1}^n a_{ii} = \text{tr}(A^T) = h_A(I).$$

Thus, one of the maximizers is $B = I$. For a general nonsingular matrix A , we will use the *singular value decomposition* (SVD). Write $A = UDV^T$, where U and V are $n \times n$ orthogonal matrices and D is a diagonal matrix with positive entries. For any $B \in O(n, \mathbb{R})$, using the cyclic property of trace, we can see that

$$\text{tr}(A^T B) = \text{tr}(D(U^T B V)).$$

Since $U^T B V$ is an orthogonal matrix, maximizing over B reduces to the earlier observation that B will be a maximizer if $U^T B V = I$, which implies $B = U V^T$.

Since A is invertible, by the polar decomposition, there exists an orthogonal matrix Q and a symmetric positive-definite matrix $S = \sqrt{A^T A}$ such that $A = QS$. Since S is a symmetric matrix, we can diagonalize it, ie $S = P \tilde{D} P^T$, where $P \in O(n, \mathbb{R})$ and \tilde{D} is a diagonal matrix with the eigenvalues of S as its diagonal entries. Thus,

$$A = QS = QP \tilde{D} P^T.$$

Set $U = QP$ and $V = P$ to obtain the SVD of A . In particular, the minimizer is given by

$$B = Q = A \sqrt{A^T A}^{-1}.$$

Therefore,

$$f(A) = n + \text{tr}(A^T A) - 2 \text{tr}(\sqrt{A^T A})$$

for invertible matrices.

In order to compute f for a noninvertible matrix A , we note that $GL(n, \mathbb{R})$ is dense in $M(n, \mathbb{R})$ and that $\sqrt{A^T A}$ is well defined for $A \in M(n, \mathbb{R})$. The continuity of the map $A \mapsto \sqrt{A^T A}$ on $M(n, \mathbb{R})$ implies that the same formula (2-1) for f applies to A as well. □

In order to understand the differentiability of f , it suffices to analyze the function $A \mapsto \text{tr}(\sqrt{A^T A})$.

Lemma 2.10 *The map $g: M(n, \mathbb{R}) \rightarrow \mathbb{R}, A \mapsto \text{tr}(\sqrt{A^T A})$, is differentiable if and only if A is invertible.*

The proof of this postponed to the appendix (see Lemma B.2).

Let us define the function

$$\phi: M(n, \mathbb{R}) \rightarrow \mathbb{R}, \quad A \mapsto \text{tr}(\sqrt{A^T A}).$$

We claim that

$$(2-3) \quad D\phi_A(H) = \text{tr}\left(\int_0^\infty e^{-t\sqrt{A^T A}}(A^T H + H^T A)e^{-t\sqrt{A^T A}} dt\right).$$

The following lemma (see Lemma B.1 for a proof) along with chain rule will prove our claim:

Lemma 2.11 *Let A be a positive-definite matrix and $\psi: A \mapsto \sqrt{A}$. Then*

$$D\psi_A(H) = \int_0^\infty e^{-t\sqrt{A}} H e^{-t\sqrt{A}} dt$$

for any symmetric matrix H .

We may drastically simplify, using basic analysis and linear algebra, the derivative of ϕ given by (2-3) to obtain

$$D\phi_A(H) = \langle A\sqrt{A^T A}^{-1}, H \rangle.$$

For any $A \in \text{GL}(n, \mathbb{R})$,

$$Df_A = 2A - 2A\sqrt{A^T A}^{-1} = -2A(\sqrt{A^T A}^{-1} - I).$$

Hence, the negative gradient of the function f , restricted to $\text{GL}(n, \mathbb{R})$, is given by

$$-\nabla f|_A = 2A(\sqrt{A^T A}^{-1} - I).$$

The critical points are orthogonal matrices. If $\gamma(t)$ is an integral curve of $-\nabla f$ initialized at A , then $\gamma(0) = A$ and

$$(2-4) \quad \frac{d\gamma}{dt} = -2\gamma(t) + 2\gamma(t)\sqrt{\gamma(t)^T \gamma(t)}^{-1} = -2\gamma(t) + 2(\gamma(t)^T)^{-1}\sqrt{\gamma(t)^T \gamma(t)}.$$

Take the test solution of (2-4) given by

$$(2-5) \quad \gamma(t) = Ae^{-2t} + (1 - e^{-2t})(A^T)^{-1}\sqrt{A^T A} = Ae^{-2t} + (1 - e^{-2t})A\sqrt{A^T A}^{-1}.$$

In order to show that $\gamma(t)$ satisfies (2-4), we may verify the simplifications

$$\begin{aligned} \gamma(t)^T \gamma(t) &= (\sqrt{A^T A} e^{-2t} + (1 - e^{-2t}) I)^2, \\ \sqrt{\gamma(t)^T \gamma(t)}^T &= (\sqrt{A^T A} A^{-1} \gamma(t))^T = \gamma(t)^T (A^T)^{-1} \sqrt{A^T A}. \end{aligned}$$

This implies that

$$(\gamma(t)^T)^{-1} \sqrt{\gamma(t)^T \gamma(t)} = (A^T)^{-1} \sqrt{A^T A}.$$

The right-hand side of (2-4), with the test solution, can be simplified to

$$-2Ae^{-2t} + 2e^{-2t} (A^T)^{-1} \sqrt{A^T A},$$

which is the derivative of γ . Thus, $\gamma(t)$, as defined in (2-5), is the required flow line which deforms $GL(n, \mathbb{R})$ to $O(n, \mathbb{R})$. In particular, $GL^+(n, \mathbb{R})$ deforms to $SO(n, \mathbb{R})$ and the other component of $GL(n, \mathbb{R})$ deforms to $O(n, \mathbb{R}) \setminus SO(n, \mathbb{R})$. We note, however, that this deformation takes infinite time to perform the retraction.

Remark 2.12 A modified curve

$$(2-6) \quad \eta(t) = A(1 - t) + tA\sqrt{A^T A}^{-1},$$

with the same image as γ , defines an actual deformation retraction of $GL(n, \mathbb{R})$ to $O(n, \mathbb{R})$. Apart from its origin via the distance function, this is a geometric deformation in the following sense. Given $A \in GL(n, \mathbb{R})$, consider its columns as an ordered basis. This deformation deforms the ordered basis according to the length of the basis vectors and mutual angles between pairs of basis vectors in a geometrically uniform manner. This is in sharp contrast with Gram–Schmidt orthogonalization, also a deformation of $GL(n, \mathbb{R})$ to $O(n, \mathbb{R})$, which is asymmetric as it never changes the direction of the first column, the modified second column only depends on the first two columns, and so on.

We now show that f is Morse–Bott. The tangent space $T_I O(n, \mathbb{R})$ consists of skew-symmetric matrices while the normal vectors at I_n are the symmetric matrices. As left translation by an orthogonal matrix is an isometry of $M(n, \mathbb{R})$, normal vectors at $A \in O(n, \mathbb{R})$ are of the form AW for symmetric matrices W . Since

$$Df_A(H) = 2\langle A, H \rangle - 2\langle A\sqrt{A^T A}^{-1}, H \rangle,$$

the relevant Hessian is

$$\text{Hess}(f)_A(H, H') = \lim_{t \rightarrow 0} \frac{Df_{A+tH'}(H) - Df_A(H)}{t}$$

with $H = AW$ and $H' = AW'$ for symmetric matrices W and W' . A standard computation leads to

$$\text{Hess}(f)_A(H, H') = 2 \text{tr}(H^T H') = 2\langle H, H' \rangle.$$

Therefore, the Hessian matrix restricted to $(T_A O(n, \mathbb{R}))^\perp$ is $2I_{n(n+1)/2}$. This is a recurring feature of distance-squared functions associated to embedded submanifolds (see Proposition 3.5).

There is a relationship between the local homology of cut loci and the reduced Čech cohomology of the *link* of a point in the cut locus. This is due to Theorem 1.4 of Hebda [9] and the remark following it.

Definition 2.13 Let N be an embedded submanifold of a complete smooth Riemannian manifold M . For each $q \in \text{Cu}(N)$, consider the set $\Lambda(q, N)$ of unit tangent vectors at q such that the associated geodesics realize the distance between q and N . This set is called the *link* of q with respect to N .

The set of points in N obtained by the endpoints of the geodesics associated to $\Lambda(q, N)$ will be called the *equidistant set*, denoted by $\text{Eq}(q, N)$, of q with respect to N .

Since the equidistant set $\text{Eq}(q, N)$, consisting of points which realize the distance $d_N(q)$, is obtained by exponentiating the points in $\Lambda(q, N)$, there is a natural surjection map from $\Lambda(q, N)$ to $\text{Eq}(q, N)$.

Theorem 2.14 (Hebda 1983) *Let N be a properly embedded submanifold of a complete Riemannian manifold M of dimension n . If $q \in \text{Cu}(N)$ and v is an element of $\Lambda := \Lambda(q, N)$, then there is an isomorphism*

$$(2-7) \quad \check{H}^i(\Lambda, v) \cong H_{n-1-i}(\text{Cu}(N), \text{Cu}(N) - q).$$

We are interested in computing $\Lambda(A, O(n, \mathbb{R}))$ for singular matrices A . Note that geodesics in $M(n, \mathbb{R})$, initialized at A , are straight lines and any two such geodesics can never meet other than at A . Therefore, there is a natural identification between the link and the equidistant set of A .

Lemma 2.15 *If $A \in M(n, \mathbb{R})$ is singular of rank k , then $\text{Eq}(A, O(n, \mathbb{R}))$ is homeomorphic to $O(n - k, \mathbb{R})$.*

Proof Using the singular value decomposition, we write $A = UDV^T$, where $U, V \in O(n, \mathbb{R})$ and D is a diagonal matrix with entries the eigenvalues of $\sqrt{A^T A}$. If we specify that the diagonal entries of D are arranged in decreasing order, then D is unique. Moreover, as A has rank $k < n$, the first k diagonal entries of D are positive while the last $n - k$ diagonal entries are zero. In order to find the matrices in $O(n, \mathbb{R})$ which realize the distance $d(A, O(n, \mathbb{R}))$, by (2-2), it suffices to find $B \in O(n, \mathbb{R})$ such that

$$\sup_{B \in O(n, \mathbb{R})} \text{tr}(A^T B) = \sup_{B \in O(n, \mathbb{R})} \text{tr}(VDU^T B) = \sup_{B \in O(n, \mathbb{R})} \text{tr}(DU^T B V)$$

is maximized. However, $U^T B V \in O(n, \mathbb{R})$ has orthonormal rows and the specific form of D implies that the maximum happens if and only if $U^T B V$ has e_1, \dots, e_k as the first k rows, in order. Therefore, $U^T B V$ is a block orthogonal matrix, with blocks of I_k and $C \in O(n - k, \mathbb{R})$, ie $B \in U(I_k \times O(n - k, \mathbb{R}))V^T$. □

Corollary 2.16 *Let Sing denote the space of singular matrices in $M(n, \mathbb{R})$. If $A \in \text{Sing}$ is of rank $k < n$, then there is an isomorphism*

$$(2-8) \quad \tilde{H}^i(O(n - k, \mathbb{R})) \cong H_{n^2 - 1 - i}(\text{Sing}, \text{Sing} - A).$$

Proof It follows from Lemma 2.15 that $\Lambda(A, O(n, \mathbb{R})) \cong O(n - k, \mathbb{R})$ if A has rank k . Since $O(n - k, \mathbb{R})$ is a manifold, the Čech and singular cohomology groups are isomorphic. The space Sing is a star-convex set, whence all homotopy and homology groups are that of a point. Applying (2-7) in our case, we obtain an isomorphism

$$\tilde{H}^i(O(n - k, \mathbb{R})) \cong H_{n^2 - 1 - i}(\text{Sing}, \text{Sing} - A)$$

between the reduced cohomology and local homology groups. In particular, the local homology of the cut locus at A detects the rank of A . □

Similar computations hold for $U(n, \mathbb{C})$ and singular $n \times n$ complex matrices.

3 Main results

We recall some results about exponential maps and Fermi coordinates in Section 3.1. A result of Wolter [29] may be generalized to prove (Lemma 3.7) that the distance-squared function from a submanifold is not differentiable on the separating set. This result may be well known to experts, but the proof, following Wolter, is elementary. Buchner’s result [4] may be generalized to prove (Theorem 3.9) that the cut locus is a simplicial

complex for real analytic pairs. In Section 3.2 we recall the notion of Thom space and apply it to the normal bundle of an embedded submanifold in a closed, connected Riemannian manifold. Our first main result, Theorem 3.16, states that the quotient of the ambient manifold by the cut locus of the submanifold results in the Thom space of the normal bundle. As a consequence we obtain Theorem 3.24, which says that a homology k -sphere inside a manifold homeomorphic to S^d has cut locus weakly homotopy equivalent to S^{d-k-1} provided $d - k \geq 3$, $k > 0$ and $H_{d-1}(\text{Cu}(N))$ is torsion-free. Theorem 3.27 is another consequence about analytic surfaces. In Section 3.3 we prove (see Theorem 3.30) that the cut locus of a submanifold is closed, essentially following Wolter’s arguments [29]. This leads us to the other main result, Theorem 3.32, which proves that the complement of the cut locus $\text{Cu}(N)$ deforms to N .

3.1 Basic results

For understanding the geometry in the neighborhood of a submanifold, it is convenient to use Fermi coordinates, a generalization of normal coordinates. We shall briefly introduce Fermi coordinates and state some of their relevant properties. Let N be an embedded submanifold of a Riemannian manifold M . Let ν be the normal bundle of $N \subseteq M$, ie

$$\nu := \{(p, v) : p \in N, v \in (T_p N)^\perp\}.$$

In fact, ν is a subbundle of the restriction of TM to N . We define the *exponential map of the normal bundle* as

$$(3-1) \quad \exp_\nu : \nu \rightarrow M, \quad \exp_\nu(p, v) := \exp_p(v) \quad \text{for } (p, v) \in \nu.$$

We may write $\exp_\nu(v)$ in short and call this the *normal exponential map*.

Now we will list some lemmas; for proofs we refer to Gray [8, Sections 2.1 and 2.3].

Lemma 3.1 *Let N be a topologically embedded submanifold of a Riemannian manifold M . Then the normal exponential map $\exp_\nu : \nu \rightarrow M$ maps a neighborhood of N in ν diffeomorphically onto a neighborhood of N in M .*

Let \mathcal{O}_N denote the largest neighborhood of the zero section of ν for which \exp_ν is a diffeomorphism. We shall later be able to describe this neighborhood in terms of a function s ; see (3-2). To define a system of Fermi coordinates, we need an arbitrary system of coordinates (y_1, \dots, y_k) defined in a neighborhood $\mathcal{U} \subset N$ of $p \in N$ together with orthogonal sections E_{k+1}, \dots, E_n of the restriction of ν to \mathcal{U} .

Definition 3.2 (Fermi coordinates) The Fermi coordinates (x_1, \dots, x_n) of $N \subset M$ centered at p (relative to a given coordinate system (y_1, \dots, y_k) on N and orthogonal sections E_{k+1}, \dots, E_n of ν) are defined by

$$x_l \left(\exp_\nu \left(\sum_{j=k+1}^n t_j E_j(p') \right) \right) = y_l(p') \quad \text{for } l = 1, \dots, k,$$

$$x_i \left(\exp_\nu \left(\sum_{j=k+1}^n t_j E_j(p') \right) \right) = t_i \quad \text{for } i = k + 1, \dots, n,$$

for $p' \in U$ provided the numbers t_{k+1}, \dots, t_n are small enough that

$$t_{k+1} E_{k+1}(p') + \dots + t_n E_n(p') \in \mathcal{O}_N.$$

Since \exp_ν is a diffeomorphism on \mathcal{O}_N , $(x_1, \dots, x_k, x_{k+1}, \dots, x_n)$ defines a coordinate system near p . In fact, the restrictions to N of the coordinate vector fields $\partial/\partial x_{k+1}, \dots, \partial/\partial x_n$ are orthonormal.

