Nonlinear Analysis 239 (2024) 113421

Contents lists available at ScienceDirect

Nonlinear Analysis

www.elsevier.com/locate/na

Local non-injectivity of the exponential map at critical points in sub-Riemannian geometry

Samuël Borza [a,*], Wilhelm Klingenberg [b]

[a] Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy
[b] Department of Mathematics, Durham University, United Kingdom

ARTICLE INFO

Article history:
Received 19 July 2023
Accepted 13 October 2023
Communicated by Enrico Valdinoci

MSC:
53C17
58E10
53C22

Keywords:
Sub-Riemannian geometry
Conjugate points
Metric geometry

ABSTRACT

We prove that the sub-Riemannian exponential map is not injective in any neighbourhood of certain critical points. Namely that it does not behave like the injective map of reals given by $f(x) = x^3$ near its critical point $x = 0$. As a consequence, we characterise conjugate points in ideal sub-Riemannian manifolds in terms of the metric structure of the space. The proof uses the Hilbert invariant integral of the associated variational problem.

1. Introduction

In their 1932 landmark paper, Morse and Littauer [18] showed that the exponential map $\exp_p : \mathrm{Dom}(\exp_p) \subseteq \mathrm{T}_{p(M)} \to M$, based at a point p of an analytic Finsler manifold M, is never injective on any neighbourhood of critical value $v \in \mathrm{T}_{p(M)}$. The critical values of the exponential map are also called conjugate vectors at p, while the image $\exp_{p(v)} \in M$ of a conjugate vector v is what is usually called a conjugate point to p (along the geodesic curve $t \mapsto \exp_p(tv)$. Morse and Littauer showed that conjugate points occur precisely when certain families of extremals (called fields of extremals) fail to cover the neighbourhood of these points in a one-to-one manner. Savage [22] extended this result to smooth Finsler manifolds, while Warner [26] gave a different proof of the same result by obtaining normal forms of the exponential map near those conjugate vectors, at which Whitney's singularity theory can be applied.

The present work aims to extend this result to sub-Riemannian geometry. There are important differences between Riemannian and sub-Riemannian geometry that need to be taken into account. We explain some of them in simple terms here; they will be detailed rigorously afterwards.

* Corresponding author.
E-mail addresses: sborza@sissa.it (S. Borza), wilhelm.klingenberg@dur.ac.uk (W. Klingenberg).

https://doi.org/10.1016/j.na.2023.113421

The geometry of a sub-Riemannian manifold M is better described from the Hamiltonian point of view. Riemannian geodesics solve a second-order differential equation while the sub-Riemannian geodesic equation is a Hamiltonian system on the cotangent bundle $\mathrm{T}^*(M)$. Pontryagin's maximum principle, stated in Theorem 6, shows that a sub-Riemannian length minimiser $\gamma : [0,T] \to M$ has a lift $\lambda : [0,T] \to \mathrm{T}^*(M)$ that is either *normal*, or else is *abnormal*. A normal lift is one that satisfy the sub-Riemannian geodesic equation. Abnormal lifts on the other hands satisfy another non-differential condition. A geodesic γ is normal (resp. abnormal) if it has a normal (resp. abnormal) lift. In general, a curve γ could be both normal and abnormal, although a normal lift λ of γ must be unique. A geodesic can be abnormal and not normal, we say that it is *strictly* abnormal. A geodesic γ that is normal but not abnormal is said to be *strictly* normal. A strictly normal geodesic γ is therefore not abnormal, but this does not mean that the restriction of γ to a subsegment of $[0,T]$ cannot be abnormal. A geodesic γ that is normal on every subsegment is said to be *strongly* normal.

Abnormal geodesics are still not completely understood, and there are famous open problems related to them. It is still unknown in general if abnormal geodesics are always smooth, and if the size of the set of points that can be reached by abnormal curve starting from a given point is a null set. These are the so-called regularity and Sard conjectures in sub-Riemannian geometry. In this work, we will mainly deal with normal geodesics. A sub-Riemannian manifold that does not have non-trivial abnormal geodesics is said to be *ideal*.

The sub-Riemannian exponential map is the projection of the flow of the sub-Riemannian equation, that is to say

$$\exp_p : \mathrm{Dom}(\exp_p) \subseteq \mathrm{T}^*_p(M) \to M : \lambda_0 \to \gamma(1),$$

where $\gamma : [0,1] \mapsto M$ is the projection of the solution $\lambda : [0,1] \to \mathrm{T}^*(M)$ to the sub-Riemannian geodesic equation with initial condition $\lambda(0) = (p, \lambda_0)$. The curve $t \in [0,1] \mapsto \exp_p(t\lambda_0)$ is a normal geodesic. A critical value of $\exp_p$ is defined in the same way as in Riemannian geometry, a covector λ_0 is conjugate if $\dim(\mathrm{Ker}\mathrm{d}_{\lambda_0}\exp_p) \neq 0$. The order of conjugacy, or *multiplicity*, of λ_0 is precisely defined as the dimension of $\mathrm{Ker}(\mathrm{d}_{\lambda_0}\exp_p)$. Our goal is to show that, under some generic and nondegenerate conditions, the covector λ_0 is a critical value of $\exp_p$ if and only if $\exp_p$ is not injective in any neighbourhood of $\lambda_0 \in \mathrm{T}^*_p(M)$, generalising Morse and Littauer's theorem to the sub-Riemannian exponential map. We describe in a few words those generic and nondegenerate conditions that we will assume for our main result.

In our previous work [8], we showed that the sub-Riemannian exponential map satisfies a *continuity property*, similar to what we have in Riemannian geometry. If $\lambda_0 \in \mathrm{T}^*_p(M)$ is a conjugate covector, then there is a neighbourhood $\mathcal{U}$ of this covector such that the following holds: for all $\lambda_0' \in \mathcal{U}$, the number of conjugate covectors, counted with their multiplicities, on the intersection of $\mathcal{U}$ with the straight line passing through the origin and λ_0' is equal to the multiplicity of λ_0. Following Warner [26, Section 3], we say that a conjugate covector λ_0 is *regular* if, in a neighbourhood $\mathcal{U}$ of this covector, there exists precisely one conjugate covector, which, due to the continuity property, has the same multiplicity as λ_0, on every straight line connecting the origin and the neighbourhood $\mathcal{U}$. A conjugate vector that is not regular will be said to be *singular*.

Furthermore, and unlike what happens in Riemannian geometry, the multiplicity of a critical value λ_0 of $\exp_p$ is usually not equal to the order of the root $t = 1$ of the map

$$[0,1] \to \mathbb{R} : t \mapsto \det(\mathrm{d}_{t\lambda_0}\exp_p).$$

It is typically smaller. It will be key to our argument to nevertheless assume that this order is still finite. We will prove in Theorem 29 that the subset of conjugate vectors in $\mathrm{T}^*_p(M)$ is a smooth hypersurface in a small neighbourhood of a conjugate covecter λ_0 that has finite order which is also of independent interest.

Theorem 1. *Let M be a sub-Riemannian manifold and $p \in M$. If $\lambda_0 \in \mathrm{Dom}(\exp_p) \subseteq \mathrm{T}^*_p(M)$ is a strongly normal and regular conjugate covector of finite order, then the exponential map $\exp_p$ is not injective in any neighbourhood of λ_0.*

Theorem 1 is a refinement of the following known fact in Riemannian and sub-Riemannian geometry: if the structure does not admit abnormal extremals, and if γ is a length-minimising geodesic with conjugate endpoints $p = \gamma(0)$ and $q = \gamma(1)$, then on any neighbourhood $V \subseteq M$ of q the exponential map $\exp_p|_{\exp_p^{-1}(V)} : \exp_p^{-1}(V) \to M$ is not injective (see [1, Theorem 8.73 and Corollary 8.74]). A priori, this statement does not exclude the possibility that the pairs of geodesics connecting a point p with certain points $q \in V$ may be distant from each other at intermediate times. Theorem 1 (and analogously the Morse–Littauer–Savage theorem in Riemannian geometry) demonstrates this.