Lemma 3.3 Let γ be a unit-speed geodesic normal to N with $\gamma(0) = p \in N$. If $u = \gamma'(0)$, then there is a system of Fermi coordinates (x_1, \dots, x_n) such that, for small enough t , ie for $(p, tu) \in \mathcal{O}_N$, we have

$$\frac{\partial}{\partial x_{k+1}} \Big|_{\gamma(t)} = \gamma'(t), \quad \frac{\partial}{\partial x_l} \Big|_p \in T_p N, \quad \frac{\partial}{\partial x_i} \Big|_p \in (T_p N)^\perp$$

for $1 \leq l \leq k$ and $k + 1 \leq i \leq n$. Furthermore, for $1 \leq j \leq n$,

$$(x_j \circ \gamma)(t) = t \delta_{j(k+1)}.$$

Definition 3.4 Let (x_1, \dots, x_n) be a system of Fermi coordinates for $N \subset M$. Define $\sigma(x_1, \dots, x_n)$ to be the nonnegative number satisfying

$$\sigma^2 = \sum_{i=k+1}^n x_i^2.$$

It is known that σ does not depend on the choice of Fermi coordinates.

Proposition 3.5 Let U be a neighborhood of N such that each point in U admits a unique unit-speed N -geodesic. If $p \in U$, then

$$\sigma(p) = d_N(p).$$

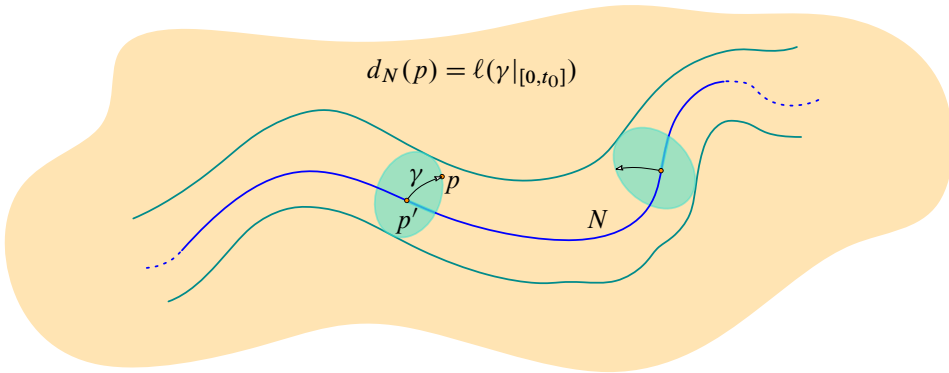


Figure 3: Distance via Fermi coordinates.

Proof Since the expression of σ is independent of the choice of the Fermi coordinates, we will make a special choice of the Fermi coordinates (x_1, \dots, x_n) . For $p \in U$, choose the unique unit-speed N -geodesic γ joining p to N . This geodesic meets N orthogonally at $\gamma(0) = p'$. Choose t_0 so that $\gamma(t_0) = p$; see Figure 3. According to Lemma 3.3, there is a system of Fermi coordinates (x_1, \dots, x_n) centered at p' such that $x_i(\gamma(t)) = t\delta_{i(k+1)}$. The sequence of equalities

$$\sigma(p) = x_{k+1}(\gamma(t_0)) = t_0 = d_N(p)$$

completes the proof. □

Corollary 3.6 Consider the distance-squared function with respect to a submanifold N in M . The Hessian of the distance-squared function at the critical submanifold N is nondegenerate in the normal direction.

Towards the regularity of the distance-squared function, the following observation will be useful. It is a routine generalization of [29, Lemma 1].

Lemma 3.7 Let M be a connected, complete Riemannian manifold and N be an embedded submanifold of M . Suppose two N -geodesics exist joining N to $q \in M$. Then $d_N^2 : M \rightarrow \mathbb{R}$ has no directional derivative at q for vectors in direction of those two N -geodesics.

Proof Let us assume that all the geodesics are parametrized by arc length. Let $\gamma_i : [0, \hat{t}] \rightarrow M$ for $i = 1, 2$ be two distinct geodesics with $\gamma_1(0), \gamma_2(0) \in N$ and $\gamma_1(l) = q = \gamma_2(l)$, where $l = d_N(q)$ and $0 < l < \hat{t}$. Let us suppose that the two

geodesics start at p_1 and p_2 and so $d(p_1, q) = l = d(p_2, q)$. Note that the directional derivative of d^2 at q in the direction of $\gamma'_i(q)$ from the left is given by

$$(d^2)'_-(q) := \lim_{\varepsilon \rightarrow 0^+} \frac{d_N^2(\gamma_i(l)) - d_N^2(\gamma_i(l - \varepsilon))}{\varepsilon} = \lim_{\varepsilon \rightarrow 0^+} \frac{l^2 - (l - \varepsilon)^2}{\varepsilon} = 2l.$$

Next, we claim that the derivative of the same function from the right is strictly bounded above by $2l$. Let $\omega \in (0, \pi]$ be the angle between the two geodesics γ_1 and γ_2 at q . Define the function

$$u(\tau) := d_N(\gamma_1(l - \varepsilon)) + d(\gamma_1(l - \varepsilon), \gamma_2(\tau + l)).$$

By the triangle inequality, we observe that

$$f(\tau) := (u(\tau))^2 \geq d^2(p_1, \gamma_2(\tau + l)) \geq d_N^2(\gamma_2(\tau + l)),$$

and equality holds at $\tau = 0$ and $(u(0))^2 = d_N^2(q) = l^2$. Thus, in order to prove the claim, it suffices to show that the derivative of f from the right, at $\tau = 0$, is bounded below by $2l$. We need to invoke a version of the cosine law for small geodesic triangles. Although this may be well known to experts, we will use the version that appears in Sharafutdinov’s work [27] (see also Daniilidis et al [6, Lemma 2.4] for a detailed proof). In our case, this means that

$$d^2(\gamma_1(l - \varepsilon), \gamma_2(\tau + l)) = \varepsilon^2 + \tau^2 + 2\varepsilon\tau \cos \omega + K(\tau)\varepsilon^2\tau^2,$$

where $|K(\tau)|$ is bounded and the side lengths are sufficiently small. Note that we are considering geodesic triangles with two vertices constant and the varying vertex being $\gamma_2(l + \tau)$. It follows from taking a square root and then expanding in powers of τ that

$$d(\gamma_1(l - \varepsilon), \gamma_2(\tau + l)) = \sqrt{\varepsilon^2 + \tau^2 + 2\varepsilon\tau \cos \omega} (1 + O(\tau^2)).$$

It follows that

$$u(\tau) = l - \varepsilon + \sqrt{\varepsilon^2 + \tau^2 + 2\varepsilon\tau \cos \omega} (1 + O(\tau^2)).$$

Therefore, $u'_+(\tau) = \cos \omega = d'_+(\gamma_1(l - \varepsilon), \gamma_2(l))$. Observe that

$$\begin{aligned} f'_+(\tau)|_{\tau=0} &= 2d_N(\gamma_1(l - \varepsilon))d'_+(\gamma_1(l - \varepsilon), \gamma_2(l)) + 2d(\gamma_1(l - \varepsilon), \gamma_2(l))d'_+(\gamma_1(l - \varepsilon), \gamma_2(l)) \\ &= 2d_N(\gamma_1(l - \varepsilon)) \cos \omega + 2d(\gamma_1(l - \varepsilon), \gamma_2(l)) \cos \omega \\ &= 2d_N(\gamma_1(l)) \cos \omega < 2l. \end{aligned}$$

Thus, we have proved the claim and subsequently the result. □

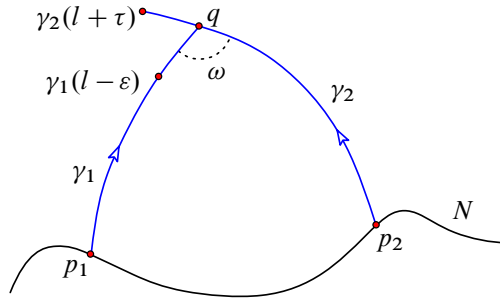


Figure 4: When two N -geodesics meet.

The above lemma prompts us to define the following set, the notation being consistent with Wolter’s paper [29]:

Definition 3.8 Let N be a subset of a Riemannian manifold M . The set $\text{Se}(N)$, called the *separating set*,¹ consists of all points $q \in M$ such that at least two distance-minimal geodesics from N to q exist.

If $q \in \text{Se}(N)$ but $q \notin \text{Cu}(N)$, then we have Figure 4, ie γ_1 is an N -geodesic beyond q while γ_2 is another N -geodesic for q . The triangle inequality applied to $\gamma_1(0)$, $q = \gamma_1(l)$ and $\gamma_2(l + \tau)$ implies that

$$d_N(\gamma_2(l + \tau)) < l + \tau,$$

while, for τ small enough, $d_N(\gamma_2(l + \tau)) = l + \tau$ as γ_2 is an N -geodesic beyond q . This contradiction establishes the well-known fact $\text{Se}(N) \subseteq \text{Cu}(N)$. In quite a few examples, these two sets are equal. In the case of $M = \mathbb{S}^n$ with $N = \{p\}$, the set $\text{Se}(N)$ consists of $-p$. There is an infinite family of minimal geodesics joining p to $-p$. An appropriate choice of a pair of such minimal geodesics would create a loop, which is permissible in the definition of $\text{Se}(N)$.

Regarding the question of cut loci being triangulable, we recall the result of Buchner [4] that the cut locus (of a point) of a real analytic Riemannian manifold (of dimension d) is a simplicial complex of dimension at most $d - 1$. It follows, without many changes, that the result holds for cut loci of submanifolds as well. Hence, we attribute the following result to Buchner:

¹We could not find any name for this set in the literature. This terminology is our own although this nomenclature is rarely used in the paper.

Theorem 3.9 (Buchner 1977) *Let N be an analytic submanifold of a real analytic manifold M . If M is of dimension d , then the cut locus $\text{Cu}(N)$ is a simplicial complex of dimension at most $d - 1$.*

The obvious modifications to the proof by Buchner are the following:

- (i) Choose ε to be such that there is a unique geodesic from p to q if $(p, q) < \varepsilon$ and, if $d_N(q) < \varepsilon$, then there is a unique N -geodesic to q .
- (ii) Consider the set $\Omega_N(t_0, t_1, \dots, t_k)$, the space of piecewise broken geodesics starting at N , and define $\Omega_N(t_0, t_1, \dots, t_k)^S$ analogously.
- (iii) The map

$$\Omega_N(t_0, t_1, \dots, t_k)^S \rightarrow N \times M \times \dots \times M, \quad \omega \mapsto (\omega(t_0), \omega(t_1), \dots, \omega(t_k)),$$

determines an analytic structure on $\Omega_N(t_0, t_1, \dots, t_k)^S$.

The remainder of the proof works essentially verbatim.

Remark 3.10 As we have seen in Example 2.7, the dimension of the cut locus of a k -dimensional submanifold is $d - k - 1$. However, generically, we may not expect this to be true. In fact, for real analytic knots (except the unknot) in \mathbb{S}^3 , it is always the case that the cut locus cannot be homotopic to a (connected) 1-dimensional simplicial complex (see Example 3.29).

3.2 Thom space via cut locus

Let (M, g) be a complete Riemannian manifold with distance function d . The exponential map at p ,

$$\exp_p: T_p M \rightarrow M,$$

is defined on the tangent space. Moreover, there exists a minimal geodesic joining any two points in M . However, not all geodesics are distance-realizing. Given $v \in T_p M$ with $\|v\| = 1$, let γ_v be the geodesic initialized at p with velocity v . Let $S(TM)$ denote the unit tangent bundle and let $[0, \infty]$ be the one-point compactification of $[0, \infty)$. Define

$$s: S(TM) \rightarrow [0, \infty], \quad s(v) := \sup\{t \in [0, \infty) : \gamma_v|_{[0,t]} \text{ is minimal}\}.$$

Definition 3.11 (cut locus) Let M be a complete, connected Riemannian manifold. If $s(v) < \infty$ for some $v \in S(T_p M)$, then $\exp_p(s(v)v)$ is called a *cut point*. The collection of cut points is defined to be the cut locus of p .

As geodesics are locally distance-realizing, $s(v) > 0$ for any $v \in S(TM)$. The following result [26, Proposition 4.1] will be important for the underlying ideas in its proof:

Proposition 3.12 *The map $s : S(TM) \rightarrow [0, \infty]$, $u \mapsto s(u)$, is continuous.*

The proof relies on a characterization of $s(v)$ provided $s(v) < \infty$. A positive real number T is $s(v)$ if and only if $\gamma_v : [0, T] \rightarrow M$ is minimal and at least one of the following holds:

- (i) $\gamma_v(T)$ is the first conjugate point of p along γ_v .
- (ii) There exists $u \in S(T_pM)$ with $u \neq v$ and $\gamma_u(T) = \gamma_v(T)$.

Recall that, if $\gamma : [0, a] \rightarrow M$ is a geodesic, then $q = \gamma(t_0)$ is conjugate to $p = \gamma(0)$ along γ if \exp_p is singular at $t_0\dot{\gamma}(0)$, ie $(D \exp_p)(t_0\dot{\gamma}(0))$ is not of full rank.

Remark 3.13 If M is compact, then it has bounded diameter, which implies that $s(v) < \infty$ for any $v \in S(TM)$. The converse is also true: if M is complete and connected with $s(v) < \infty$ for any $v \in S(TM)$, then M has bounded diameter, whence it is compact.

We shall be concerned with closed Riemannian manifolds in what follows. Let N be an embedded submanifold inside a closed, ie compact without boundary, manifold M . Let ν denote the normal bundle of N in M with $D(\nu)$ denoting the unit disk bundle. In the context of $S(\nu)$, the unit normal bundle and the cut locus of N , distance-minimal geodesics or N -geodesics are relevant (see Definitions 2.4 and 2.5). We want to consider

$$(3-2) \quad s : S(\nu) \rightarrow [0, \infty), \quad s(v) := \sup\{t \in [0, \infty) : \gamma_v|_{[0,t]} \text{ is an } N\text{-geodesic}\}.$$

Notice that $0 < s(v) \leq s(v)$ for any $v \in S(\nu)$. In the special case when $N = \{p\}$, s is simply the restriction of s to T_pM . Analogous to Proposition 3.12, we have the following result:

Proposition 3.14 *The map $s : S(\nu) \rightarrow [0, \infty)$, as defined in (3-2), is continuous.*

As expected, the proof of Proposition 3.14 relies on a characterization of $s(v)$ similar to that of $s(v)$ (refer to Lemma A.2 and Bishop and Crittenden’s book [3, Exercise 23, page 241]).

Let us postpone the proofs (see Appendix A) and proceed with some immediate applications.

Definition 3.15 (rescaled exponential) The *rescaled exponential* or *s*-exponential map is defined to be

$$\widetilde{\text{exp}}: D(v) \rightarrow M, \quad (p, v) \mapsto \begin{cases} \exp_p(s(\hat{v})v) & \text{if } v = \|v\|\hat{v} \neq 0, \\ p & \text{if } v = 0. \end{cases}$$

We are now ready to prove the main result of this section.

Theorem 3.16 Let N be an embedded submanifold inside a closed, connected Riemannian manifold M . If ν denotes the normal bundle of N in M , then there is a homeomorphism

$$\widetilde{\text{exp}}: D(\nu)/S(\nu) \xrightarrow{\cong} M/\text{Cu}(N).$$

Proof It follows from Proposition 3.14 that the rescaled exponential is continuous. Moreover, $\widetilde{\text{exp}}$ is surjective and $\widetilde{\text{exp}}(S(\nu)) = \text{Cu}(N)$. If there exist $(p, v) \neq (q, w) \in D(\nu)$ such that

$$\widetilde{\text{exp}}(p, v) = \widetilde{\text{exp}}(q, w) = p',$$

then $d_N(p')$ can be computed in two ways to obtain

$$d_N(p') = s(\hat{v})\|v\| = s(\hat{w})\|w\|.$$

Thus, $T = d(p', N)$ is a number such that $\gamma_v: [0, T] \rightarrow M$ is an N -geodesic and $\gamma_v(T) = \gamma_w(T) = p'$. By Lemma A.2, we conclude that $T = s(\hat{v}) = s(\hat{w})$, whence $\|v\| = \|w\| = 1$. Therefore, $\widetilde{\text{exp}}$ is injective on the interior of $D(\nu)$.

As $\text{Cu}(N)$ is closed and M is a compact metric space, the quotient space $M/\text{Cu}(N)$ is Hausdorff. As the quotient $D(\nu)/S(\nu)$ is compact, standard topological arguments imply the map induced by the rescaled exponential is a homeomorphism. □

Recall that the *Thom space* $\text{Th}(E)$ of a real vector bundle $E \rightarrow B$ of rank k is $D(E)/S(E)$, where it is understood that we have chosen a Euclidean metric on E . If B is compact, then the Thom space $\text{Th}(E)$ is the one-point compactification of E . In general, we compactify the fibers and then collapse the section at infinity to a point to obtain $\text{Th}(E)$. Thus, Thom spaces obtained via two different metrics are homeomorphic. We will now revisit a basic property of Thom space via its connection to the cut locus. It can be seen that

$$(3-3) \quad \text{Cu}(N_1 \times N_2) = (\text{Cu}(N_1) \times M_2) \cup (M_1 \times \text{Cu}(N_2))$$

for an embedding $N_1 \times N_2$ inside $M_1 \times M_2$. If ν_j is the normal bundle of N_j inside M_j , then Theorem 3.16 along with (3-3) implies that

$$\begin{aligned} \text{Th}(\nu_1 \oplus \nu_2) &\cong \frac{M_1 \times M_2}{(M_1 \times \text{Cu}(N_2)) \cup (\text{Cu}(N_1) \times M_2)} \cong \frac{M_1/\text{Cu}(N_1) \times M_2/\text{Cu}(N_2)}{M_1/\text{Cu}(N_1) \vee M_2/\text{Cu}(N_2)} \\ &\cong \text{Th}(\nu_1) \wedge \text{Th}(\nu_2). \end{aligned}$$

Let $N = N_1 \sqcup N_2$ be a disjoint union of connected manifolds of the same dimension. If $N \hookrightarrow M$, then let ν_j denote the normal bundle of N_j in M . If ν is the normal bundle of N in M , then

$$(3-4) \quad \text{Th}(\nu) \cong \text{Th}(\nu_1) \vee \text{Th}(\nu_2).$$

This implies that

$$M/\text{Cu}(N) \cong M/\text{Cu}(N_1) \vee M/\text{Cu}(N_2).$$

Example 3.17 Consider the two circles

$$N_1 = \{(\cos t, \sin t, 0, 0) \mid t \in \mathbb{R}\}, \quad N_2 = \{(0, 0, \cos t, \sin t) \mid t \in \mathbb{R}\}$$

in \mathbb{S}^3 . The link $N := N_1 \sqcup N_2$ has linking number 1. It can be checked that

$$\text{Cu}(N) = \left\{ \frac{1}{\sqrt{2}}(\cos s, \sin s, \cos t, \sin t) \mid s, t \in \mathbb{R} \right\}$$

is a torus. Note that $\text{Cu}(N_1) = N_2$ and vice versa as well as

$$\mathbb{S}^3/\text{Cu}(N_j) \cong (S^1 \times S^2)/(S^1 \times \infty),$$

where $S^1 \times S^2$ is the fiberwise compactification of the normal bundle of N_j . We conclude that

$$\mathbb{S}^3/\text{Cu}(N) \cong \left(\frac{S^1 \times S^2}{S^1 \times \infty} \right) \vee \left(\frac{S^1 \times S^2}{S^1 \times \infty} \right).$$

There are some topological similarities between $\text{Cu}(N)$ and $M - N$.

Lemma 3.18 *The cut locus $\text{Cu}(N)$ is a strong deformation retract of $M - N$. In particular, $(M, \text{Cu}(N))$ is a good pair and the number of path components of $\text{Cu}(N)$ equals that of $M - N$.*

Proof Consider the map $H : (M - N) \times [0, 1] \rightarrow M - N$ defined via the normal exponential map

$$H(q, t) = \begin{cases} \exp_\nu \left[\left\{ t \cdot s \left(\frac{\exp_\nu^{-1}(q)}{\|\exp_\nu^{-1}(q)\|} \right) + (1-t) \|\exp_\nu^{-1}(q)\| \right\} \frac{\exp_\nu^{-1}(q)}{\|\exp_\nu^{-1}(q)\|} \right] & \text{if } q \in M - (\text{Cu}(N) \cup N), \\ q & \text{if } q \in \text{Cu}(N). \end{cases}$$

If $q \in M - (\text{Cu}(N) \cup N)$, then let γ be the unique N -geodesic joining N to q . The path $H(q, t)$ is the image of this geodesic from q to the first cut point along γ . The continuity of s implies that H is continuous. It also satisfies $H(q, 0) = q$ and $H(q, 1) \in \text{Cu}(N)$. The claims about good pair and path components are clear. \square

Corollary 3.19 *If two embeddings $f, g: N \rightarrow M$ are ambient isotopic, then $\text{Cu}(f(N))$ and $\text{Cu}(g(N))$ are homotopy equivalent.*

Proof The hypothesis implies that there is a diffeomorphism $\varphi: M \rightarrow M$ such that $\varphi(f(N)) = g(N)$. Thus, $M - \text{Cu}(f(N))$ is homeomorphic to $M - \text{Cu}(g(N))$ and the claim follows from the lemma above. Note that, in the smooth category, the notion of isotopic and ambient isotopic are equivalent (refer to Section 8.1 of Hirsch’s book [12]). Thus, the same conclusion holds if we assume that the embeddings are isotopic. \square

Remark 3.20 Without the assumption of M being closed, the above result fails to be true. One may consider $M = S^1 \times \mathbb{R}$ with the natural product metric and $N = S^1$. In fact, the universal cover of M is $\mathbb{R} \times \mathbb{R}$ while that of N is \mathbb{R} . If we choose a periodic curve in \mathbb{R}^2 which is isotopic to the x -axis and has nonempty cut locus in \mathbb{R}^2 , then we may pass via the covering map to obtain an embedding g of N isotopic to the embedding f identifying N with $S^1 \times \{0\}$. For this pair, $\text{Cu}(f(N)) = \emptyset$ while $\text{Cu}(g(N)) \neq \emptyset$.

Several other identifications between topological invariants can be explored. For instance, if $\iota: N^k \hookrightarrow M^d$ is, as before, such that $M - N$ is path-connected, then

$$(3-5) \quad \iota_*: \pi_j(\text{Cu}(N)) \xrightarrow{\cong} \pi_j(M)$$

if $0 \leq j \leq d - k - 2$ while ι_* is a surjection for $j = d - k - 1$. The proof of this relies on a general position argument, ie being able to find a homotopy of the sphere that avoids N , followed by Lemma 3.18. Surjectivity of ι_* if $j \leq d - k - 1$ is imposed by the requirement that a sphere S^j in general position must not intersect N^k . Injectivity of ι for $j \leq d - k - 2$ is imposed by the condition that a homotopy $S^j \times [0, 1]$ in general position must not intersect N^k . This observation (3-5) generalizes a result of Sakai [26, Proposition 4.5(1)].

The inclusion $i: \text{Cu}(N) \hookrightarrow M$ induces a long exact sequence in homology

$$\dots \rightarrow H_j(\text{Cu}(N)) \xrightarrow{i_*} H_j(M) \rightarrow H_j(M, \text{Cu}(N)) \xrightarrow{\partial} H_{j-1}(\text{Cu}(N)) \rightarrow \dots$$

As $(M, \text{Cu}(N))$ is a good pair (see Lemma 3.18), we replace the relative homology of $(M, \text{Cu}(N))$ with the reduced homology of $M/\text{Cu}(N) \cong \text{Th}(v)$. This results in the

long exact sequence

$$(3-6) \quad \cdots \rightarrow H_j(\text{Cu}(N)) \xrightarrow{i_*} H_j(M) \xrightarrow{q} \tilde{H}_j(\text{Th}(v)) \xrightarrow{\partial} H_{j-1}(\text{Cu}(N)) \rightarrow \cdots .$$

If $N = \{p\}$ is a point, then $\text{Th}(v) = S^d$ and (3-6) imply isomorphisms

$$i_* : H_j(\text{Cu}(p)) \xrightarrow{\cong} H_j(M), \quad i^* : H^j(M) \xrightarrow{\cong} H^j(\text{Cu}(p))$$

for $j \neq d, d - 1$ (see [26, Proposition 4.5(2)]).

Remark 3.21 The long exact sequence (3-6) can be interpreted as the dual to the long exact sequence in cohomology of the pair (M, N) . If $N = N_1 \sqcup \cdots \sqcup N_l$ is a disjoint union of submanifolds of dimension k_1, \dots, k_l , respectively, then the Thom isomorphism implies that

$$\begin{aligned} \tilde{H}_j(\text{Th}(v)) &\cong \tilde{H}_j(\text{Th}(v_1)) \oplus \cdots \oplus \tilde{H}_j(\text{Th}(v_l)) \\ &\cong H_{j-(d-k_1)}(N_1) \oplus \cdots \oplus H_{j-(d-k_l)}(N_l), \end{aligned}$$

where v_j is the normal bundle of N_j . Applying Poincaré duality to each N_j , we obtain isomorphisms

$$\tilde{H}_j(\text{Th}(v)) \cong \bigoplus_{i=1}^l H^{d-j}(N_i) = H^{d-j}(N).$$

Poincaré–Lefschetz duality applied to the pair (M, N) provides isomorphisms

$$(3-7) \quad \check{H}^j(M, N) \cong H_{d-j}(M - N).$$

As M and N are triangulable, Čech cohomology may be replaced by singular cohomology. Since $M - N$ deforms to $\text{Cu}(N)$ by Lemma 3.18, we have isomorphisms

$$(3-8) \quad H^j(M, N) \cong H_{d-j}(\text{Cu}(N)).$$

Combining all these isomorphisms, we obtain the long exact sequence in cohomology for (M, N) from (3-6).

Lemma 3.22 *Let N be a closed submanifold of M with l components. If M has dimension d , then $H_{d-1}(\text{Cu}(N))$ is free abelian of rank $l - 1$ and $H_{d-j}(\text{Cu}(N)) \cong H^j(M)$ if $j - 2 \geq k$, where k is the maximum of the dimensions of the components of N .*

Proof It follows from (3-7) that

$$H_{d-1}(\text{Cu}(N)) \cong H^1(M, N).$$

Consider the long exact sequence associated to the pair (M, N) ,

$$0 \rightarrow H^0(M, N) \rightarrow H^0(M) \xrightarrow{i^*} H^0(N) \rightarrow H^1(M, N) \rightarrow H^1(M) \rightarrow H^1(N) \rightarrow \dots$$

If N has l components, ie $N = N_1 \sqcup \dots \sqcup N_l$, where N_j has dimension k_j , then $H^1(M, N)$ is torsion-free. This follows from the fact that $i^*(1) = (1, \dots, 1)$ and $H^1(M)$ is free abelian. In particular, if $H^1(M) = 0$, then $H_{d-1}(\text{Cu}(N)) \cong \mathbb{Z}^{l-1}$.

The long exact sequence for the pair (M, N) implies that there are isomorphisms

$$(3-9) \quad H_{d-j}(\text{Cu}(N)) \cong H^j(M, N) \xrightarrow{\cong} H^j(M)$$

if $j \geq k + 2$, where $k = \max\{k_1, \dots, k_l\}$. □

Remark 3.23 The cut locus can be very hard to compute. For a general space, we have the notion of topological dimension. This notion coincides with the usual notion if the space is triangulable. However, Barratt and Milnor [2] proved that the singular homology of a space may be nonzero beyond its topological dimension. Čech (co)homology is better equipped to detect topological dimension and is the reason why one may prefer it over singular homology due to the generic fractal-like nature of cut loci (see the remarks following Theorem C in Section 1). Although the topological dimension of $\text{Cu}(N)$ is at most $d - 1$, it is not apparent that $H_{d-1}(\text{Cu}(N))$ is a free abelian group.

There are several applications of this discussion.

Theorem 3.24 *Let N be a smooth homology k -sphere embedded in a Riemannian manifold homeomorphic to S^d . If $d \geq k + 3$, then the cut locus $\text{Cu}(N)$ is homotopy equivalent to S^{d-k-1} .*

Proof As N has codimension at least 3, its complement is path-connected. It follows from (3-5) and Lemma 3.18 that $M - N$ is $(d - k - 2)$ -connected. In particular, $M - N$ is simply connected and, by the Hurewicz isomorphism, $H_j(M - N) = 0$ if $j \leq d - k - 2$. Note that $H_d(M - N) = 0$ as $M - N$ is a noncompact manifold of dimension d .

If $k > 0$, then, by Lemma 3.22, $H_{d-1}(M - N) = 0$. Moreover, by Poincaré-Lefschetz duality (3-7), the only nonzero higher homology of $M - N$ is $H_{d-k-1}(M - N) \cong \mathbb{Z}$. By the Hurewicz theorem, there is an isomorphism $\pi_{d-k-1}(M - N) \cong \mathbb{Z}$. Let

$$\alpha: S^{d-k-1} \rightarrow M - N$$

be a generator. The map α_* induces an isomorphism on all homology groups between two simply connected CW complexes. It follows from Whitehead’s theorem that α is a homotopy equivalence. Using Lemma 3.18, we obtain our homotopy equivalence $H_1 \circ \alpha: S^{d-k-1} \rightarrow \text{Cu}(N)$.

If $k = 0$, then, by Lemma 3.22, $H_{d-1}(M - N) \cong \mathbb{Z}$. Arguments similar to the $k > 0$ case now apply to obtain a homotopy equivalence with S^{d-1} . □

The above result was foreshadowed by Example 2.7, where we showed that the cut locus of $N = S_i^k$ inside $M = S^d$ is S_i^{d-k-1} . It also differs from Poincaré–Lefschetz duality in that we are able to detect the exact homotopy type of the cut locus. In fact, when M and N are real analytic and the embedding is also real analytic, then, by Theorem 3.9, we infer that $\text{Cu}(N)$ is a simplicial complex of dimension at most $d - 1$. Towards this direction, Theorem 3.24 can be pushed further.