Already in Riemannian geometry, it does not seem possible to obtain Theorem 1 in full generality without going through a complex proof (see the discussion in Section 4.3 alongside the description of a flawed, simpler proof present in the literature). Only in the next very specific case (see Theorem 39), we know of a more straightforward proof. Let $\mathrm{Cut}(p)$ denote the cotangent cut locus, i.e. the set of initial covectors corresponding to geodesics that are minimising up to time $t = 1$ but not minimising up to time $t = 1 + \varepsilon$, for any $\varepsilon > 0$. The subset $\mathrm{Cut}^1(p)$ of $\mathrm{Cut}(p)$ consists of the covectors λ_0 for which there exists another $\lambda_0' \in \mathrm{Cut}(p)$ such that $\exp_p(\lambda_0) = \exp_p(\lambda_0')$.

Theorem 2. *Let M be an ideal sub-Riemannian manifold, and $p \in M$. If $\lambda_0 \in \mathrm{Cut}(p) \setminus \mathrm{Cut}^1(p)$, then the exponential map $\exp_p$ fails to be injective in any neighbourhood of $\lambda_0 \in \mathrm{T}_p^*(M)$.*

Note that in the previous statement, we do not need to assume that λ_0 is a regular conjugate covector, or that it has finite order. This is because here we do not need to prove that the conjugate locus is a smooth hypersurface of $\mathrm{T}_p^*(M)$ to conclude, contrary to our proof of Theorem 1.

The local non-injectivity of the sub-Riemannian exponential map at critical points has applications in metric geometry. In [23] Shankar and Sormani introduced different metric definitions of conjugate points along geodesics. If $\gamma : [0,1] \to X$ is a geodesic in a metric space X, then the point $q := \gamma(1)$ is said to be *one-sided conjugate* to $p := \gamma(0)$ along γ if there exists a sequence of points (q_n) converging to q such that for every n, there are two distinct geodesics γ_n^1 and γ_n^2 joining p to q_n and both converging to γ. There are variants of this definition, that we recall in Definition 36, which will give meaning to q being *symmetrically conjugate*, or *unreachably conjugate*, to p along γ.

We prove that when our version of Morse–Littauer–Savage theorem applies, for example in the ideal case, then there is a correspondence between these metric notions of conjugate points and the differential one. This shows that if discrepancies in metric notions of conjugate points are to be found in sub-Riemannian geometry, abnormal segments or infinite order of conjugate points must play a role.

Theorem 3. *Let M be an ideal sub-Riemannian manifold, $\gamma : [0,1] \to M$ be a normal geodesic such that its initial covector λ_0 is regular conjugate and has finite order, and denote $p := \gamma(0)$ and $q := \gamma(1)$. Then, the following statements are equivalent:*

(i) *q is conjugate to p along γ;*
(ii) *q is one-sided conjugate to p along γ;*
(iii) *q is symmetrically conjugate to p along γ.*

Furthermore, if p and q are unreachable conjugate points along γ, then q is also conjugate to p along γ.

The paper is organised as follows. In Section 2, we summarise notions from sub-Riemannian geometry that will be important in our argument. The family of extremals is introduced in Section 3.1 while in Section 3.2 we define the Hilbert integrals. In Section 3.3, we address the regularity of the sub-Riemannian conjugate locus. This result is a contribution to sub-Riemannian geometry of independent interest. Finally, we prove Theorem 1 in Section 3.4 and discuss Theorem 3, as well as possible future work and related open questions, in Section 4.

2. Preliminaries

We begin with a general description of *sub-Riemannian geometry*, one that includes rank-varying structures. Nowadays, there are many good references on sub-Riemannian geometry. We refer the reader to the seminal paper [25], as well as the reference books [11,15]. In this work, we will adopt the point of view and notations from [1].

A sub-Riemannian structure is a connected manifold M equipped with a set of smooth global vector fields $X_1, \dots, X_m$ called the *generating family*. To be more precise, we are defining what is sometimes called a *free* sub-Riemannian structure (see [1, Chapter 3]). However, the most general definition of sub-Riemannian structure is always equivalent to a free one, as shown in [1, Corollary 3.27].

The distribution at a point $p \in M$ is $\mathcal{D}_p := \operatorname{span}\{X_1(p), \dots, X_m(p)\}$. The *rank* of the sub-Riemannian structure at $p \in M$ is $\operatorname{rank}(p) := \dim(\mathcal{D}_p)$. Observe that in our definition, a sub-Riemannian manifold may be rank-varying. An inner product on $\mathcal{D}_p$ is induced by the polarisation formula applied to the norm

$$\|u\|^2_{\mathcal{D}_p} := \min\left\{ \sum_{k=1}^m u_i^2 \;\middle|\; \sum_{k=1}^m u_i X_k(p) = u \right\}.$$

A curve $\gamma : [0,T] \to M$ with initial value $\gamma(0) = p \in M$ is horizontal if there exists $u \in \mathrm{L}^2([0,T], \mathbb{R}^m)$, called a *control*, such that $\dot\gamma(t) = \sum_{k=0}^m u_k(t) X_k(\gamma(t))$ for almost every $t \in [0,T]$. In fact, from Carathéodory's existence theorem, there exists a unique maximal Lipschitz solution to the Cauchy problem

$$\begin{cases} \dot\gamma(t) = \sum_{k=0}^m u_k(t) X_k(\gamma(t)) \\ \gamma(0) = p \end{cases} \tag{1}$$

for every $u \in \mathrm{L}^2([0,T], \mathbb{R}^m)$ and $p \in M$.

The *sub-Riemannian length* and the *sub-Riemannian energy* of γ are defined by

$$\mathrm{L}(\gamma) = \int_0^T \|\dot\gamma(t)\|_{\mathcal{D}_{\gamma(t)}} \mathrm{d}t, \;\; \mathrm{J}(\gamma) = \frac{1}{2} \int_0^T \|\dot\gamma(t)\|^2_{\mathcal{D}_{\gamma(t)}} \mathrm{d}t. \tag{2}$$

In the case where every two points can be joined by a horizontal curve, we have a well-defined distance function on M.

Definition 4. The distance function of a sub-Riemannian manifold M, also called the *Carnot-Carathéodory distance*, is defined by

$$\mathrm{d}(x,y) := \inf\{\mathrm{L}(\gamma) \mid \gamma : [0,T] \to M \text{ is horizontal and } \gamma(0) = x \text{ and } \gamma(T) = y\}.$$

In this work, we assume that the sub-Riemannian structures satisfy *Hörmander's condition* (introduced by Hörmander in [12]), that is to say $\mathrm{Lie}_p(\mathcal{D}) = \mathrm{T}_p(M)$ for all $p \in M$, where $\mathrm{Lie}_p(\mathcal{D})$ is defined as the smallest Lie algebra equipped with the Lie bracket of vector fields that contains $\mathcal{D}$. In that case, we also say that $\mathcal{D}$ is *bracket-generating*. This is motivated by the well-known result from the independent works of Chow [9] and Rashevsky [21] (see also [1, Theorem 3.31]).

Theorem 5 (*Chow–Rashevsky Theorem*)**.** *Let M be a sub-Riemannian manifold such that its distribution $\mathcal{D}$ is $\mathcal{C}^\infty$ and satisfies the Hörmander condition. Then, (M, d) is a metric space and the manifold and metric topology of M are equivalent.*

On the space of controls $\mathrm{L}^2([0,T], \mathbb{R}^m)$, we can define a *length functional*, as well as a corresponding *energy functional*

$$\mathrm{L}(u) := \int_0^T \|u(t)\|_{\mathbb{R}^m} \mathrm{d}t, \;\; \mathrm{J}(u) := \frac{1}{2} \int_0^T \|u(t)\|^2_{\mathbb{R}^m} \mathrm{d}t.$$

Given an horizontal curve $\gamma : [0,T] \to M$, we define at every differentiability point of γ the *minimal control* $\overline{u}$ associated with γ

$$\overline{u}(t) := \arg\min \left\{ \|u\|_{\mathbb{R}^m} \mid u \in \mathbb{R}^m, \dot{\gamma}(t) = \sum_{k=0}^{m} u_k X_k(\gamma(t)) \right\}.$$

The relationship with the functionals defined in (2) is the following: $\mathrm{L}(\gamma) = \mathrm{L}(\overline{u})$ and $\mathrm{J}(\gamma) = \mathrm{J}(\overline{u})$ when $\overline{u}$ is the minimal control associated with γ. Through the Cauchy problem (1), it can be seen that finding a length minimiser for L among the horizontal curves with fixed endpoints, where $\gamma(0) = p$ and $\gamma(T) = q$, is equivalent to finding a control that minimises the length functional L among all the controls for which the associated path connects points p and q. Furthermore, we have the following classical correspondence: a horizontal curve $\gamma : [0,T] \to M$ joining p to q is a minimiser of J if and only if it is a minimiser of L and is parametrised by constant speed, see [1, Lemma 3.64]. Therefore, it is also equivalent to finding a control that minimises the energy J among all the controls for which the associated path connects points p and q, this is the *sub-Riemannian energy minimisation* problem.