Proposition 3.25 *Let N be a real analytic homology k -sphere embedded in a real analytic homology d -sphere M . If $d \geq k + 3$, then the cut locus $\text{Cu}(N)$ is a simplicial complex of dimension at most $d - 1$, having the homology of the $(d - k - 1)$ -sphere with fundamental group isomorphic to that of M .*

The proof of this is a combination of ideas used in the proof of Theorem 3.24. The homotopy type cannot be deduced here due to the presence of a nontrivial fundamental group. An intriguing example can be obtained by combining Proposition 3.25 and the Poincaré homology sphere.

Example 3.26 (cut locus of 0-sphere in the Poincaré sphere) Let \tilde{I} be the binary icosahedral group. It is a double cover of I , the icosahedral group, and can be realized a subgroup of $\text{SU}(2)$. It is known that $H_1(\tilde{I}; \mathbb{Z}) = H_1(I; \mathbb{Z}) = 0$, ie it is perfect and the second homology of the classifying space $B\tilde{I}$ is zero. A presentation of \tilde{I} is given by

$$\tilde{I} = \langle s, t \mid (st)^2 = s^3 = t^5 \rangle.$$

In fact, if we construct a cell complex X of dimension 2 using the presentation above, then X has one 0-cell, two 1-cells and two 2-cells. The cellular chain complex, as computed from the presentation, is given by

$$0 \rightarrow \mathbb{Z}^2 \xrightarrow{\begin{pmatrix} -1 & 2 \\ 3 & -5 \end{pmatrix}} \mathbb{Z}^2 \xrightarrow{0} \mathbb{Z} \rightarrow 0.$$

Therefore, $H_1(X) = H_2(X) = 0$ while $\pi_1(X) = \tilde{I}$.

In contrast, consider the cut locus C of the 0–sphere in $SU(2)/\tilde{I}$, the Poincaré homology sphere. As $SU(2)$ is real analytic, so is the homology sphere. By Proposition 3.25, C is a finite, connected simplicial complex of dimension 2 such that $\pi_1(C) \cong \tilde{I}$ and $H_\bullet(C; \mathbb{Z}) \cong H_\bullet(S^2; \mathbb{Z})$. The existence of this space is interesting for the following reason: although $X \vee S^2$ has the same topological invariants we are unable to determine whether $X \vee S^2$ is homotopy equivalent to C .

In the codimension two case, we have two results.

Theorem 3.27 *Let Σ be a closed, orientable, real analytic surface of genus g and N a nonempty finite subset. Then $\text{Cu}(N)$ is a connected graph, homotopy equivalent to a wedge product of $|N| + 2g - 1$ circles.*

Proof As $M - N$ is connected, Lemma 3.18 implies that $\text{Cu}(N)$ is connected. It follows from Theorem 3.9 that $\text{Cu}(N)$ is a finite 1–dimensional simplicial complex, ie a finite graph. In this case, $\text{Th}(v)$ is a wedge product of $|N|$ copies of S^2 (see (3-4)). We consider (3-6) with $j = 2$:

$$0 \xrightarrow{i_*} \mathbb{Z} \xrightarrow{q} \tilde{H}_2(\vee_{|N|} S^2) \xrightarrow{\partial} H_1(\text{Cu}(N)) \xrightarrow{i_*} H_1(\Sigma) \rightarrow 0.$$

Note that $H_{d-1}(M)$ is torsion-free, whence all the groups appearing in the long exact sequence are free abelian groups. This implies that

$$\dim_{\mathbb{Z}} H_1(\text{Cu}(N)) = 2g + |N| - 1.$$

As $\text{Cu}(N)$ is a connected finite graph, collapsing a maximal tree T results in a quotient space $\text{Cu}(N)/T$ which is homotopic to $\text{Cu}(N)$ as well as being a wedge product of $|N| + 2g - 1$ circles. □

Remark 3.28 Itoh and Vîlcu [15] proved that every finite, connected graph can be realized as the cut locus (of a point) of some surface. There remains the question of orientability of the surface. As noted in the proof of Theorem 3.27, if the surface is orientable and $|N| = 1$, then the graph has an even number of generating cycles. If Σ is nonorientable, then $\Sigma \cong (\mathbb{R}P^2)^{\#k}$ has nonorientable genus k and the oriented double cover of Σ has genus $g = k - 1$. Recall that $H_1(\Sigma) \cong \mathbb{Z}^{k-1} \oplus \mathbb{Z}_2$ and $H_2(\Sigma) = 0$. Looking at (3-6) with $j = 2$, we obtain

$$0 \rightarrow \mathbb{Z} \rightarrow H_1(\text{Cu}(p)) \rightarrow \mathbb{Z}^{k-1} \oplus \mathbb{Z}_2 \rightarrow 0.$$

Thus, $H_1(\text{Cu}(p)) \cong \mathbb{Z}^k$ as homology groups of graphs are free abelian. Let $B_\varepsilon(\text{Cu}(p))$ denote the ε -neighborhood of $\text{Cu}(p)$ in Σ . For ε sufficiently small, this is a surface such that $\overline{B_\varepsilon(\text{Cu}(p))}$ has one boundary component. The compact surface $B_\varepsilon(\text{Cu}(p))$ is reminiscent of ribbon graphs. The surface Σ can be obtained as the connect-sum of a disk centered at p and the closure of $B_\varepsilon(\text{Cu}(p))$. Therefore, nonorientability of Σ is equivalent to nonorientability of $B_\varepsilon(\text{Cu}(p))$. A similar observation appears in the unpublished work of Itoh and Vîlcu [14, Theorem 3.7].

Example 3.29 (homology spheres of codimension two) In continuation of Theorem 3.24, let $N \hookrightarrow S^{k+2}$ be a homology sphere of dimension $k \geq 1$. Since N has codimension two, $S^{k+2} - N$ is path-connected and so is $\text{Cu}(N)$. We are not assuming that the metric on S^{k+2} is real analytic. Using (3-8) and the long exact sequence in cohomology of (M, N) , we infer that $H_1(\text{Cu}(N)) \cong \mathbb{Z}$ and all higher homology groups vanish. However, the Hurewicz theorem cannot be used here to establish that $\pi_1(\text{Cu}(N)) \cong \mathbb{Z}$.

In particular cases, we may conclude that $\text{Cu}(N)$ is homotopic to a circle. It was proved by Plotnick [22] that certain homology 3-spheres N , obtained by a Dehn surgery of type $1/2a$ on a knot, smoothly embed in S^5 with complement a homotopy circle. Since $M - N$ deforms to $\text{Cu}(N)$, it follows that there is a map $\alpha: S^1 \rightarrow \text{Cu}(N)$ inducing isomorphisms on homotopy and homology groups.

If $k = 1$, then a homology 1-sphere is just a knot K in S^3 . Since $S^3 - K$ deforms to $\text{Cu}(K)$, the fundamental group of the cut locus is the knot group. Moreover, in the case of real analytic knots in S^3 , the cut locus is a finite simplicial complex of dimension at most 2 (see Theorem 3.9). Except for the unknot, the knot group is never a free group, while the fundamental group of a connected, finite graph is free. This observation establishes that $\text{Cu}(K)$ is always a 2-dimensional simplicial complex whenever K is a nontrivial (real analytic) knot in S^3 .

3.3 Morse–Bott function associated to distance function

We first prove that the closure of $\text{Se}(N)$ is the cut locus, closely following the proof given in [29] for the case of a point.

Theorem 3.30 *Let $\text{Cu}(N)$ be the cut locus of a compact submanifold N of a complete Riemannian manifold M . The subset $\text{Se}(N)$ of $\text{Cu}(N)$ is dense in $\text{Cu}(N)$.*

Proof Let $q \in \text{Cu}(N)$ but not in $\text{Se}(N)$. Choose an N -geodesic γ , joining N to q , such that any extension of γ is not an N -geodesic. This geodesic γ is unique as $q \notin \text{Se}(N)$.

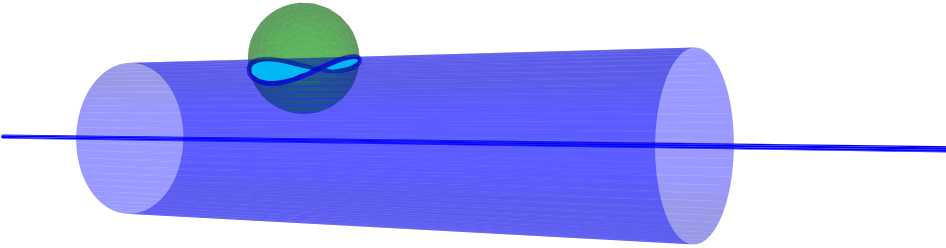


Figure 5: $\text{Co}(x_0, \delta)$.

We may write $\gamma(t) = \exp_v(tx)$, where $\gamma(0) = p \in N$ and $\gamma'(0) = x_0 \in S(v_p)$. It follows from the definition of s that $q = \exp_v(s(x_0)x_0)$. We need to show that every neighborhood of q in $\text{Cu}(N)$ must intersect $\text{Se}(N)$. Suppose it is false. Let $\delta > 0$ and consider $\overline{B}(x_0, \delta)$, the closed ball with center x_0 and radius δ . Define the cone

$$\text{Co}(x_0, \delta) := \{tx : 0 \leq t \leq 1, x \in \overline{B}(x_0, \delta) \cap S(v)\};$$

see Figure 5. Since $\overline{B}(x_0, \delta) \cap S(v)$ is homeomorphic to a closed $(n-1)$ -ball for sufficiently small δ , the cone will be homeomorphic to a closed Euclidean n -ball. Similarly, define another cone

$$\text{Co}^*(x_0, \delta) := \left\{ s\left(\frac{x}{\|x\|}\right)x \mid x \in \text{Co}(x_0, \delta), x \neq 0 \right\} \cup \{0\}.$$

Note that $s(x_0)$ is finite. As s is continuous, due to Proposition 3.14, for sufficiently small δ the term $s(x/\|x\|)$ is still finite, whence $\text{Co}^*(x_0, \delta)$ is well defined. We claim that $\text{Co}^*(x_0, \delta)$ is also homeomorphic to a closed Euclidean $(n-k)$ -ball. Indeed, a nonzero $x \in \text{Co}(x_0, \delta)$ implies $x = \lambda \hat{x}$ for some $\lambda \in (0, 1]$ and $\hat{x} \in \overline{B}(x_0, \delta) \cap S(v)$. Since $s(\hat{x})x = \lambda s(\hat{x})\hat{x}$, it follows that $\text{Co}^*(x_0, \delta)$ is the cone of the set

$$\{s(\hat{x})\hat{x} \mid \hat{x} \in \overline{B}(x_0, \delta) \cap S(v)\},$$

which is homeomorphic to $\overline{B}(x_0, \delta) \cap S(v)$. Now we have a dichotomy:

- (a) for a fixed small $\delta > 0$, the restriction of \exp_v to $\text{Co}^*(x_0, \delta)$ is a homeomorphism to its image because it is injective, or
- (b) for any $\delta > 0$, the restriction of \exp_v to $\text{Co}^*(x_0, \delta)$ is not injective.

If (b) holds, choose $v_n \neq w_n \in \text{Co}^*(x_0, 1/n)$ such that these map to q_n under \exp_v . Thus, $q_n \in \text{Se}(N)$ and compactness of $S(v)$ ensures that q_n converges to q . If (a) holds, then let $B(q, \varepsilon)$ denote the open ball in M centered at q with radius $\varepsilon > 0$. We claim that it intersects the complement of $\exp_v(\text{Co}^*(x_0, \delta))$ in M . But it is true as $s(x_0)x_0$ lies on the boundary of $\text{Co}^*(x_0, \delta)$ and hence it has a neighborhood in $\text{Co}^*(x_0, \delta)$

which is homeomorphic to a closed n -dimensional Euclidean half plane. Since \exp_ν restricted to $\text{Co}^*(x_0, \delta)$ is a homeomorphism, the open ball $B(q, \varepsilon)$ must intersect the points outside the image of $\exp_\nu(\text{Co}^*(x_0, \delta))$.

Now take $\varepsilon = 1/n$. For each n , there exists $q_n \in B(q, 1/n)$ with $q_n \notin \exp_\nu(\text{Co}^*(x_0, \delta))$. Since M is complete, for each point q_n let γ_n be an N -geodesic joining $p_n \in N$ to q_n . We may invoke the following result from Busemann’s book [5, Theorem 5.16, page 24]. Let $\{\gamma_n\}$ be a sequence of rectifiable curves in a finitely compact set X such that the lengths $\ell(\gamma_n)$ are bounded. If the initial points p_n of γ_n form a bounded set, then $\{\gamma_n\}$ contains a subsequence γ_{n_k} which converges uniformly to a rectifiable curve $\tilde{\gamma}$ in X and

$$\ell(\tilde{\gamma}) \leq \liminf \ell(\gamma_{n_k}).$$

Since $\{p_n\}$ lie in the compact set N , we obtain a rectifiable curve $\tilde{\gamma}$ such that

$$\ell(\tilde{\gamma}) \leq \liminf \ell(\gamma_{n_k}) = \lim_k \ell(\gamma_{n_k}) = \lim_k d_N(q_{n_k}) = d_N(q).$$

Thus, $\tilde{\gamma}$ is actually an N -geodesic joining $p' = \lim_k p_{n_k}$ to q and the unit tangent vectors $x_{n_k} = \gamma'_{n_k}(0)$ at p_{n_k} converges to the unit tangent vector $\tilde{x} = \tilde{\gamma}'(0)$ at p' . Since x_0 is an interior point of the set $\overline{B(x_0, \delta)} \cap S(\nu)$, any sequence in $S(\nu)$ converging to x_0 must eventually lie in $\text{Co}(x_0, \delta)$. According to our choice, $q_{n_k} \notin \exp_\nu(\text{Co}^*(x_0, \delta))$ and the x_{n_k} all lie outside of $\text{Co}(x_0, \delta)$. Hence, $x_0 \neq \tilde{x}$ and $\gamma \neq \tilde{\gamma}$. Thus, there are two distinct N -geodesics γ and $\tilde{\gamma}$ joining N to q , a contradiction to $q \notin \text{Se}(N)$. \square

We have seen (in Lemma 3.7) that d_N^2 is smooth away from the cut locus. It follows from Theorem 3.30 that the cut locus is the closure of the singularity of d_N^2 . The following example suggests that d_N^2 can be differentiable at points in $\text{Cu}(N) - \text{Se}(N)$ but not twice differentiable:

Example 3.31 (cut locus of an ellipse) We discuss the regularity of the distance-squared function from an ellipse $x^2/a^2 + y^2/b^2 = 1$ (with $a > b > 0$) in \mathbb{R}^2 . For a discussion of the cut locus for ellipses inside \mathbb{S}^2 and ellipsoids, see Hebda [10, pages 90–91]. Let (x_0, y_0) be a point inside the ellipse lying in the first quadrant. The point closest to (x_0, y_0) and lying on the ellipse is given by

$$x = \frac{a^2 x_0}{t + a^2}, \quad y = \frac{b^2 y_0}{t + b^2},$$

where t is the unique root of the quartic

$$\left(\frac{ax_0}{t+a^2}\right)^2 + \left(\frac{by_0}{t+b^2}\right)^2 = 1$$

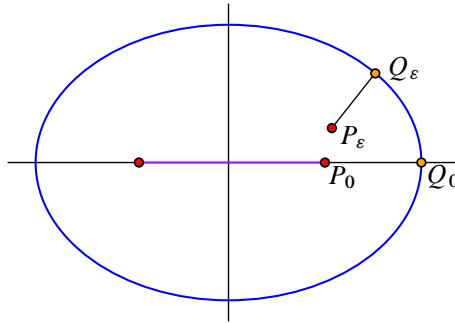


Figure 6: Cut locus of an ellipse.

in the interval $(-b^2, \infty)$. Given (α, β) with $\beta > 0$, we set

$$P_\varepsilon(\alpha, \beta) = \left(\frac{a^2 - b^2}{a} + \varepsilon\alpha, \varepsilon\beta \right);$$

this defines a straight line passing through $P_0(\alpha, \beta)$ in the direction of (α, β) . For $\varepsilon > 0$, $P_\varepsilon(\alpha, \beta)$ lies in the first quadrant and we denote by $t = t(\varepsilon)$ the unique relevant root of the quartic

$$\left(\frac{a((a^2 - b^2)/a + \varepsilon\alpha)}{t + a^2} \right)^2 + \left(\frac{b\varepsilon\beta}{t + b^2} \right)^2 = 1.$$

Simplifying this after dividing by ε and taking a limit $\varepsilon \rightarrow 0^+$, we obtain

$$\frac{2a\alpha}{a^2 - b^2} = \lim_{\varepsilon \rightarrow 0^+} \left(\left(\frac{2}{a^2 - b^2} \right) \frac{t + b^2}{\varepsilon} - b^2\beta^2 \frac{\varepsilon}{(t + b^2)^2} \right).$$

On the other hand, the point $Q_\varepsilon(\alpha, \beta)$ on the ellipse closest to $P_\varepsilon(\alpha, \beta)$ is given by

$$x_\varepsilon = \frac{a^2((a^2 - b^2)/a + \varepsilon\alpha)}{t + a^2}, \quad y_\varepsilon = \frac{b^2\varepsilon\beta}{t + b^2}.$$

It follows that

$$(3-10) \quad d_\varepsilon^2(\alpha, \beta) := d^2(P_\varepsilon, Q_\varepsilon) = \frac{t^2}{a^2} \left(\frac{a^2 - b^2 + a\varepsilon\alpha}{t + a^2} \right)^2 + \frac{t^2}{b^2} \left(\frac{b\varepsilon\beta}{t + b^2} \right)^2.$$

Using $t(0) = -b^2$, simplifications lead us to

$$\begin{aligned} \lim_{\varepsilon \rightarrow 0^+} \frac{d_\varepsilon^2 - d_0^2}{\varepsilon} &= \frac{2ab^4\alpha}{a^2(a^2 - b^2)} - \lim_{\varepsilon \rightarrow 0^+} \left(\frac{(t + b^2)(a^2b^2 - a^2t + 2b^2t)}{\varepsilon(t + a^2)^2} - \beta^2 \frac{t^2\varepsilon}{(t + b^2)^2} \right) \\ &= \frac{2ab^4\alpha}{a^2(a^2 - b^2)} - \frac{2b^2}{a^2 - b^2} \lim_{\varepsilon \rightarrow 0} \frac{t + b^2}{\varepsilon} + \beta^2 b^4 \lim_{\varepsilon \rightarrow 0} \frac{\varepsilon}{(t + b^2)^2} \\ &= \frac{2ab^4\alpha}{a^2(a^2 - b^2)} - \frac{2ab^2\alpha}{a^2 - b^2} = -\frac{2b^2\alpha}{a}. \end{aligned}$$

On the other hand, for $\varepsilon < 0$, the point $P_\varepsilon(\alpha, \beta)$ lies in the fourth quadrant. By symmetry, the distance between $P_\varepsilon(\alpha, \beta)$ and $Q_\varepsilon(\alpha, \beta)$ is the same as that between $P_{-\varepsilon}(-\alpha, \beta)$ and $Q_{-\varepsilon}(-\alpha, \beta)$. However, it is seen that

$$d^2(P_{-\varepsilon}(-\alpha, \beta), Q_{-\varepsilon}(-\alpha, \beta)) = d_{-\varepsilon}^2(-\alpha, \beta),$$

as defined in (3-10). Therefore,

$$\begin{aligned} \lim_{\varepsilon \rightarrow 0^-} \frac{d^2(P_\varepsilon(\alpha, \beta), Q_\varepsilon(\alpha, \beta)) - d^2(P_0(\alpha, \beta), Q_0(\alpha, \beta))}{\varepsilon} &= \lim_{\varepsilon \rightarrow 0^-} \frac{d_{-\varepsilon}^2(-\alpha, \beta) - d_0^2(-\alpha, \beta)}{\varepsilon} \\ &= - \lim_{-\varepsilon \rightarrow 0^+} \frac{d_{-\varepsilon}^2(-\alpha, \beta) - d_0^2(-\alpha, \beta)}{-\varepsilon} \\ &= - \frac{2b^2\alpha}{a}, \end{aligned}$$

where the last equality follows from the right-hand derivative of d^2 , as computed previously.

When $\beta = 0$, we would like to compute $d_\varepsilon^2(\alpha, 0)$. If $\varepsilon > 0$, then

$$(3-11) \quad d_\varepsilon^2(\alpha, 0) = (b^2/a - \varepsilon\alpha)^2 = \frac{b^4}{a^2} - \frac{2b^2\alpha\varepsilon}{a} + \alpha^2\varepsilon^2.$$

On the other hand, if $\varepsilon < 0$ is sufficiently small, then there are two points on the ellipse closest to $P_\varepsilon(\alpha, 0) = ((a^2 - b^2)/a + \varepsilon\alpha, 0)$, with exactly one on the first quadrant, say Q_ε . Since the segment $P_\varepsilon Q_\varepsilon$ must be orthogonal to the tangent to the ellipse at Q_ε , we obtain the coordinates for Q_ε :

$$x_\varepsilon = \frac{a^2((a^2 - b^2)/a + \varepsilon\alpha)}{a^2 - b^2}, \quad y_\varepsilon^2 = b^2 \left(1 - \frac{x_\varepsilon^2}{a^2} \right), \quad y_\varepsilon > 0.$$

We may compute the distance

$$(3-12) \quad d_\varepsilon^2(\alpha, 0) := d^2(P_\varepsilon, Q_\varepsilon) = \frac{b^4}{a^2} - \frac{2b^2\alpha\varepsilon}{a} - \frac{b^2\alpha^2\varepsilon^2}{a^2 - b^2},$$

where $\varepsilon < 0$. Combining (3-11) and (3-12), we conclude that d^2 is differentiable at $P_0 = ((a^2 - b^2)/a, 0)$, a point in $\text{Cu}(N)$ but not in $\text{Se}(N)$. However, comparing the quadratic part of d^2 in (3-11)–(3-12), we conclude that d^2 is not twice differentiable at P_0 .

Theorem 3.32 *Let N be a closed embedded submanifold of a complete Riemannian manifold M . Let $d_N : M \rightarrow \mathbb{R}$ be the distance function with respect to N . If $f = d_N^2$, then its restriction to $M - \text{Cu}(N)$ is a Morse–Bott function, with N as the critical submanifold. Moreover, the gradient flow of f deforms $M - \text{Cu}(N)$ to N .*

Proof It follows from Lemma A.2 that the map $\exp_v^{-1} : M - (\text{Cu}(N) \cup N) \rightarrow v - \{0\}$ is an (into) diffeomorphism and $d_N(q) = \|\exp_v^{-1}(q)\|$ and hence the distance function is of class C^∞ at $q \in M - (\text{Cu}(N) \cup N)$. Using Fermi coordinates (see Proposition 3.5), we have seen that the distance-squared function is smooth around N and therefore it is smooth on $M - \text{Cu}(N)$. By Corollary 3.6, the Hessian of this function at N is nondegenerate in the normal direction. It is well known [26, Proposition 4.8] that $\|\nabla d(q)\| = 1$ if d_N is differentiable at $q \in M$. Thus, for $q \in M - (\text{Cu}(N) \cup N)$, we have

$$(3-13) \quad \|\nabla f(q)\| = 2d_N(q)\|\nabla d_N(q)\| = 2d_N(q).$$

Let γ be the unique unit-speed N -geodesic that joins N to q , ie

$$\gamma : [0, d_N(q)] \rightarrow M, \quad \gamma(0) = p, \quad \gamma(d_N(q)) = q, \quad \|\gamma'\| = 1.$$

We may write $\nabla f(q) = \lambda\gamma'(d_N(q)) + w$, where w is orthogonal to $\gamma'(d_N(q))$. But

$$\begin{aligned} \langle \nabla f|_q, \gamma'(d_N(q)) \rangle &= \left. \frac{d}{dt} f(\gamma(d_N(q) + t)) \right|_{t=0} = \left. \frac{d}{dt} (d_N(q)^2 + 2d_N(q)t + t^2) \right|_{t=0} \\ &= 2d_N(q). \end{aligned}$$

Thus, $\lambda = 2d(q)$ and, combined with (3-13), we conclude that

$$\nabla f(q) = 2d_N(q)\gamma'(d_N(q)).$$

Therefore, the negative gradient flow line initialized at $q \in M - \text{Cu}(N)$ is given by

$$\eta(t) = \gamma(d_N(q)e^{-2t}).$$

These flow lines define a flow which deforms $M - \text{Cu}(N)$ to N in infinite time. □

The reader may choose to revisit the example of $\text{GL}(n, \mathbb{R})$ discussed in Section 2.2 and treat it as a concrete illustration of the theorem above.

4 Applications to Lie groups

Due to classical results of Cartan, Iwasawa and others, we know that any connected Lie group G is diffeomorphic to the product of a maximally compact subgroup K and an Euclidean space. In particular, G deforms to K . For semisimple groups, this decomposition is stronger and is attributed to Iwasawa. The Killing form on the Lie algebra \mathfrak{g} is nondegenerate and negative-definite for compact semisimple Lie algebras.

For such a Lie group G , consider the Levi-Civita connection associated to the bi-invariant metric obtained from the negative of the Killing form. This connection coincides with the Cartan connection.

We consider two examples, both of which are noncompact and nonsemisimple. We prove that these Lie groups G deformation retract to maximally compact subgroups K via gradient flows of appropriate Morse–Bott functions. This requires a choice of a left-invariant metric which is right- K -invariant and a careful analysis of the geodesics associated with the metric. In particular, we provide a possibly new proof of the surjectivity of the exponential map for $U(p, q)$.

4.1 Invertible matrices with positive determinant

Let g be a left-invariant metric on $GL(n, \mathbb{R})$, the set of invertible matrices. Recall that a left-invariant metric g on a Lie group is determined by its restriction at the identity. For $A \in GL(n, \mathbb{R})$, consider the left multiplication map $l_A: GL(n, \mathbb{R}) \rightarrow GL(n, \mathbb{R})$, $B \mapsto AB$. This extends to a linear isomorphism from $M(n, \mathbb{R})$ to itself. Thus, the differential $(Dl_A)_I: T_I GL(n, \mathbb{R}) \rightarrow T_A GL(n, \mathbb{R})$ is an isomorphism and given by l_A itself. For $X, Y \in T_I GL(n, \mathbb{R})$,

$$g_I(X, Y) = g_A((Dl_A)_I X, (Dl_A)_I Y) = g_A(AX, AY).$$

We choose the left-invariant metric on $GL(n, \mathbb{R})$ generated by the Euclidean metric at I . Therefore,

$$g_{A^{-1}}(X, Y) = \langle AX, AY \rangle_I := \text{tr}((AX)^T AY) = \text{tr}(X^T A^T AY).$$

Note that this metric is right- $O(n, \mathbb{R})$ -invariant. We are interested in the distance between an invertible matrix A (with $\det(A) > 0$) and $SO(n, \mathbb{R})$. Since $SO(n, \mathbb{R})$ is compact, there exists $B \in SO(n, \mathbb{R})$ such that $d(A, B) = d_{SO(n, \mathbb{R})}(A)$.

Lemma 4.1 *If D is a diagonal matrix with positive diagonal entries $\lambda_1, \dots, \lambda_n$, then*

$$d_{SO(n, \mathbb{R})}(D) = d(D, I).$$

Moreover, I is the unique minimizer and the associated minimal geodesic is given by $\gamma(t) = e^{t \log D}$.

Proof Choose $B \in SO(n, \mathbb{R})$ satisfying $d(A, B) = d_{SO(n, \mathbb{R})}(A)$. Since, with respect to the left-invariant metric, $GL^+(n, \mathbb{R})$ is complete, there exists a minimal geodesic $\gamma: [0, 1] \rightarrow GL^+(n, \mathbb{R})$ joining B to D , ie

$$\gamma(0) = B, \quad \gamma(1) = D \quad \text{and} \quad \ell(\gamma) = d(D, B).$$

The first variational principle implies that $\gamma'(0)$ is orthogonal to $T_B\text{SO}(n, \mathbb{R})$. It follows from Martin and Neff [19, Section 2.1] that $\eta(t) = e^{tW}$ is a geodesic if W is a symmetric matrix. Moreover, $\eta'(0) = W$ is orthogonal to $T_I\text{SO}(n, \mathbb{R})$. As left translation is an isometry and isometry preserves geodesic, it follows that $\gamma(t) = Be^{tW}$ is a geodesic with $\gamma'(0)$ orthogonal to $T_B\text{SO}(n, \mathbb{R})$. By the defining properties of γ , $D = \gamma(1) = Be^W$. Since e^W is symmetric positive-definite, we obtain two polar decompositions of D , ie $D = ID$ and $D = Be^W$. By the uniqueness of the polar decomposition for invertible matrices, $B = I$ and $D = e^W$.

In order to compute $d(I, D)$, note that

$$e^W = D = e^{\log D},$$

where $\log D$ denotes the diagonal matrix with entries $\log \lambda_1, \dots, \log \lambda_n$. As W and $\log D$ are symmetric, and matrix exponential is injective on the space of symmetric matrices, we conclude that $W = \log D$. The geodesic is given by $\gamma(t) = e^{t \log D}$ and

$$(4-1) \quad d_{\text{SO}(n, \mathbb{R})}(D) = \|\gamma'(0)\|_I = \|\log D\|_I = \left(\sum_{i=1}^n (\log \lambda_i)^2 \right)^{1/2}.$$

Thus, the distance-squared function will be given by $\sum_{i=1}^n (\log \lambda_i)^2$. □

Now, for any $A \in \text{GL}^+(n, \mathbb{R})$, we can apply the SVD decomposition, ie $A = UDV^T$ with $\sqrt{A^T A} = VDV^T$ and $\log \sqrt{A^T A} = V(\log D)V^T$. Note that $U, V \in \text{SO}(n, \mathbb{R})$ and D is a diagonal matrix with positive entries. The left-invariant metric is right-invariant with respect to orthogonal matrices. Thus,

$$d_{\text{SO}(n, \mathbb{R})}(A) = d_{\text{SO}(n, \mathbb{R})}(D) = \|\log D\|_I,$$

where the last equality follows from the lemma (see (4-1)). As

$$\|\log D\|_I = \|V(\log D)V^T\|_I = \|\log \sqrt{A^T A}\|_I,$$

It follows from the arguments of the lemma and the metric being bi- $O(n, \mathbb{R})$ -invariant that

$$\gamma(t) = Ue^{t \log D}V^T$$

is a minimal geodesic joining UV^T to A , realizing $d_{\text{SO}(n, \mathbb{R})}(A)$. As the minimizer UV^T is unique, $\text{Se}(\text{SO}(n, \mathbb{R}))$ is empty, implying that $\text{Cu}(\text{SO}(n, \mathbb{R}))$ is empty as well. In fact, $UV^T = A\sqrt{A^T A}^{-1}$ and

$$(4-2) \quad \gamma(t) = Ue^{t \log D}V^T = UV^T V e^{t \log D} V^T = A\sqrt{A^T A}^{-1} e^{t \log \sqrt{A^T A}}.$$

If we compare (2-6)—the deformation of $GL(n, \mathbb{R})$ to $O(n, \mathbb{R})$ inside $M(n, \mathbb{R})$ —with (4-2), then, in both cases, an invertible matrix A deforms to $A\sqrt{A^T A}^{-1}$. Finally, observe that the normal bundle of $SO(n, \mathbb{R})$ is diffeomorphic to $GL^+(n, \mathbb{R})$.