In metric geometry, a length minimiser that is parametrised by constant speed is often called a minimising geodesic, while a geodesic is a curve that is only locally length minimising, i.e. it is a length minimiser between every pair of points close enough in its image. Because of the lack of a torsion-free metric connection, we cannot have a geodesic equation through a covariant derivative. Rather, sub-Riemannian geodesics are characterised via Hamilton's equations. The *Hamiltonian* of the sub-Riemannian structure is defined by

$$H : \mathrm{T}^*(M) \to \mathbb{R} : \lambda \mapsto H(\lambda) := \max_{u \in \mathbb{R}^m} \left(\sum_{k=1}^{m} u_k \langle \lambda, X_k(\pi(\lambda)) \rangle - \frac{1}{2} \sum_{k=1}^{m} u_k^2 \right). \tag{3}$$

The Hamiltonian H may be written in terms of the generating family of the sub-Riemannian structure $(X_1, \ldots, X_m)$ as follows

$$H(\lambda) = \frac{1}{2} \sum_{k=1}^{m} h_k(p, \lambda_0)^2, \qquad \forall \lambda \in T^*(M),$$

where $h_k(p, \lambda_0) := \langle \lambda_0, X_k(p) \rangle$. For $p \in M$, we will also write H_p for the restriction of H to the cotangent space $\mathrm{T}^*_p(M)$. The following result can also be viewed as an application of Pontryagin's Maximum Principle first proved by Pontryagin in [20]. It provides a necessary condition for a control to be a solution to the sub-Riemannian energy minimisation problem.

Theorem 6 (*Pontryagin's Maximum Principle*). *Let $\gamma : [0,T] \to M$ be a horizontal curve which is a length minimiser among horizontal curves, and parametrised by constant speed. Then, there exists a Lipschitz curve $\lambda(t) \in \mathrm{T}^*_{\gamma(t)}(M)$ such that one and only one of the following is satisfied:*

(N) $\dot{\lambda} = \overrightarrow{H}(\lambda)$, where $\overrightarrow{H}$ is the unique vector field in $\mathrm{T}^*(M)$ such that $\sigma(\cdot, \overrightarrow{H}(\lambda)) = \mathrm{d}_\lambda H$ for all $\lambda \in \mathrm{T}^*(M)$ and σ denotes the canonical symplectic form on the cotangent bundle $\mathrm{T}^*(M)$;

(A) $\sigma_{\lambda(t)}(\dot{\lambda}(t), \cap_{k=1}^{n} \ker(\mathrm{d}_{\lambda(t)} h_k)) = 0$ for all $t \in [0,T]$.

Moreover, in case (A), we have $\lambda(t) \neq 0$ for every $t \in [0,T]$.

If λ is a curve in $\mathrm{T}^*(M)$ that satisfies (N) (resp. (A)), we will also say that λ is a normal extremal (resp. abnormal extremal). The same terminology is used for the corresponding curve $\gamma = \pi(\lambda)$ and for the minimal control associated to γ. Here, the map $\pi : \mathrm{T}^*(M) \to M$ denotes the bundle projection. We note that an extremal could be both normal and abnormal. A normal trajectory $\gamma : [0,T] \to M$ is called *strictly normal* if it is not abnormal. If, in addition, the restriction $\gamma|_{[0,s]}$ is strictly normal for every $s > 0$, we say that γ is *strongly normal*. It can be seen that γ is strongly normal if and only if the normal geodesic γ does not contain any abnormal segment. The projection of a normal extremal to M is locally minimising, that is to say it is a geodesic (for the sub-Riemannian distance) parametrised by constant-speed. More precisely, it holds $\|\dot{\gamma}(t)\|^2 = 2H(\lambda(t))$ for every $t \in [0, T]$. The projection of an abnormal extremal to M might not necessarily be a geodesic (or a minimising geodesic).

Sub-Riemannian structures which do not admit any non-trivial (i.e. non-constant) abnormal geodesics (the trivial geodesic is always abnormal as soon as $\text{rank}(\mathcal{D}_p) < \dim(M)$) are said to be *ideal*. Examples of ideal sub-Riemannian manifolds include the Heisenberg groups, Sasakian manifolds, and more generally any contact sub-Riemannian manifold, as well as the α-Grushin plane (see [6]). The Martinet flat structure, the Engel group, and in fact any Carnot group of step 3 and higher, are not ideal.

The theory of ordinary differential equations proves the existence of a maximal solution to (N) for every given initial condition $\lambda(0) \in \mathrm{T}^*(M)$. The time t - flow of Hamilton's equation (N) is the semigroup denoted by $\mathrm{e}^{t\vec{H}} : \mathrm{T}^*(M) \to \mathrm{T}^*(M)$. The restriction of the time 1 - map of this flow to the fibre $\mathrm{T}^*_p(M)$, followed by projection to the base, is called the sub-Riemannian *exponential map* based at p.

Definition 7. The sub-Riemannian *exponential map* at $p \in M$ is the map

$$\exp_p : U_p \to M : \lambda \mapsto \pi(\mathrm{e}^{\vec{H}}(\lambda)),$$

where $U_p \subseteq \mathrm{T}^*_p(M)$ is the open set of covectors such that the corresponding solution of (N) is defined on the interval $[0, 1]$.

The sub-Riemannian exponential map $\exp_p$ is smooth. If $\lambda : [0, T] \to \mathrm{T}^*(M)$ is the normal extremal that satisfies the initial condition $\lambda(0) = (p, \lambda_0) \in \mathrm{T}^*(M)$, then the corresponding normal extremal path $\gamma(t) = \pi(\lambda(t))$ by definition satisfies $\gamma(t) = \exp_p(t\lambda_0)$ for all $t \in [0, T]$. If M is complete for the Carathéodory distance, then $U_p = \mathrm{T}^*_p(M)$, and if in addition there are no strictly abnormal length minimisers, then the exponential map $\exp_p$ is surjective. Contrary to the Riemannian case, the sub-Riemannian exponential map is not necessarily a diffeomorphism of a small ball in $\mathrm{T}^*_p(M)$ onto a small geodesic ball in M. In fact, $\mathrm{Im}(\mathrm{d}_0\exp_p) = \mathcal{D}_p$ and $\exp_p$ is a local diffeomorphism at $0 \in \mathrm{T}^*_p(M)$ if and only if $\mathcal{D}_p = \mathrm{T}_p(M)$.

Definition 8. The critical points of the exponential map $\exp_p : U_p \to M$ are called *conjugate covectors at* p. We denote by $\mathrm{Conj}(p) \subseteq U_p$ the collection of all such covectors. If $s\lambda_0 \in \mathrm{T}^*_p(M)$ is conjugate, we say that the point $q := \exp_p(s\lambda_0)$ is conjugate to $p = \exp_p(0 \cdot \lambda_0)$ along the normal geodesic $\gamma(t) := \exp_p(t\lambda_0)$, and that s is a conjugate time.

We mention here some important properties related to conjugacy. The restriction of a normal extremal γ to an interval $[t, t + \varepsilon]$ is abnormal if and only if $\gamma(s)$ is a conjugate point to $\gamma(0)$ for all $s \in [t, t + \varepsilon]$ ([1, Theorem 8.47]). Furthermore, the set of conjugate times of γ is discrete if γ does not contain abnormal segments ([1, Corollary 8.51]). We will also use the following result from [1, Theorem 8.61].

Theorem 9. *Let $\gamma : [0, T] \to M$ be a normal extremal that does not contain abnormal segments. If γ has no conjugate points, then it is a local minimiser for the length on the space of admissible trajectories with the same endpoints. If γ contains at least one conjugate point to $\gamma(0)$, then it is not a local minimiser on the space of admissible trajectories with the same endpoints.*

3. Non-local injectivity of the sub-Riemannian exponential map

3.1. A family of normal extremals and the associated flow

For the rest of this work, we fix a sub-Riemannian manifold M and a point $p \in M$.

Definition 10. The *family of extremals* from p in augmented space is the map given by

$$F_p : \mathcal{U}_p \to \mathbb{R} \times M : (t, \lambda_0) \mapsto (t, \exp_p(t\lambda_0)),$$

where $\mathcal{U}_p \subseteq \mathbb{R} \times \mathrm{T}^*_p(M)$ is the maximal open set on which F_p is well-defined.

The non-local injectivity of the exponential map may be related to the non-local injectivity of the family of extremals.

Definition 11. We say that the family of extremals containing $p \in M$ *simply covers* $(t, q) \in \mathbb{R} \times M$ if there exists an open $\mathcal{U}' \subseteq \mathcal{U}_p$ such that F_p is injective on $\mathcal{U}'$ and $(t, q) \in F_p(\mathcal{U}') \subseteq \mathbb{R} \times M$.

Proposition 12. *The exponential map* $\exp_p : U_p \to M$ *is injective in a neighbourhood of a covector* $\lambda_0 \in U_p \subseteq \mathrm{T}^*_p(M)$ *if and only if the family of extremals* F_p *simply covers* $(1, \exp_p(\lambda_0)) \in \mathbb{R} \times M$.

Proof. Suppose that $\exp_p$ is injective in a neighbourhood $\lambda_0 \in U' \subseteq U_p$. Then for small $\varepsilon > 0$, the restriction of F_p to $]1-\varepsilon, 1+\varepsilon[\times U'$ is also injective. Indeed, the equality $F_p(t, \lambda_0) = F_p(t', \lambda_0')$, for some $t, t' \in]1-\varepsilon, 1+\varepsilon[$ and $\lambda_0, \lambda_0' \in U'$, implies that $t = t'$, $\exp_p(t\lambda_0) = \exp_p(t'\lambda_0')$, and thus $\lambda_0 = \lambda_0'$. The other implication is proved in a similar fashion. □

Similarly, we may parametrise the lift of normal extremals.

Definition 13. *The flow of extremals* based at $p \in M$ is the map given by

$$\Phi_p : \mathcal{V}_p \to \mathbb{R} \times \mathrm{T}^*(M) : (t, \lambda_0) \mapsto (t, \mathrm{e}^{t\vec{H}}(p, \lambda_0)),$$

where $\mathcal{V}_p \subseteq \mathbb{R} \times \mathrm{T}^*_p(M)$ is the maximal open set on which Φ_p is well-defined.

Remark 14. From the definition of $\exp_p(\lambda_0)$ and $\mathrm{e}^{t\vec{H}}(p, \lambda_0)$, we know that the domains of the maps F_p and Φ_p coincide: $\mathcal{U}_p = \mathcal{V}_p$.

The maps F_p and Φ_p defined in this section are analogues of the concept of a *family of extremals* in the classical calculus of variations. Using the standard terminology, the family of extremals considered here is *central* since the extremals all start from the same point p.

A cotangent version of Gauss' lemma reads in this context as follows, see [1, Proposition 8.42].

Theorem 15. *Let* $p \in M$, $t \in \mathbb{R}$, *and* $\delta :]-\varepsilon, \varepsilon[\to \mathrm{T}^*_p(M)$ *such that* $t\delta(s) \in U_p$ *for all* $s \in]-\varepsilon, \varepsilon[$. *Let* $\lambda(t, s) := \mathrm{e}^{t\vec{H}}(p, \delta(s))$ *and* $\gamma(t, s) := \pi(\lambda(t, s)) = \exp_p(t\delta(s))$,

$$\left\langle \lambda(t, s), \frac{\mathrm{d}}{\mathrm{d}s}\gamma(t, s) \right\rangle = \frac{\mathrm{d}}{\mathrm{d}s} H(\lambda(t, s)),$$

where $\langle \cdot, \cdot \rangle$ *denotes the natural pairing of covectors and vectors.*

3.2. Hilbert invariant integrals

The *Poincaré-Cartan one-form* on $\mathbb{R} \times \mathrm{T}^*(M)$, associated to the Hamiltonian of the sub-Riemannian structure, is defined to be

$$\theta - H\mathrm{d}t \in \Omega^1(\mathbb{R} \times \mathrm{T}^*(M)),$$

where $\theta \in \Omega^1(\mathrm{T}^*(M))$ denotes the tautological one-form of the cotangent bundle. The pullback map associated to the smooth map Φ_p will be denoted by Φ_p^*.

Definition 16. The Hilbert integral I_p^*, defined on an open set $\mathcal{U} \subseteq \mathcal{U}_p \subseteq \mathbb{R} \times \mathrm{T}_p^*(M)$, is the line integral obtained by integrating the one-form $\eta_p^* := \Phi_p^*(\theta - H\mathrm{d}t) \in \Omega^1(\mathcal{U})$. In other words, if Γ^* is a smooth parametrised curve in $\mathcal{U}$, then

$$I_p^*[\Gamma^*] = \int_{\Gamma^*} \eta_p^* = \int_{\Gamma^*} \Phi_p^*(\theta - H\mathrm{d}t).$$

Let us give precise description to the action of the form η_p^*. Fix $(t, \lambda_0) \in \mathcal{U}$ and $v \in \mathrm{T}_{(t,\lambda_0)}(\mathcal{U})$, which we identify with $\mathrm{T}_{(t,\lambda_0)}(\mathbb{R} \times \mathrm{T}_p^*(M))$. As usual, we set $\lambda(t) := \mathrm{e}^{t\vec{H}}(p, \lambda_0)$ and $\gamma(t) := \exp_p(t\lambda_0)$. By definition, we have that

$$(\eta_p^*)_{(t,\lambda_0)}[v] = (\theta - H\mathrm{d}t)_{(t,\lambda(t))} \left[\mathrm{d}_{(t,\lambda_0)} \Phi_p(v)\right].$$

If we write $v = w + s\frac{\partial}{\partial t}(t, \lambda_0)$ for unique $s \in \mathbb{R}$ and $w \in \mathrm{T}_{\lambda_0}(\mathrm{T}_p^*(M))$, seen as a subspace of $\mathrm{T}_{(t,\lambda_0)}(\mathbb{R} \times \mathrm{T}_p^*(M))$, we can then express $\mathrm{d}_{(t,\lambda_0)} \Phi_p(v) \in \mathrm{T}_{(t,\lambda(t))}(\mathbb{R} \times \mathrm{T}^*(M))$ as

$$\mathrm{d}_{(t,\lambda_0)} \Phi_p(v) = \mathrm{d}_{(p,\lambda_0)}\mathrm{e}^{t\vec{H}} \circ \mathrm{d}_{\lambda_0}\iota_p(w) + s\left(\frac{\partial}{\partial t}(t, \lambda(t)) + \vec{H}(\lambda(t))\right), \tag{4}$$

where $\iota_p : \mathrm{T}_p^*(M) \to \mathrm{T}^*(M)$ is the injection $\iota_p(\lambda_0) := (p, \lambda_0)$, satisfying $\pi \circ \iota_p = p$ for all $p \in M$.

Now by recalling Hamilton's equation from Theorem 6, we obtain

$$(\eta_p^*)_{(t,\lambda_0)}[v] = \big\langle \lambda(t), \mathrm{d}_{\lambda_0}\exp_p(t\ \cdot)[w]\big\rangle + s\big(\langle\lambda(t), \dot\gamma(t)\rangle - H(\lambda(t))\big). \tag{5}$$

The Hilbert integral I_p^* has some useful properties. When evaluated along a ray in the augmented space $\mathbb{R} \times \mathrm{T}_p^*(M)$, it evaluates the length of the corresponding extremal in M.