4.2 Indefinite unitary groups

Let n be a positive integer with $n = p + q$. Consider the inner product on \mathbb{C}^n given by

$$\langle (w_1, \dots, w_n), (z_1, \dots, z_n) \rangle = z_1 \bar{w}_1 + \dots + z_p \bar{w}_p - z_{p+1} \bar{w}_{p+1} - \dots - z_n \bar{w}_n.$$

This is given by the matrix $I_{p,q}$ as

$$\langle \mathbf{w}, \mathbf{z} \rangle = \bar{\mathbf{w}}^t I_{p,q} \mathbf{z} = (\bar{w}_1 \ \dots \ \bar{w}_n) \begin{pmatrix} I_p & 0 \\ 0 & -I_q \end{pmatrix} \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}.$$

Let $U(p, q)$ denote the subgroup of $GL(n, \mathbb{C})$ preserving this indefinite form, ie $\mathcal{A} \in U(p, q)$ if and only if $\mathcal{A}^* I_{p,q} \mathcal{A} = I_{p,q}$. In particular, $\det \mathcal{A}$ is a complex number of unit length. By convention, $I_{n,0} = I_n$ and $I_{0,n} = -I_n$, both of which correspond to $U(n, 0) = U(n) = U(0, n)$, the unitary group. In all other cases, the inner product is indefinite.

The group $U(1, 1)$ is given by matrices of the form

$$\mathcal{A} = \begin{pmatrix} \alpha & \beta \\ \lambda \bar{\beta} & \lambda \bar{\alpha} \end{pmatrix} \quad \text{with } \lambda \in S^1 \text{ and } |\alpha|^2 - |\beta|^2 = 1.$$

More generally, we shall use

$$\mathcal{A} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

to denote an element of $U(p, q)$. It follows from the definition that $\mathcal{A} \in U(p, q)$ if and only if

$$A^*A - C^*C = I_p, \quad A^*B - C^*D = 0_{p \times q}, \quad B^*B - D^*D = -I_q.$$

Observe that, if $Av = 0$, then

$$0 = A^*Av = C^*Cv + v,$$

which implies that C^*C , a positive semidefinite matrix, has -1 as an eigenvalue unless $v = 0$. Therefore, A is invertible and the same argument works for D .

Lemma 4.2 *The intersection of $U(p + q)$ with $U(p, q)$ is $U(p) \times U(q)$. Moreover, if $\mathcal{A} \in U(p, q)$, then \mathcal{A}^* , $\sqrt{\mathcal{A}^* \mathcal{A}} \in U(p, q)$.*

Proof If $\mathcal{A} \in U(p) \times U(q)$, then

$$\mathcal{A}^* \mathcal{A} + C^* C = I_p, \quad B^* B + D^* D = I_q.$$

This implies that both B and C are zero matrices. If $\mathcal{A} \in U(p, q)$, then $\mathcal{A}^* = I_{p,q} \mathcal{A}^{-1} I_{p,q}$ and

$$\begin{aligned} (\mathcal{A}^* \mathcal{A})^* I_{p,q} (\mathcal{A}^* \mathcal{A}) &= (\mathcal{A}^* \mathcal{A}) I_{p,q} (\mathcal{A}^* \mathcal{A}) = I_{p,q} \mathcal{A}^{-1} I_{p,q} \mathcal{A} I_{p,q} I_{p,q} \mathcal{A}^{-1} I_{p,q} \mathcal{A} \\ &= I_{p,q} = \mathcal{A}^* I_{p,q} \mathcal{A}. \end{aligned}$$

This also implies that $\mathcal{A} I_{p,q} \mathcal{A}^* = I_{p,q}$.

All the eigenvalues of $\mathcal{A}^* \mathcal{A}$ are positive. Moreover, if λ is an eigenvalue of $\mathcal{A}^* \mathcal{A}$ with eigenvector $\mathbf{v} = (v_1, \dots, v_p, v_{p+1}, \dots, v_n)$, then

$$I_{p,q} \mathbf{v} = \mathcal{A}^* \mathcal{A} I_{p,q} \mathcal{A}^* \mathcal{A} \mathbf{v} = \lambda (\mathcal{A}^* \mathcal{A} I_{p,q} \mathbf{v}),$$

which implies that λ^{-1} is also an eigenvalue with eigenvector

$$\mathbf{v}' = (v_1, \dots, v_p, -v_{p+1}, \dots, -v_n).$$

If $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ is an eigenbasis of $\mathcal{A}^* \mathcal{A}$ with (possibly repeated) eigenvalues $\lambda_1, \dots, \lambda_n$, then

$$\sqrt{\mathcal{A}^* \mathcal{A}} I_{p,q} \sqrt{\mathcal{A}^* \mathcal{A}} \mathbf{v}_j = \sqrt{\mathcal{A}^* \mathcal{A}} I_{p,q} \sqrt{\lambda_j} \mathbf{v}_j = \sqrt{\lambda_j} \sqrt{\mathcal{A}^* \mathcal{A}} \mathbf{v}'_j = \mathbf{v}'_j = I_{p,q} \mathbf{v}_j.$$

Thus, $\sqrt{\mathcal{A}^* \mathcal{A}}$ satisfies the defining relation for a matrix to be in $U(p, q)$. □

We may use the polar decomposition (for matrices in $GL(n, \mathbb{C})$) to write

$$\mathcal{A} = U |\mathcal{A}|, \quad \text{where } U = \mathcal{A} \sqrt{\mathcal{A}^* \mathcal{A}}^{-1} \text{ and } |\mathcal{A}| = \sqrt{\mathcal{A}^* \mathcal{A}},$$

where $U, |\mathcal{A}| \in U(p, q)$. For $U(1, 1)$, this decomposition takes the form

$$\begin{pmatrix} \alpha & \beta \\ \lambda \bar{\beta} & \lambda \bar{\alpha} \end{pmatrix} = \begin{pmatrix} \alpha/|\alpha| & 0 \\ 0 & \lambda \bar{\alpha}/|\alpha| \end{pmatrix} \begin{pmatrix} |\alpha| & |\alpha| \beta/\alpha \\ |\alpha| \bar{\beta}/\bar{\alpha} & |\alpha| \end{pmatrix}.$$

The Lie algebra $\mathfrak{u}_{p,q}$ is given by matrices $X \in M_n(\mathbb{C})$ such that

$$X^* I_{p,q} + I_{p,q} X = 0.$$

This is a real Lie subalgebra of $M_{p+q}(\mathbb{C})$. It contains the subalgebras \mathfrak{u}_p and \mathfrak{u}_q as Lie algebras of the subgroups $U(p) \times I_q$ and $I_p \times U(q)$. Consider the inner product

$$\langle \cdot, \cdot \rangle: \mathfrak{u}_{p,q} \times \mathfrak{u}_{p,q} \rightarrow \mathbb{R}, \quad \langle X, Y \rangle := \text{tr}(X^* Y).$$

Lemma 4.3 *The inner product is symmetric and positive-definite.*

Proof Note that

$$\langle X, Y \rangle = \text{tr}(-I_{p,q} X I_{p,q} Y) = \text{tr}(-I_{p,q} Y I_{p,q} X) = \langle Y, X \rangle.$$

Since $\overline{\langle X, Y \rangle} = \langle Y, X \rangle$ due to the invariance of trace under transpose, we conclude that the inner product is real and symmetric. It is positive-definite as $\langle X, X \rangle = \text{tr}(X^* X) \geq 0$ and equality holds if and only if X is the zero matrix. \square

The Riemannian metric obtained by left translations of $\langle \cdot, \cdot \rangle$ will also be denoted by $\langle \cdot, \cdot \rangle$. We shall analyze the geodesics for this metric. The Lie algebra $\mathfrak{u}_p \oplus \mathfrak{u}_q$ of $U(p) \times U(q)$ consists of matrices

$$\begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix} \quad \text{with } A + A^* = 0 \text{ and } D + D^* = 0.$$

Let \mathfrak{n} denote the orthogonal complement of $\mathfrak{u}_p \oplus \mathfrak{u}_q$ inside $\mathfrak{u}_{p,q}$. As \mathfrak{n} is of (complex) dimension pq and

$$\left\{ \begin{pmatrix} 0 & B \\ B^* & 0 \end{pmatrix} \mid B \in M_{p,q}(\mathbb{C}) \right\}$$

is contained in \mathfrak{n} , this is all of it. We may verify that

$$\begin{aligned} \left[\begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix}, \begin{pmatrix} 0 & B \\ B^* & 0 \end{pmatrix} \right] &= \begin{pmatrix} 0 & AB - BD \\ DB^* - B^*A & 0 \end{pmatrix} \in \mathfrak{n}, \\ \left[\begin{pmatrix} 0 & B \\ B^* & 0 \end{pmatrix}, \begin{pmatrix} 0 & C \\ C^* & 0 \end{pmatrix} \right] &= \begin{pmatrix} BC^* - CB^* & 0 \\ 0 & B^*C - C^*B \end{pmatrix} \in \mathfrak{u}_p \oplus \mathfrak{u}_q. \end{aligned}$$

Lemma 4.4 *Let γ be the integral curve, initialized at e , for a left-invariant vector field Y . This curve is a geodesic if $Y(e)$ belongs either to \mathfrak{n} or to $\mathfrak{u}_p \oplus \mathfrak{u}_q$.*

Proof The Levi-Civita connection ∇ is given by the Koszul formula

$$2\langle X, \nabla_Z Y \rangle = Z\langle X, Y \rangle + Y\langle X, Z \rangle - X\langle Y, Z \rangle + \langle Z, [X, Y] \rangle + \langle Y, [X, Z] \rangle - \langle X, [Y, Z] \rangle.$$

Putting $Z = Y$ and $Z = X$, two left-invariant vector fields, in the above, we obtain

$$\langle X, \nabla_Y Y \rangle = \langle Y, [X, Y] \rangle.$$

To prove our claim, it suffices to show that $\nabla_Y Y = 0$, ie $\langle Y, [X, Y] \rangle = 0$ for any X . Let us assume that $Y(e) \in \mathfrak{n}$. If $X(e) \in \mathfrak{n}$, then $[X(e), Y(e)] \in \mathfrak{u}_p \oplus \mathfrak{u}_q$, which implies that

$\langle Y(e), [X(e), Y(e)] \rangle = 0$. If $X(e) \in \mathfrak{u}_p \oplus \mathfrak{u}_q$, then

$$\begin{aligned} \langle Y, [X, Y] \rangle &= \left\langle \begin{pmatrix} 0 & B \\ B^* & 0 \end{pmatrix}, \begin{pmatrix} 0 & AB - BD \\ DB^* - B^*A & 0 \end{pmatrix} \right\rangle \\ &= \text{tr} \begin{pmatrix} B(DB^* - B^*A) & 0 \\ 0 & B^*(AB - BD) \end{pmatrix} \\ &= \text{tr}(BDB^* - BB^*A) + \text{tr}(B^*AB - B^*BD) \\ &= 0 \end{aligned}$$

by the cyclic property of trace. Thus, $\nabla_Y Y = 0$ if $Y(e) \in \mathfrak{n}$; a similar proof works if $Y(e) \in \mathfrak{u}_p \oplus \mathfrak{u}_q$. □

Remark 4.5 An integral curve of a left-invariant vector field (also called one-parameter subgroups) need not be a geodesic in $U(p, q)$. For instance, if $X + Y$ is a left-invariant vector field given by $X(e) \in \mathfrak{u}_p \oplus \mathfrak{u}_q$ and $Y(e) \in \mathfrak{n}$, then $\nabla_{X+Y}(X + Y) = 0$ if and only if $\nabla_X Y = \frac{1}{2}[X, Y]$ and $\nabla_Y X = \frac{1}{2}[Y, X]$. This happens if and only if the metric is bi-invariant, ie

$$\langle [X, Z], Y \rangle = \langle X, [Z, Y] \rangle.$$

This is not true; for instance, for $X(e) \in \mathfrak{u}_p \oplus \mathfrak{u}_q$ and linearly independent $Y(e), Z(e) \in \mathfrak{n}$, we get $\langle [X, Z], Y \rangle - \langle X, [Z, Y] \rangle \neq 0$.

Consider the matrix

$$Y = \begin{pmatrix} 0 & B \\ B^* & 0 \end{pmatrix} \in \mathfrak{n}.$$

Let $B = U\sqrt{B^*B}$ and $B^* = \sqrt{B^*B}U^*$ be polar decompositions, where U and U^* are partial isometries. It follows from direct computation that

$$\begin{aligned} e^Y &= \begin{pmatrix} I_p + \frac{1}{2!}BB^* + \frac{1}{4!}(BB^*)^2 + \dots & B + \frac{1}{3!}B(B^*B) + \frac{1}{5!}B(B^*B)^2 + \dots \\ B^* + \frac{1}{3!}(B^*B)B^* + \frac{1}{5!}(B^*B)^2B^* + \dots & I_q + \frac{1}{2!}B^*B + \frac{1}{4!}(B^*B)^2 + \dots \end{pmatrix} \\ &= \begin{pmatrix} \cosh(\sqrt{BB^*}) & U \sinh(\sqrt{B^*B}) \\ \sinh(\sqrt{B^*B})U^* & \cosh(\sqrt{B^*B}) \end{pmatrix}. \end{aligned}$$

It can be checked that

$$e^{\mathfrak{n}} \cap (U(p) \times U(q)) = \{I_n\}.$$

It is known that the nonzero eigenvalues of Y are the nonzero eigenvalues of $\sqrt{BB^*}$ and their negatives.

Theorem 4.6 For any element $A \in U(p, q)$, the associated matrix $\sqrt{A^*A}$ can be expressed uniquely as e^Y for $Y \in \mathfrak{n}$. Moreover, there is a unique way to express A as a product of a unitary matrix and an element of $e^{\mathfrak{n}}$, and it is given by the polar decomposition.