Proposition 17. *Given an open neighbourhood $\mathcal{U}$ in $\mathcal{U}_p$, a covector $\lambda_0 \in \mathrm{T}_p^*(M)$, and the curve $\Gamma^* : [t_0, t_1] \to \mathbb{R} \times \mathrm{T}_p^*(M)$ defined by $\Gamma^*(t) = (t, \lambda_0)$, then we have*

$$I_p^*[\Gamma^*] = \mathrm{L}(\gamma|_{[t_0,t_1]}),$$

where $\gamma(t) := \exp_p(t\lambda_0)$, whenever $\Gamma^*(t) \in \mathcal{U}$ and $t\lambda_0 \in U_p$ for all $t \in [t_0, t_1]$.

Proof. Since $\dot\Gamma^*(t) = (1, 0)$, (5) implies that

$$I_p^*[\Gamma^*] = \int_{t_0}^{t_1} \langle\lambda(t), \dot\gamma(t)\rangle - H(\lambda(t))\mathrm{d}t,$$

where $\lambda(t) := \mathrm{e}^{t\vec{H}}(p, \lambda_0)$. By (3), we can write the maximised Hamiltonian as

$$H(\lambda(t)) = \langle\lambda(t), \dot\gamma(t)\rangle - \frac{1}{2}|u(t)|^2, \text{ for all } t \in [t_0, t_1],$$

where $u(t)$ is the minimal control of $\gamma(t) = \exp_p(t\lambda_0)$, i.e. $u_i(t) = \langle\lambda(t), X_i(\gamma(t))\rangle$ for $i = 1, \dots, N$. Therefore,

$$I_p^*[\Gamma^*] = \int_{t_0}^{t_1} \left[\langle\lambda(t), \dot\gamma(t)\rangle - \left(\langle\lambda(t), \dot\gamma(t)\rangle - \frac{1}{2}|u(t)|^2\right)\right] \mathrm{d}t = \frac{1}{2}\int_{t_0}^{t_1} |u(t)|^2\mathrm{d}t = \mathrm{L}(\gamma|_{[t_0,t_1]}). \quad \square$$

Our goal now is to establish that the Hilbert integral I_p^* defines an *invariant* integral, namely that it is independent of path relative to endpoints.

Proposition 18. *The one-form η_p^* defined on an open connected subset $\mathcal{U}$ of $\mathcal{U}_p \subseteq \mathbb{R} \times \mathrm{T}_p^*(M)$ is closed. Equivalently, the Hilbert integral I_p^* is homotopy-invariant in $\mathcal{U}$.*

Proof. We prove that η_p^* is closed by showing that $\mathrm{d}\eta_p^* = 0$ vanishes identically. We have

$$\begin{aligned} \mathrm{d}\eta_p^* &= \mathrm{d}\left[\Phi_p^*(\theta - H\mathrm{d}t)\right] = \Phi_p^*\left[\mathrm{d}(\theta - H\mathrm{d}t)\right] = \Phi_p^*\left[\mathrm{d}(\theta - H\mathrm{d}t)\right] \\ &= \Phi_p^*\left[\mathrm{d}\theta - \mathrm{d}H \wedge \mathrm{d}t\right] = \Phi_p^*\left[\sigma - \mathrm{d}H \wedge \mathrm{d}t\right], \end{aligned}$$

where σ denotes the Poincaré two-form on $\mathrm{T}^*(M)$. If $(t, \lambda_0) \in \mathcal{U} \subseteq \mathbb{R} \times \mathrm{T}_p^*(M)$ and $v_1, v_2 \in \mathrm{T}_{(t,\lambda_0)}(\mathcal{U}) \simeq \mathbb{R} \times \mathrm{T}_p^*(M)$, then

$$(\mathrm{d}\eta_p^*)_{(t,\lambda_0)}(v_1, v_2) = [\sigma - \mathrm{d}H \wedge \mathrm{d}t]_{(t,\lambda(t))}\,(\mathrm{d}_{(t,\lambda_0)}\Phi_p(v_1), \mathrm{d}_{(t,\lambda_0)}\Phi_p(v_2)).$$

As previously in (4), we let $v_i = w_i + s_i \frac{\partial}{\partial t}(t, \lambda_0)$ for unique $s_i \in \mathbb{R}$ and $w_i \in \mathrm{T}_{\lambda_0}(\mathrm{T}_p^*(M))$.

Without loss of generality, we only need to treat the following two cases: $s_1 = s_2 = 0$, and $s_1 = 0$ while $s_2 \neq 0$.

In the first case, we use the invariance of the symplectic form under the Hamiltonian flow, i.e. the equality $\mathrm{e}^{t\vec{H}*}\sigma = \sigma$, as well as the fact that $\mathrm{d}_{(p,\lambda_0)}\pi \circ \mathrm{d}_{\lambda_0}\iota_p = 0$ to find

$$\begin{aligned} (\mathrm{d}\eta_p^*)_{(t,\lambda_0)}(v_1, v_2) &= \sigma_{\lambda(t)}(\mathrm{d}_{(p,\lambda_0)}\mathrm{e}^{t\vec{H}} \circ \mathrm{d}_{\lambda_0}\iota_p(w_1), \mathrm{d}_{(p,\lambda_0)}\mathrm{e}^{t\vec{H}} \circ \mathrm{d}_{\lambda_0}\iota_p(w_2)) \\ &= (\mathrm{e}^{t\vec{H}*}\sigma)_{(p,\lambda_0)}(\mathrm{d}_{\lambda_0}\iota_p(w_1), \mathrm{d}_{\lambda_0}\iota_p(w_2)) = \sigma_{(p,\lambda_0)}(\mathrm{d}_{\lambda_0}\iota_p(w_1), \mathrm{d}_{\lambda_0}\iota_p(w_2)) \\ &= (\iota_p\sigma)^*_{\lambda_0}(w_1, w_2) = \iota_p^*(\mathrm{d}\theta)_{\lambda_0}(w_1, w_2) = \mathrm{d}\iota_p^*(\theta)_{\lambda_0}(w_1, w_2) = 0. \end{aligned}$$

In the second case, we use Hamilton's equation alongside the definition of the symplectic gradient, and we obtain

$$\begin{aligned} (\mathrm{d}\eta_p^*)_{(\lambda_0,t)}(v_1, v_2) &= \sigma_{\lambda(t)}\left(\mathrm{d}_{(p,\lambda_0)}\mathrm{e}^{t\vec{H}} \circ \mathrm{d}_{\lambda_0}\iota_p(w_1), \vec{H}(\lambda(t))\right) \\ &\quad - (\mathrm{d}H \wedge \mathrm{d}t)_{(t,\lambda(t))}\left(\mathrm{d}_{(p,\lambda_0)}\mathrm{e}^{t\vec{H}} \circ \mathrm{d}_{\lambda_0}\iota_p(w_1), \frac{\partial}{\partial t}(t, \lambda(t))\right) \\ &\quad - (\mathrm{d}H \wedge \mathrm{d}t)_{(t,\lambda(t),t)}\left(\mathrm{d}_{(p,\lambda_0)}\mathrm{e}^{t\vec{H}} \circ \mathrm{d}_{\lambda_0}\iota_p(w_1), \vec{H}(\lambda(t))\right) \\ &= (\mathrm{e}^{t\vec{H}}(p,\cdot)^*\mathrm{d}H)_{\lambda_0}(w_1) - (\mathrm{e}^{t\vec{H}}(p,\cdot)^*\mathrm{d}H)_{\lambda_0}(w_1) \wedge \mathrm{d}t(\vec{H}(\lambda(t))) \\ &\quad + (\mathrm{d}H)_{\lambda(t)}(\vec{H}(\lambda(t))) \wedge \mathrm{d}t(\mathrm{d}_{(p,\lambda_0)}\mathrm{e}^{t\vec{H}} \circ \mathrm{d}_{\lambda_0}\iota_p(w_1)) \\ &\quad - (\mathrm{e}^{t\vec{H}}(p,\cdot)^*\mathrm{d}H)_{\lambda_0}(w_1) \wedge \mathrm{d}t\left(\frac{\partial}{\partial t}(t, \lambda(t))\right) \\ &\quad + (\mathrm{d}H)_{\lambda(t)}\left(\frac{\partial}{\partial t}(t, \lambda(t))\right) \wedge \mathrm{d}t(\mathrm{d}_{(p,\lambda_0)}\mathrm{e}^{t\vec{H}} \circ \mathrm{d}_{\lambda_0}\iota_p(w_1)) = 0, \end{aligned}$$

which concludes the proof. □

We will now complete the proof of the invariance of I_p^*.

Proposition 19. *Let $\mathcal{U} \subseteq \mathbb{R} \times \mathrm{T}_p^*(M)$ be an open neighbourhood which is convex in the $\mathbb{R}$-direction (i.e. if $(t_0, \lambda_0), (t_1, \lambda_0) \in \mathcal{U}$, then $(t, \lambda_0) \in \mathcal{U}$ for all $t \in [t_0, t_1]$). Then, one-form η_p^* defined on $\mathcal{U}$ is exact. Equivalently, the Hilbert integral I_p^* is path-independent in $\mathcal{U}$.*

Proof. Let $\Gamma^*(s) = (t(s), \lambda_0(s))$ be closed path in $\mathcal{U}$ parametrised on the interval $[a, b]$. According to the convexity hypothesis, this curve is homotopic to the curve $\Gamma_0^*(s) := (t(a), \lambda_0(s))$. It follows from Proposition 18, (5) and the cotangent version Gauss' lemma as seen in Theorem 15 that

$$\begin{aligned} \int_{\Gamma^*} \eta_p^* = \int_{\Gamma_0^*} \eta_p^* &= \int_a^b \frac{\mathrm{d}}{\mathrm{d}s}\left[H\left(\mathrm{e}^{t(a)\vec{H}}(p, \lambda_0(s))\right)\right]\mathrm{d}s \\ &= H\left(\mathrm{e}^{t(a)\vec{H}}(p, \lambda_0(b))\right) - H\left(\mathrm{e}^{t(a)\vec{H}}(p, \lambda_0(a))\right) = 0, \end{aligned}$$

since $\lambda_0(\cdot)$ is a closed curve. □

When a family of extremals containing p simply covers an open neighbourhood of $\mathbb{R} \times M$, it can be used to define a Hilbert integral I_p on a subset of $\mathbb{R} \times M$.

Definition 20. Assume that the family of extremals F_p is injective on a neighbourhood $\lambda_0 \in \mathcal{U} \subseteq \mathbb{R} \times \mathrm{T}_p^*(M)$. Then the Hilbert integral I_p, defined on $F_p(\mathcal{U}) \subseteq \mathbb{R} \times M$, is the line integral obtained by integrating the continuous one-form

$$(\eta_p)_{(t,q)}(s, w) = \langle \lambda_t, w \rangle - H(q, \lambda_t)s, \tag{6}$$

where $(q, \lambda_t) := \mathrm{e}^{t\overrightarrow{H}}(p, \lambda_0)$, $(t, \lambda_0) := F_p^{-1}(t, q)$, and $s\frac{\partial}{\partial t} + w \in T_{(t,q)}(\mathbb{R} \times M)$.

Remark 21. While $\eta_p^* \in \Omega^1(\mathcal{U})$ is smooth, the form $\eta_p \in \Omega^1(F_p(\mathcal{U}))$ needs to be only continuous. Namely the functions (q, λ_t) in its definition depend smoothly on t, p, λ_0, but the value λ_0 in turn depends only *continuously* on (t, q). This is because F_p is only simply covering, thereby a homeomorphism. This lack of smoothness will require a detailed analysis, as given in Section 3.4.

In a similar way to Proposition 17, evaluating the Hilbert integral I_p on the graph of one of the normal extremals of the family F_p simply gives its length.

Proposition 22. *Given an open neighbourhood $\mathcal{U} \subseteq \mathcal{U}_p$ such that the family of extremals F_p is injective on $\mathcal{U}$, a covector $\lambda_0 \in \mathrm{T}_p^*(M)$, and the curve $\Gamma : [t_0, t_1] \to F_p(\mathcal{U}_p)$ defined by $\Gamma(t) := (t, \exp_p(t\lambda_0))$. Then the following equality holds :*

$$I_p[\Gamma] = \mathrm{L}(\gamma|_{[t_0,t_1]}).$$

Here $\gamma(t) := \exp_p(t\lambda_0)$, whenever $\Gamma(t) \in F_p(\mathcal{U})$ and $t\lambda_0 \in U_p$ for all $t \in [t_0, t_1]$.

Proof. Reasoning as in the proof of Proposition 17 and using (6), it is seen that

$$I_p[\Gamma] = \int_\Gamma \eta_p = \int_{t_0}^{t_1} \left(\langle \mathrm{e}^{t\overrightarrow{H}}(p, \lambda_0), \dot{\gamma}(t) \rangle - H(\mathrm{e}^{t\overrightarrow{H}}(p, \lambda_0)) \right) \mathrm{d}t = \mathrm{L}(\gamma|_{[t_0,t_1]}). \quad \square$$

Furthermore, the Hilbert integrals and I_p^* on $\mathcal{U}$ and I_p on $F_p(\mathcal{U})$ coincide on smooth curves (assuming that F_p is injective), as shown in the following proposition.

Proposition 23. *Suppose that the family of extremals $F_p : \mathcal{U} \to \mathbb{R} \times M$ is an injective map. Assume also that either Γ^* is a smooth curve in $\mathcal{U}$ and $\Gamma := F_p \circ \Gamma^*$ is the corresponding smooth curve in $F_p(\mathcal{U})$, or assume that both Γ is a smooth curve in $F_p(\mathcal{U})$ and the corresponding curve $\Gamma^* = F_p^{-1} \circ \Gamma$ in $\mathcal{U}$ are smooth. Then $I_p^*[\Gamma^*] = I_p[\Gamma]$.*

Proof. Let us write $\Gamma(s) := (t(s), q(s))$, as well as $\Gamma^*(s) := (t(s), \lambda_0(s))$, with $q(s) := \exp_p(t_0(s)\lambda_0(s))$, where the functions $t(\cdot), q(\cdot)$ and $\lambda_0(\cdot)$ are smooth. As before we use the notation $\lambda(s) := \mathrm{e}^{t(s)\overrightarrow{H}}(p, \lambda_0(s))$ and $\gamma(s) := \exp_p(t_0(s), \lambda_0(s))$.

Notice first that

$$\dot{\gamma}(s) = \frac{\mathrm{d}}{\mathrm{d}s} \left(\pi \circ \mathrm{e}^{t(s)\overrightarrow{H}}(p, \lambda_0(s)) \right) = \mathrm{d}_{\lambda(s)}\pi \left[\dot{t}(s)\overrightarrow{H}(\lambda(s)) + \mathrm{d}_{\lambda_0(s)}\mathrm{e}^{t(s)\overrightarrow{H}}(p, \cdot)(\dot{\lambda}_0(s)) \right].$$

Therefore, (5) and (6) yield

$$(\eta_p)_{\Gamma(s)}(\dot{\Gamma}(s)) = \langle \lambda(s), \dot{\gamma}(s) \rangle - \dot{t}(s)H(\lambda(s))$$

$$
\begin{aligned}
&= \langle \lambda(s), \mathrm{d}_{\lambda(s)}\pi \circ \mathrm{d}_{\lambda_0(s)} \mathrm{e}^{t(s)\overrightarrow{H}}(p,\cdot)(\dot\lambda_0(s))\rangle \\
&\quad + \dot t(s)\left[\langle \lambda(s), \mathrm{d}_{\lambda(s)}\pi \circ \overrightarrow{H}(\lambda(s))\rangle - H(\lambda(s))\right] \\
&= (\eta_p^*)_{\Gamma^*(t)}(\dot\Gamma^*(t)),
\end{aligned}
$$

which proves the statement. □

The relations between I_p and I_p^* given by the last two propositions are insufficient to extract the exactness of η_p. In order to progress, we will now study the nature of the sub-Riemannian conjugate locus.