In order to prove the result, we discuss some preliminaries on logarithms of complex matrices. In general, there is no unique logarithm. However, the Gregory series

$$\log A = - \sum_{m=0}^{\infty} \frac{2}{2m+1} [(I - A)(I + A)^{-1}]^{2m+1}$$

converges if all the eigenvalues of $A \in M_n(\mathbb{C})$ have positive real part; see Higham [11, Section 11.3, page 273]. In particular, $\log A$ is well defined for Hermitian positive-definite matrices. This is often called the *principal logarithm* of A . This logarithm satisfies $e^{\log A} = A$. There is an integral form of the logarithm that applies to matrices without real or zero eigenvalues; it is given by

$$\log A = (A - I) \int_0^1 [s(A - I) + I]^{-1} ds.$$

Lemma 4.7 The inverse of $A^*A + I_n$ for $A \in U(p, q)$ is given by

$$[A^*A + I_n]^{-1} = \frac{1}{2} \begin{pmatrix} I_p & -A^{-1}B \\ -B^*(A^*)^{-1} & I_q \end{pmatrix}.$$

Proof Since A^*A has only positive eigenvalues, $A^*A + I_n$ has no kernel. We note that

$$A^*A + I_n = \begin{pmatrix} 2C^*C + 2I_p & 2A^*B \\ 2B^*A & 2B^*B + 2I_q \end{pmatrix} = \begin{pmatrix} 2A^*A & 2A^*B \\ 2B^*A & 2D^*D \end{pmatrix}.$$

The inverse matrix satisfies

$$\begin{pmatrix} 2A^*A & 2A^*B \\ 2B^*A & 2D^*D \end{pmatrix} \begin{pmatrix} E & F \\ F^* & G \end{pmatrix} = \begin{pmatrix} I_p & 0 \\ 0 & I_q \end{pmatrix}.$$

As the matrices are Hermitian, the three constraints that E , F and G must satisfy (and are uniquely determined by) are

$$E = \frac{1}{2}(A^*A)^{-1} - A^{-1}BF^*, \quad G = \frac{1}{2}(D^*D)^{-1} - D^{-1}CF, \quad F = -A^{-1}BG.$$

We note that $E = \frac{1}{2}I_p$, $G = \frac{1}{2}I_q$ and $F = -\frac{1}{2}A^{-1}B$ satisfy the above equations. For instance,

$$\begin{aligned} \frac{1}{2}(A^*A)^{-1} - A^{-1}BF^* &= \frac{1}{2}(A^*A)^{-1} + \frac{1}{2}A^{-1}BB^*(A^*)^{-1} \\ &= \frac{1}{2}(A^*A)^{-1} + \frac{1}{2}A^{-1}(AA^* - I_p)(A^*)^{-1} = \frac{1}{2}I_p, \end{aligned}$$

where $BB^* = AA^* - I_p$ is a consequence of $\mathcal{A}^* \in U(p, q)$. Yet another consequence is $AC^* = BD^*$, which is equivalent to

$$A^{-1}B = (D^{-1}C)^*.$$

In a similar vein,

$$\begin{aligned} \frac{1}{2}(D^*D)^{-1} - D^{-1}CF &= \frac{1}{2}(D^*D)^{-1} + \frac{1}{2}D^{-1}CC^*(D^*)^{-1} \\ &= \frac{1}{2}(D^*D)^{-1} + \frac{1}{2}D^{-1}(DD^* - I_q)(D^*)^{-1} = \frac{1}{2}I_q, \end{aligned}$$

where $CC^* = DD^* - I_q$ is due to $\mathcal{A}^* \in U(p, q)$. □

Proof of Theorem 4.6 We use Gregory series expansion for computing the principal logarithm of $\mathcal{A}^*\mathcal{A}$ along with Lemma 4.7:

$$\begin{aligned} \log(\mathcal{A}^*\mathcal{A}) &= \sum_{m=0}^{\infty} \frac{2}{2m+1} \left[2 \begin{pmatrix} A^*A - I_p & A^*B \\ B^*A & D^*D - I_q \end{pmatrix} \frac{1}{2} \begin{pmatrix} I_p & -A^{-1}B \\ -B^*(A^*)^{-1} & I_q \end{pmatrix} \right]^{2m+1} \\ &= \sum_{m=0}^{\infty} \frac{2}{2m+1} \begin{pmatrix} 0 & A^{-1}B \\ B^*(A^*)^{-1} & 0 \end{pmatrix}^{2m+1}. \end{aligned}$$

We set $Y = \frac{1}{2} \log(\mathcal{A}^*\mathcal{A})$. It is clear that $Y \in \mathfrak{n}$ and $e^Y = \sqrt{\mathcal{A}^*\mathcal{A}}$. It is known that the exponential map is injective on Hermitian matrices. This implies the uniqueness of Y .

If $U_1e^{Y_1} = U_2e^{Y_2}$ are two decompositions of $\mathcal{A} \in U(p, q)$ with $U_i \in U(p) \times U(q)$ and $Y_i \in \mathfrak{n}$, then

$$e^{2Y_1} = e^{Y_1}U_1^*U_1e^{Y_1} = e^{Y_2}U_2^*U_2e^{Y_2} = e^{2Y_2}.$$

By the injectivity of the exponential map (on Hermitian matrices), we obtain $Y_1 = Y_2$, which implies that $U_1 = U_2$. □

We infer the following result (see Yakubovich and Starzhinskii [30, Lemma 1, page 211] for a different proof):

Corollary 4.8 *The exponential map $\exp: \mathfrak{u}_{p,q} \rightarrow U(p, q)$ is surjective.*

Proof Using the polar decomposition and Theorem 4.6,

$$A = \mathcal{A}\sqrt{\mathcal{A}^*\mathcal{A}}^{-1}\sqrt{\mathcal{A}^*\mathcal{A}} = \mathcal{A}\sqrt{\mathcal{A}^*\mathcal{A}}^{-1}e^Y.$$

Since the matrix exponential is surjective for $U(p) \times U(q)$, choose $Z \in \mathfrak{u}_p \oplus \mathfrak{u}_q$ such that $e^Z = \mathcal{A}\sqrt{\mathcal{A}^*\mathcal{A}}^{-1}$. By the Baker–Campbell–Hausdorff formula, we may express e^Ze^Y as the exponential of an element in $\mathfrak{u}_{p,q}$. □

The distance from any matrix $\mathcal{A} \in U(p, q)$ to $U(p) \times U(q)$ is given by the length of the curve

$$\gamma(t) = \mathcal{A} \sqrt{\mathcal{A}^* \mathcal{A}}^{-1} e^{tY},$$

which can be computed (and simplified via left-invariance) as

$$\ell(\gamma) = \int_0^1 \|\gamma'(t)\|_{\gamma(t)} dt = \int_0^1 \|Y\| dt = \|Y\|.$$

Note that

$$\|Y\|^2 = \text{tr}(Y^*Y) = \text{tr}\left[\frac{1}{4}(\log(\mathcal{A}^*\mathcal{A}))^2\right].$$

Thus, the distance-squared function is given by

$$d_{U(p) \times U(q)}^2: U(p, q) \rightarrow \mathbb{R}, \quad \mathcal{A} \mapsto \frac{1}{4} \text{tr}[(\log(\mathcal{A}^*\mathcal{A}))^2].$$

Appendix A The continuity of the map (3-2)

Recall the statement of Proposition 3.14:

Proposition A.1 *The map $s: S(v) \rightarrow [0, \infty)$, as defined in (3-2), is continuous.*

The proof relies on a characterization of $s(v)$.

Lemma A.2 *Let $u \in S_p(v)$. A positive real number T is $s(u)$ if and only if $\gamma_u: [0, T] \rightarrow M$ is an N -geodesic and at least one of the following holds:*

- (i) $\gamma_u(T)$ is the first focal point of N along γ_u .
- (ii) There exists $v \in S(v)$ with $v \neq u$ such that $\gamma_v(T) = \gamma_u(T)$.

Note that $\gamma_u(T)$ being a focal point of N along γ_u means that $(D \exp_v)(uT)$ is not of full rank, where \exp_v is the normal exponential, as defined in (3-1). When N is a point, this notion of focal points reduces to that of conjugate points.

In order to prove the lemma, we need the following observations:

Observation A [26, Lemma 2.11, page 96] *Let N be a submanifold of M and $\gamma: [a, \infty) \rightarrow M$ a geodesic emanating perpendicularly from N . If $\gamma(b)$ is the first focal point of N along γ , then, for $t > b$, $\gamma|_{[a,t]}$ cannot be an N -geodesic, ie $\ell(\gamma|_{[a,t]}) > d_N(\gamma(t))$.*

Recall that a sequence $\{\gamma_n\}$ of geodesics, defined on closed intervals, is said to converge to a geodesic γ if $\gamma_n(0) \rightarrow \gamma(0)$ and $\gamma'_n(0) \rightarrow \gamma'(0)$. It follows from the continuity of the exponential map that, if $t_n \rightarrow t$, then $\gamma_n(t_n) \rightarrow \gamma(t)$.

Observation B Let γ_n be unit-speed N -geodesics joining $p_n = \gamma_n(0)$ to $q_n = \gamma_n(t_n)$. If γ_n converges to a geodesic γ and $t_n \rightarrow l$, then γ is a unit-speed N -geodesic joining $p = \lim_n p_n$ to $q := \gamma(l) = \lim_n \gamma_n(t_n)$.

Proof The unit normal bundle $S(v)$ is closed. Since $\gamma'_n(0) \rightarrow \gamma'(0)$, it follows that $\gamma'(0) \in S(v)$. Note that

$$d_N(q) = \lim_{n \rightarrow \infty} d_N(q_n) = \lim_{n \rightarrow \infty} d(p_n, q_n) = \lim_{n \rightarrow \infty} t_n = l = \ell(\gamma|_{[0,l]})$$

implies that γ is an N -geodesic. □

Proof of Lemma A.2 If $\gamma_u(t)$ is the first focal point of N along γ_u , then Observation A implies that γ_u cannot be minimal beyond this value. If (ii) holds, then we need to show that, for sufficiently small $\varepsilon > 0$, $\gamma_u|_{[0, T+\varepsilon]}$ is not minimal. Suppose, on the contrary, that γ_u is minimal beyond T . Take a minimal geodesic β joining $\gamma_v(T - \varepsilon)$ to $\gamma_u(T + \varepsilon)$. Observe that

$$2\varepsilon = d(\gamma_u(T + \varepsilon), \gamma_u(T)) + d(\gamma_v(T), \gamma_v(T - \varepsilon)) > d(\gamma_u(T + \varepsilon), \gamma_v(T - \varepsilon)).$$

If $p, q, r \in M$ are such that $d(p, q) + d(q, r) = d(p, r)$ and there exist shortest normal geodesics γ_1 and γ_2 joining p to q and q to r , respectively, then $\gamma_1 \cup \gamma_2$ is smooth at q and defines a shortest normal geodesic joining p to r . Therefore, we have

$$\ell(\gamma_v|_{[0, T-\varepsilon]} \cup \beta) = T - \varepsilon + d(\gamma_v(T - \varepsilon), \gamma_u(T + \varepsilon)) < T + \varepsilon = \ell(\gamma_u|_{[0, T+\varepsilon]}).$$

This contradiction establishes that $\gamma_u|_{[0, T+\varepsilon]}$ is not minimal.

For the converse, set $T = s(u)$ and observe that $\gamma_u|_{[0, T]}$ is an N -geodesic. Assuming that $q := \gamma_u(T)$ is not the first focal point of N along γ_u , we will prove that (ii) holds. Let $p = \gamma_u(0)$ and choose a neighborhood \tilde{U} of Tu in v such that $\exp_v|_{\tilde{U}}$ is a diffeomorphism. For sufficiently large n , $q_n := \gamma_u(T + 1/n) \in \exp_v(\tilde{U})$. Take N -geodesics γ_n parametrized by arc length joining p_n to q_n and set $u_n := \dot{\gamma}_n(0) \in S((T_{p_n}N)^\perp)$. Since $S((T_{p_n}N)^\perp)$ is compact, by passing to a subsequence, we may assume that u_n converges to $v \in S(N_p)$. By Observation B,

$$\gamma_v(T) = \lim_{n \rightarrow \infty} \gamma_{u_n}\left(T + \frac{1}{n}\right) = \gamma_u(T).$$

If $v = u$, then, for sufficiently large n , $d(p, q_n)u_n \in \tilde{U}$, whence

$$\left(T + \frac{1}{n}\right)u = d(p, q_n)u_n.$$

Taking absolute values on both sides implies $T + 1/n > d(p, q_n)$. This contradiction implies $v \neq u$. □

Proof of Proposition A.1 We will prove that $s(u_n) \rightarrow s(u)$ whenever $(p_n, u_n) \rightarrow (p, u)$ in the unit normal bundle $S(v)$. Let T be any accumulation point of the sequence $\{s(u_n)\}$ including ∞ . By Observation B, $\gamma_u|_{[0, T]}$ is an N -geodesic and hence $T \leq s(u)$. If $T = +\infty$, we are done. So let us assume that $T < +\infty$. From Lemma A.2, at least one of the following holds for infinitely many n :

- (i) $s(u_n)$ is the first focal point of N along γ_{u_n} .
- (ii) There exist $v_n \in S(N_{p_n})$ with $v_n \neq u_n$ and $\gamma_{u_n}(s(u_n)) = \gamma_{v_n}(s(u_n))$.

If (i) is true for infinitely many n , then choose infinitely many unit vectors $\{w_n\}$ which belong to the kernel $\ker(D \exp_v(s(u_n)u_n))$ and are contained in a compact subset of $S(v)$. Choose a convergent subsequence whose limit w is contained in $\ker(D \exp_v(Tu))$. Since $w \neq 0$, the rank of $D \exp_v(Tu)$ is less than $\dim M$. Thus, $\gamma_u(T)$ is the first focal point of N along γ_u and $T = s(u)$.

If (ii) is true for infinitely many n , then we may assume that $v_n \rightarrow v \in S(v)$. If $v \neq u$, then Lemma A.2(ii) holds for T , whence $T = s(u)$. If $v = u$, we claim that $\gamma_u(T)$ is the first focal point of N along γ_u . If not, then the map \exp_v is regular at $Tu \in v$ and hence the map

$$\Phi: v \rightarrow M \times M, \quad (p, u) \mapsto (p, \exp_v(p, u)),$$

is regular at Tu . Therefore, Φ is a diffeomorphism if restricted to an open neighborhood \tilde{U} of Tu in v . Since $v = u$, which implies, for sufficiently large n , $(p_n, s(u_n)u_n)$ and $(p_n, s(u_n)v_n)$ belong to \tilde{U} and are different. On the other hand, by assumption, $\Phi(s(u_n)u_n) = \Phi(s(u_n)v_n)$, which is a contradiction. Therefore, $\gamma_u(T)$ is the first focal point and $T = s(u)$. □

Appendix B Derivative of the square root map

Lemma B.1 Let A be a positive-definite matrix and $\psi: A \mapsto \sqrt{A}$. Then

$$D\psi_A(H) = \int_0^\infty e^{-t\sqrt{A}} H e^{-t\sqrt{A}} dt$$

for any symmetric matrix H .

Proof As $\psi(A) \cdot \psi(A) = A$, differentiating at A , we obtain

$$(B-1) \quad D\psi_A(H)\psi(A) + \psi(A)D\psi_A(H) = H.$$

(i) Given a positive-definite matrix A and a symmetric matrix H , we need to show that the equation

$$B\sqrt{A} + \sqrt{A}B = H$$

has a unique solution. For that, we will prove that the map

$$f : \text{Symm}_n \rightarrow \text{Symm}_n, \quad B \mapsto B\sqrt{A} + \sqrt{A}B,$$

is bijective, where Symm_n denotes the set of all $n \times n$ symmetric matrices. Equivalently, we will show that f is injective. Without loss of generality, we assume that \sqrt{A} is a diagonal matrix with positive entries t_1, t_2, \dots, t_n . Note that

$$\ker(f) = \{B \in \text{Symm}_n : B\sqrt{A} + \sqrt{A}B = 0\}.$$

Consider

$$\begin{aligned} B\sqrt{A} + \sqrt{A}B = 0 &\implies B \text{diag}(t_1, \dots, t_n) + \text{diag}(t_1, \dots, t_n)B = 0 \\ &\implies t_j b_{ij} + t_i b_{ij} = 0 \quad \text{for } 1 \leq i, j \leq n. \end{aligned}$$

Since $t_i > 0$ for $1 \leq i \leq n$, we see $b_{ij} = 0$. Therefore, f is injective.