3.3. The sub-Riemannian conjugate locus

Let us recall the regularity and continuity properties of the sub-Riemannian exponential map proven by the authors in [8]. If a Riemannian metric g on M is fixed, one can consider the isomorphism

$$\sharp : \mathrm{Ver}_\lambda \subseteq \mathrm{T}_\lambda(\mathrm{T}_p^*(M)) \to \mathrm{T}_p(M) : \xi \mapsto \xi^\sharp, \qquad p := \pi(\lambda),\ \lambda \in \mathrm{T}^*(M),$$

where $\xi^\sharp$ is the unique element of $\mathrm{T}_p(M)$ such that $g(\xi^\sharp, X) = \xi(X)$, for every $X \in \mathrm{T}_p(M)$, the spaces Ver_λ and $\mathrm{T}_p^*(M)$ being canonically identified. As explained in [8, Section 3.2], fixing a Riemannian metric in this way is equivalent to choosing a symplectic moving frame along a normal geodesic $\gamma(t)$, and extend the scalar product $\langle\cdot,\cdot\rangle_{\gamma(t)}$ along $\gamma(t)$ to the whole manifold.

The *ray* in $U_p \subseteq \mathrm{T}_p^*(M)$ through λ_0 is the map

$$\mathrm{r}_{p,\lambda_0} : [0,T] \to \mathrm{T}_p^*(M) : t \mapsto t\lambda_0$$

where $[0,T] \subseteq \mathbb{R}^+$ is the maximal interval containing 0 such that $t\lambda_0 \in U_p$ for every $t \in [0,T]$. In this way, $\dot{\mathrm{r}}_{p,\lambda_0}(t) \in \mathrm{T}_{t\lambda_0}(\mathrm{T}_p^*(M))$ and identifying $\mathrm{T}_{t\lambda_0}(\mathrm{T}_p^*(M))$ with $\mathrm{T}_p^*(M)$ in the usual way, we have $\dot{\mathrm{r}}_{p,\lambda_0}(t) = \lambda_0$ for every $t \in [0,T]$.

Theorem 24 (*Warner-Regularity of the Sub-Riemannian Exponential Map [8]*)**.** *Let M be a sub-Riemannian manifold and $p \in M$. Then,*

- ***(R1)*** *The map $\exp_p$ is $\mathcal{C}^\infty$ on $U_p = \mathrm{Dom}(\exp_p)$ and, for all $\lambda_0 \in U_p \setminus H_p^{-1}(0)$ and all $t \in I_{p,\lambda_0}$, we have $\mathrm{d}_{t\lambda_0}\exp_p(\dot{\mathrm{r}}_{p,\lambda_0}(t)) \neq 0_{\exp_p(t\lambda_0)}$;*
- ***(R2)*** *The map*
$$\mathrm{Ker}(\mathrm{d}_{\lambda_0}\exp_p(\lambda_0)) \to \mathrm{T}_{\exp_p(\lambda_0)}(M) : \xi_0 \mapsto \left(\mathrm{d}_{\lambda_0}\mathrm{e}^{t\overrightarrow{H}}[\xi_0]\right)^\sharp$$
has its image g-perpendicular to $\mathrm{Im}(\mathrm{d}_{\lambda_0}\exp_p)$
- ***(R3)*** *Let $\lambda_0 \in U_p \setminus H_p^{-1}(0)$ be a covector such that the corresponding geodesic $\gamma(t) := \exp_p(t\lambda_0)$ is strongly normal. Then, there exists a radially convex neighbourhood $\mathcal{V}$ of λ_0 such that for every ray $\mathrm{r}_{p,\overline{\lambda}_0}$ which intersects $\mathcal{V}$ that does not contain abnormal subsegments in $\mathcal{V}$, the number of singularities of $\exp_p$ (counted with multiplicities) on $\mathrm{Im}(\mathrm{r}_{p,\overline{\lambda}_0}) \cap \mathcal{V}$ is constant and equals the order of λ_0 as a singularity of $\exp_p$, i.e. $\dim(\mathrm{Ker}(\mathrm{d}_{\lambda_0}\exp_p))$.*

In view of the continuity property (R3), we will now adapt Warner's definition of singular and regular conjugate points and their order (see [26, Section 3.]) to the sub-Riemannian setting.

Definition 25. We say that a conjugate covector $\lambda_0 \in \mathrm{T}_p^*(M)$ is *regular* for $\exp_p$ if there is a neighbourhood U of λ_0 such that every ray of $\mathrm{T}_p^*(M)$ contains at most one covector in U which is conjugate. A conjugate covector that is not regular is said to be singular. The collection of regular conjugate (resp. singular conjugate) covectors is denoted by $\mathrm{Conj}^R(p) \subseteq \mathrm{T}_p^*(M)$ (resp. $\mathrm{Conj}^S(p)$). We write $\mathrm{Conj}_U^R(p) := \mathrm{Conj}^R(p) \cap U$ (resp. $\mathrm{Conj}_U^S(p) := \mathrm{Conj}^S(p) \cap U$).

In Riemannian geometry, the Jacobi fields along a given geodesic satisfy a second order differential equation. Consequently, the order of vanishing the function $t \mapsto \det(\mathrm{d}_{tv}\exp_p)$ at $t = 1$ is always finite and equal to the order of $v \in \mathrm{T}_p(M)$ as a singularity of $\exp_p$ (see [16]). This motivates the following definition.

Definition 26. We say that a covector $\lambda_0 \in U_p$ has order $m \in \mathbb{N} \cup \{+\infty\}$ if the map $t \mapsto \det(\mathrm{d}_{t\lambda_0}\exp_p)$ vanishes of order m at $t = 1$.

The definition above is well posed thanks to (R2) of Theorem 24. Indeed, the domain of the map

$$\mathrm{d}_{t\lambda_0}\exp_p : \mathrm{T}_{t\lambda_0}(\mathrm{T}_p^*(M)) \to \mathrm{T}_{\exp_p(t\lambda_0)}(M)$$

is first identified with the space $\mathrm{T}_p^*(M)$ as usual and then we have isomorphisms

$$\mathrm{T}_p^*(M) \cong \mathrm{Ker}(\mathrm{d}_{t\lambda_0}\exp_p) \oplus \mathrm{Im}(\mathrm{d}_{t\lambda_0}\exp_p) \cong \mathrm{T}_{\exp_p(t\lambda_0)}(M).$$

The last identification depends on the choice of a moving frame in (R2). However, the vanishing of the determinant of $\mathrm{d}_{t\lambda_0}\exp_p$, viewed as a map from $\mathrm{T}_{\exp_p(t\lambda_0)}(M)$ to itself, as well as the vanishing of its derivatives that we need for Definition 26, are independent of this choice.

Let us describe in geometric and analytic more details the content of Definition 26, and its relation to Jacobi fields. We will make use of the theory for sub-Riemannian Jacobi fields explained in [8] (see also [5]). Let us fix a symplectic moving frame $E_1(t), \dots, E_n(t), F_1(t), \dots, F_n(t)$ along the lift $\lambda(t)$ of $\gamma(t) := \exp_p(t\lambda_0)$, that is

$$\mathrm{Ker}(\mathrm{d}_{\lambda(t)}\pi) = \mathrm{span}\{E_1(t), \dots, E_n(t)\} \subseteq \mathrm{T}_{\lambda(t)}(\mathrm{T}^*(M)),$$

and

$$\sigma(E_i, E_j) = 0, \qquad \sigma(F_i, F_j) = 0, \qquad \sigma(E_i, F_j) = \delta_{i,j},$$

where σ is the usual symplectic form on $\mathrm{T}^*(M)$. The differential of the inclusion $\mathrm{T}_p^*(M) \to \mathrm{T}^*(M) : \lambda_0 \mapsto (p, t\lambda_0)$ is used to identify $\mathrm{T}_{t\lambda_0}(\mathrm{T}_p^*(M))$ with $\mathrm{span}\{tE_1(0), \dots, tE_n(0)\}$. A vector field $\mathcal{J}(t) = \sum_{i=1}^n p_i(t)E_i(t) + x_i(t)F_i(t)$ along $\lambda(t)$ is then a Jacobi field if and only if

$$\begin{pmatrix} \dot{p} \\ \dot{x} \end{pmatrix} = \begin{pmatrix} -C_1(t) & -R(t) \\ C_2(t) & C_1(t)^{\mathsf{T}} \end{pmatrix} \begin{pmatrix} p \\ x \end{pmatrix}, \tag{7}$$

for some smooth curves of $n \times n$ matrices $C_1(t), C_2(t)$ and $R(t)$ such that $C_2(t)$ and $R(t)$ are symmetric, $C_2(t) \geq 0$. The Riemannian Jacobi equation can be recovered by choosing a symplectic moving frame induced by the parallel transport, in which case $C_1(t) = 0$, $C_2(t) = \mathbb{1}$, and $R(t)$ is the curvature operator along the geodesic.