(ii) For any positive-definite matrix X and for any symmetric matrix Y , the integral

$$(B-2) \quad \int_0^\infty e^{-tX} Y e^{-tX} dt$$

converges. We note that the eigenvalues of e^{-tX} are $e^{-t\lambda_j}$, where λ_j are the eigenvalues of X . Since X is a positive-definite matrix, each of the λ_j is positive. Without loss of generality, we assume that $\lambda = \lambda_1$ is the smallest eigenvalue of X . Then we have

$$e^{-t\lambda_j} \leq e^{-t\lambda} \implies \|e^{-tX}\| = e^{-t\lambda},$$

where $\|\cdot\|$ is the operator norm. Therefore, the operator norm of the integrand in (B-2) is bounded by $2e^{-t\lambda}\|Y\|$, which is an integrable function. Hence, the integral given by (B-2) converges.

(iii) $D\psi_A(H)$ satisfies (B-1). Observe that

$$\begin{aligned} &\left(\int_0^\infty e^{-t\sqrt{A}} \cdot H \cdot e^{-t\sqrt{A}} dt\right)\sqrt{A} + \sqrt{A}\left(\int_0^\infty e^{-t\sqrt{A}} \cdot H \cdot e^{-t\sqrt{A}} dt\right) \\ &= \int_0^\infty (e^{-t\sqrt{A}} \cdot H \cdot e^{-t\sqrt{A}}\sqrt{A} + \sqrt{A}e^{-t\sqrt{A}} \cdot H \cdot e^{-t\sqrt{A}}) dt \\ &= \int_0^\infty (e^{-t\sqrt{A}} H e^{-t\sqrt{A}})' dt = H. \end{aligned}$$

From (i), (ii) and the uniqueness of the derivative, the lemma is proved. □

Lemma B.2 *The map $g : M(n, \mathbb{R}) \rightarrow \mathbb{R}, A \mapsto \text{tr}(\sqrt{A^T A})$, is differentiable if and only if A is invertible.*

Proof Let A be an invertible matrix. We will prove that the function g is differentiable at A . Let \mathcal{P} be the set of all positive-definite matrices, which is an open subset of the set of all symmetric matrices \mathcal{S} . We will prove that the map

$$r: \mathcal{P} \rightarrow \mathcal{P}, \quad A \mapsto \sqrt{A},$$

is differentiable. Define a function

$$s: \mathcal{P} \rightarrow \mathcal{P}, \quad A \mapsto A^2.$$

We will show that s is a diffeomorphism and, from the inverse function theorem, r will be differentiable. In order to show that s is a diffeomorphism, we claim that, for $A \in \mathcal{P}$, $Ds_A: T_A\mathcal{P} \rightarrow T_{A^2}\mathcal{P}$ is injective. Note that \mathcal{P} is an open subset of a vector space \mathcal{S} and, therefore, $T_A\mathcal{P} \cong \mathcal{S} \cong T_{A^2}\mathcal{P}$. So take $B \in \mathcal{S}$ such that $Ds_A(B) = 0$. We will show that $B = 0$. Recall that $Ds_A(B) = AB + BA$. Now choose an orthonormal basis $\{v_1, v_2, \dots, v_n\}$ of the eigenspace of A and let $Av_i = \lambda_i v_i$ ($\lambda_i > 0$). Then,

$$A(Bv_i) = -BAv_i = -B\lambda_i v_i = -\lambda_i(Bv_i),$$

which implies Bv_i is also an eigenvector of A with eigenvalue $-\lambda_i < 0$. Hence, $Bv_i = 0$, which implies $B = 0$.

For the converse, we will show that, if A is a singular matrix, then the map g is not directional differentiable. Let A be a singular matrix. Using the singular value decomposition, we write

$$A = U \begin{pmatrix} D & 0 \\ 0 & 0_k \end{pmatrix} V^T,$$

where D is an $(n - k) \times (n - k)$ diagonal matrix with positive entries. If

$$B = U \begin{pmatrix} 0_{n-k} & 0 \\ 0 & I_k \end{pmatrix},$$

then we claim that g is not differentiable in the direction of B . Since

$$\sqrt{(A + tB)^T(A + tB)} = V \begin{pmatrix} D & 0 \\ 0 & I_k|t| \end{pmatrix} V^T,$$

the limit

$$\begin{aligned} \lim_{t \rightarrow 0} \frac{g(A + tB) - g(A)}{t} &= \lim_{t \rightarrow 0} \frac{1}{t} \left(\operatorname{tr} \left(V \begin{pmatrix} D & 0 \\ 0 & I_k|t| \end{pmatrix} V^T \right) - \operatorname{tr} \left(V \begin{pmatrix} D & 0 \\ 0 & 0_k \end{pmatrix} V^T \right) \right) \\ &= k \lim_{t \rightarrow 0} \frac{|t|}{t} \end{aligned}$$

does not exist and hence the function g is not differentiable. □

References

- [1] **A Banyaga, D Hurtubise**, *Lectures on Morse homology*, Kluwer Texts in the Mathematical Sciences 29, Kluwer Academic, Dordrecht (2004) MR Zbl
- [2] **M G Barratt, J Milnor**, *An example of anomalous singular homology*, Proc. Amer. Math. Soc. 13 (1962) 293–297 MR Zbl
- [3] **R L Bishop, R J Crittenden**, *Geometry of manifolds*, Pure and Applied Mathematics XV, Academic, New York (1964) MR Zbl
- [4] **M A Buchner**, *Simplicial structure of the real analytic cut locus*, Proc. Amer. Math. Soc. 64 (1977) 118–121 MR Zbl
- [5] **H Busemann**, *The geometry of geodesics*, Academic, New York (1955) MR Zbl
- [6] **A Daniilidis, R Deville, E Durand-Cartagena, L Rifford**, *Self-contracted curves in Riemannian manifolds*, J. Math. Anal. Appl. 457 (2018) 1333–1352 MR Zbl
- [7] **H Gluck, D Singer**, *Scattering of geodesic fields, I*, Ann. of Math. 108 (1978) 347–372 MR Zbl
- [8] **A Gray**, *Tubes*, 2nd edition, Progr. Math. 221, Birkhäuser, Basel (2004) MR Zbl
- [9] **J J Hebda**, *The local homology of cut loci in Riemannian manifolds*, Tohoku Math. J. 35 (1983) 45–52 MR Zbl
- [10] **J J Hebda**, *Cut loci of submanifolds in space forms and in the geometries of Möbius and Lie*, Geom. Dedicata 55 (1995) 75–93 MR Zbl
- [11] **N J Higham**, *Functions of matrices: theory and computation*, Society for Industrial and Applied Mathematics, Philadelphia, PA (2008) MR Zbl
- [12] **M W Hirsch**, *Differential topology*, Graduate Texts in Math. 33, Springer (1994) MR Zbl
- [13] **J-i Itoh, S V Sabau**, *Riemannian and Finslerian spheres with fractal cut loci*, Differential Geom. Appl. 49 (2016) 43–64 MR Zbl
- [14] **J-i Itoh, C Vîlcu**, *Orientable cut locus structures on graphs*, preprint (2011) arXiv 1103.3136
- [15] **J-i Itoh, C Vîlcu**, *Every graph is a cut locus*, J. Math. Soc. Japan 67 (2015) 1227–1238 MR Zbl
- [16] **S Kobayashi**, *On conjugate and cut loci*, from “Studies in global geometry and analysis”, Math. Assoc. America, Englewood Cliffs, NJ (1967) 96–122 MR
- [17] **Y Li, L Nirenberg**, *The distance function to the boundary, Finsler geometry, and the singular set of viscosity solutions of some Hamilton–Jacobi equations*, Comm. Pure Appl. Math. 58 (2005) 85–146 MR Zbl
- [18] **C Mantegazza, A C Mennucci**, *Hamilton–Jacobi equations and distance functions on Riemannian manifolds*, Appl. Math. Optim. 47 (2003) 1–25 MR Zbl

- [19] **R J Martin, P Neff**, *Minimal geodesics on $GL(n)$ for left-invariant, right- $O(n)$ -invariant Riemannian metrics*, *J. Geom. Mech.* 8 (2016) 323–357 MR Zbl
- [20] **S B Myers**, *Connections between differential geometry and topology, I: Simply connected surfaces*, *Duke Math. J.* 1 (1935) 376–391 MR Zbl
- [21] **H Omori**, *A class of Riemannian metrics on a manifold*, *J. Differential Geometry* 2 (1968) 233–252 MR Zbl
- [22] **S Plotnick**, *Embedding homology 3-spheres in S^5* , *Pacific J. Math.* 101 (1982) 147–151 MR Zbl
- [23] **H Poincaré**, *Sur les lignes géodésiques des surfaces convexes*, *Trans. Amer. Math. Soc.* 6 (1905) 237–274 MR
- [24] **M M Postnikov**, *Geometry, VI: Riemannian geometry*, *Encyclopaedia of Mathematical Sciences* 91, Springer (2001) MR Zbl
- [25] **S V Sabau, M Tanaka**, *The cut locus and distance function from a closed subset of a Finsler manifold*, *Houston J. Math.* 42 (2016) 1157–1197 MR Zbl
- [26] **T Sakai**, *Riemannian geometry*, *Translations of Mathematical Monographs* 149, Amer. Math. Soc., Providence, RI (1996) MR Zbl
- [27] **V A Sharafutdinov**, *Complete open manifolds of nonnegative curvature*, *Sibirsk. Mat. Zh.* 15 (1974) 126–136 MR Zbl In Russian; translated with modifications as “Proof of soul theorem”, preprint (2006)
- [28] **H Singh**, *On the cut locus and the focal locus of a submanifold in a Riemannian manifold, II*, *Publ. Inst. Math. (Beograd)* 41(55) (1987) 119–124 MR Zbl
- [29] **F-E Wolter**, *Distance function and cut loci on a complete Riemannian manifold*, *Arch. Math. (Basel)* 32 (1979) 92–96 MR Zbl
- [30] **V A Yakubovich, V M Starzhinskii**, *Linear differential equations with periodic coefficients*, volume 1, Halsted, New York (1975) MR Zbl

*Department of Mathematics and Statistics, Indian Institute of Science Education and Research
Kolkata, India*

*Department of Mathematics and Statistics, Indian Institute of Science Education and Research
Kolkata, India*

somnath.basu@iiserkol.ac.in, sp17rs038@iiserkol.ac.in

Received: 4 June 2021 Revised: 15 February 2023

ALGEBRAIC & GEOMETRIC TOPOLOGY

msp.org/agt

EDITORS

PRINCIPAL ACADEMIC EDITORS

John Etnyre
etnyre@math.gatech.edu
Georgia Institute of Technology

Kathryn Hess
kathryn.hess@epfl.ch
École Polytechnique Fédérale de Lausanne

BOARD OF EDITORS

Julie Bergner	University of Virginia jeb2md@eservices.virginia.edu	Robert Lipshitz	University of Oregon lipshitz@uoregon.edu
Steven Boyer	Université du Québec à Montréal cohf@math.rochester.edu	Norihiko Minami	Nagoya Institute of Technology nori@nitech.ac.jp
Tara E Brendle	University of Glasgow tara.brendle@glasgow.ac.uk	Andrés Navas	Universidad de Santiago de Chile andres.navas@usach.cl
Indira Chatterji	CNRS & Univ. Côte d'Azur (Nice) indira.chatterji@math.cnrs.fr	Thomas Nikolaus	University of Münster nikolaus@uni-muenster.de
Alexander Dranishnikov	University of Florida dranish@math.ufl.edu	Robert Oliver	Université Paris 13 bobol@math.univ-paris13.fr
Tobias Ekholm	Uppsala University, Sweden tobias.ekholm@math.uu.se	Jessica S Purcell	Monash University jessica.purcell@monash.edu
Mario Eudave-Muñoz	Univ. Nacional Autónoma de México mario@matem.unam.mx	Birgit Richter	Universität Hamburg birgit.richter@uni-hamburg.de
David Futер	Temple University dfuter@temple.edu	Jérôme Scherer	École Polytech. Féd. de Lausanne jerome.scherer@epfl.ch
John Greenlees	University of Warwick john.greenlees@warwick.ac.uk	Vesna Stojanoska	Univ. of Illinois at Urbana-Champaign vesna@illinois.edu
Ian Hambleton	McMaster University ian@math.mcmaster.ca	Zoltán Szabó	Princeton University szabo@math.princeton.edu
Matthew Hedden	Michigan State University mhedden@math.msu.edu	Maggy Tomova	University of Iowa maggy-tomova@uiowa.edu
Hans-Werner Henn	Université Louis Pasteur henn@math.u-strasbg.fr	Nathalie Wahl	University of Copenhagen wahl@math.ku.dk
Daniel Isaksen	Wayne State University isaksen@math.wayne.edu	Chris Wendl	Humboldt-Universität zu Berlin wendl@math.hu-berlin.de
Thomas Koberda	University of Virginia thomas.koberda@virginia.edu	Daniel T Wise	McGill University, Canada daniel.wise@mcgill.ca
Christine Lescop	Université Joseph Fourier lescop@ujf-grenoble.fr		


See inside back cover or msp.org/agt for submission instructions.

The subscription price for 2023 is US \$650/year for the electronic version, and \$940/year (+ \$70, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP. Algebraic & Geometric Topology is indexed by Mathematical Reviews, Zentralblatt MATH, Current Mathematical Publications and the Science Citation Index.

Algebraic & Geometric Topology (ISSN 1472-2747 printed, 1472-2739 electronic) is published 9 times per year and continuously online, by Mathematical Sciences Publishers, c/o Department of Mathematics, University of California, 798 Evans Hall #3840, Berkeley, CA 94720-3840. Periodical rate postage paid at Oakland, CA 94615-9651, and additional mailing offices. POSTMASTER: send address changes to Mathematical Sciences Publishers, c/o Department of Mathematics, University of California, 798 Evans Hall #3840, Berkeley, CA 94720-3840.

AGT peer review and production are managed by EditFlow[®] from MSP.

PUBLISHED BY

 **mathematical sciences publishers**
nonprofit scientific publishing

<http://msp.org/>

© 2023 Mathematical Sciences Publishers

ALGEBRAIC & GEOMETRIC TOPOLOGY

Volume 23

Issue 9 (pages 3909–4400)

2023

Two-dimensional extended homotopy field theories	3909
KÜRŞAT SÖZER	
Efficient multisections of odd-dimensional tori	3997
THOMAS KINDRED	
Bigrading the symplectic Khovanov cohomology	4057
ZHECHI CHENG	
Fibrations of 3–manifolds and asymptotic translation length in the arc complex	4087
BALÁZS STRENNER	
A uniformizable spherical CR structure on a two-cusped hyperbolic 3–manifold	4143
YUEPING JIANG, JIEYAN WANG and BAOHUA XIE	
A connection between cut locus, Thom space and Morse–Bott functions	4185
SOMNATH BASU and SACHCHIDANAND PRASAD	
Staircase symmetries in Hirzebruch surfaces	4235
NICKI MAGILL and DUSA MCDUFF	
Geometric triangulations of a family of hyperbolic 3–braids	4309
BARBARA NIMERSHIEM	
Beta families arising from a v_2^9 self-map on $S/(3, v_1^8)$	4349
EVA BELMONT and KATSUMI SHIMOMURA	
Uniform foliations with Reeb components	4379
JOAQUÍN LEMA	