It can also be shown, see [8, Section 3.2] for example, that the multiplicity λ_0 as a conjugate covector is given by the dimension of the space Jacobi fields (p, x) satisfying $x(0) = x(1) = 0$. In these coordinates, the determinant given in Definition 26 is given by

$$\det(\mathrm{d}_{t\lambda_0}\exp_p) = \det\begin{pmatrix} x_1(t) & x_2(t) & \cdots & x_n(t) \end{pmatrix}, \tag{8}$$

where $(p_1, x_1), (p_2, x_2), \dots, (p_n, x_n)$ form a basis of the space of Jacobi fields (p, x) such that $x(0) = 0$. Without loss of generality, we can assume that the k first columns in (8) correspond to a basis of the space of Jacobi fields (p, x) such that $x(0) = x(1) = 0$.

At first glance, it seems to us that the possibility of the order of vanishing of $t \mapsto \det(\mathrm{d}_{t\lambda_0}\exp_p)$ at $t = 1$ being infinite cannot be ruled out in general. However, it appears challenging to construct a meaningful example. It may be that one would have to construct an example similar to the one appearing in [14] to have $t \mapsto \det(\mathrm{d}_{t\lambda_0}\exp_p)$ of infinite order at $t = 1$. Clearly, if the sub-Riemannian structure is analytic, as is the case in all Carnot group, for instance, then every strictly normal conjugate covector must have finite order.

In any case, we give further sufficient natural conditions under which a conjugate covector has finite order. Suppose that all Jacobi fields (p, x) such that $x(0) = x(1) = 0$ also satisfy $\dot{x}(1) = \ddot{x}(1) = \cdots = x^{(l-1)}(1) = 0$ and $x^{(l)}(1) \neq 0$ for some fixed $l \in \mathbb{N}$, and suppose also that in the expression $x^{(l)}(1) = C(1)p(1)$ induced from Eq. (7), the matrix $C(1)$ is invertible and positive definite. This will be the case if the matrices appearing in the Hamiltonian system Eq. (7) obey some kind of generalised Kalman rank condition. All the examples investigated in [7,8] satisfy this condition.

Under these assumptions and using Hadamard's lemma, (8) can be written as

$$\det(\mathrm{d}_{t\lambda_0}\exp_p) = \frac{(t-1)^{lk}}{l!k} \underbrace{\det\begin{pmatrix} y_1(t) & \cdots & y_k(t) & x_{k+1}(t) & \cdots & x_n(t) \end{pmatrix}}_{=:E(t)}, \tag{9}$$

where $y_1, \dots, y_k$ are smooth functions such that $y_1(1) = x_1^{(l)}(1), \dots, y_k(1) = x_k^{(l)}(1)$. We claim that $E(1) \neq 0$. If not, we would have

$$a_1 x_1^{(l)}(1) + \cdots + a_k x_k^{(l)}(1) = a_{k+1}x_1(1) + \cdots + a_n x_n(1),$$

for some $a_1, \dots, a_n \in \mathbb{R}$ not all zero. Define the Jacobi fields $(p, x) := a_{k+1}(p_{k+1}, x_{k+1}) + \cdots + a_n(p_n, x_n)$ and $(\overline{p}, \overline{x}) := a_1(p_1, x_1) + \cdots + a_k(p_k, x_k)$. We show that $x(1) \neq 0$. Suppose that $0 = a_{k+1}x_1(1) + \cdots + a_n x_n(1) = x(1)$, then we would have $a_1 = \cdots = a_n = 0$, and $a_1 x_1^{(l)}(1) + \cdots + a_k x_k^{(l)}(1) = 0$. Since the matrix $C(1)$ is assumed to be invertible, this would also imply that $0 = a_1 p_1(1) + \cdots + a_k p_k(1) = \overline{p}(1)$. Now, we have $\overline{x}(1) = \overline{p}(1) = 0$ and thus the Jacobi fields $(\overline{p}, \overline{x})$ must be identically zero, and so $a_1 = \cdots = a_k = 0$. Therefore, we must have $x(1) \neq 0$. In this case, [8, Lemma 16] shows that

$$p \cdot \overline{x} - \overline{p} \cdot x = 0, \text{ for all } t \in [0, 1],$$

and evaluating this expression at $t = 1$ gives

$$\overline{p}(1) \cdot C(1)\overline{p}(1) = \overline{p}(1) \cdot x(1) = p(1) \cdot \overline{x} = 0,$$

which is impossible since $C(1)$ is assume to be positive definite. Therefore, we must have $E(1) \neq 0$ and so λ_0 is a conjugate covector of order lk, according to Definition 26.

Proposition 27. *If $\lambda_0 \in \mathrm{Conj}^R(p)$ has finite order, then there exists a neighbourhood $U \subseteq \mathrm{T}_p^*(M)$ of λ_0 such that all vectors in $\mathrm{Conj}_U^R(p)$ have the same order.*

Proof. Assume that λ_0 has order $m \geq 1$. By Malgrange's Preparation Theorem (see [10]) there exists in a neighbourhood $]1-\varepsilon, 1+\varepsilon[\times V$ of $(1, \lambda_0) \in \mathbb{R} \times \mathrm{T}_p^*(M)$ a factorisation of the type

$$\det(\mathrm{d}_{t\eta}\exp_p) = c(t, \eta) \cdot \left[(t-1)^m + a_{m-1}(\eta)(t-1)^{m-1} + \cdots + a_0(\eta)\right], \tag{10}$$

where $c :]1-\varepsilon, 1+\varepsilon[\times V \to \mathbb{R}$, $a_i : V \to \mathbb{R}$ are smooth functions such that $c(1, \lambda_0) \neq 0$ and $a_{m-1}(\lambda_0) = \cdots = a_0(\lambda_0) = 0$. Since λ_0 is regular, then for sufficiently small $\varepsilon' \leq \varepsilon$ and $V' \subseteq V$, we may assume that all the conjugate covectors in V' are regular.

If $\lambda_0' \in \mathrm{Conj}(p)$ has order $m' \geq 1$ such that $(1, \lambda_0') \in \mathcal{V}'$, we can use Malgrange's Preparation Theorem again at λ_0' to find a neighbourhood $]1-\varepsilon', 1+\varepsilon'[\times V' \subseteq]1-\varepsilon, 1+\varepsilon[\times \mathcal{U}$ and smooth functions $c' :]1-\varepsilon', 1+\varepsilon'[\times V' \to \mathbb{R}, a_0', \dots, a_{m'-1}' : V' \to \mathbb{R}$ such that

$$\det(\mathrm{d}_{t\eta}\exp_p) = c'(t, \eta) \cdot ((t-1)^{m'} + a_{m'-1}'(\eta)(t-1)^{m'-1} + \cdots + a_0'(\eta)), \tag{11}$$

for every $(t, \eta) \in]1-\varepsilon', 1+\varepsilon'[\times U'$, with $c'(1, \lambda_0') \neq 0$, $a_0'(\lambda_0') = \cdots = a_{m'-1}'(\lambda_0') = 0$.

Now (10) and (11) both hold in the non-empty neighbourhood $]1-\varepsilon', 1+\varepsilon'[\times U'$ and therefore m must equal m'. Indeed, $a_0(\lambda_0') = 0$ since λ_0' is conjugate and the limit

$$\lim_{t\to 1} \frac{(t-1)^m + a_{m-1}(\lambda_0')(t-1)^{m-1} + \cdots + a_1(\lambda_0')(t-1)}{(t-1)^{m'}}$$

must exist and tend to $c'(1, \lambda_0')/c(1, \lambda_0')$. This implies that $m = m'$.

On the other hand, if a sequence $(\lambda_k)_{k\in\mathbb{N}}$ of conjugate covectors that have infinite order converges to λ_0, then λ_0 would also have an infinite order, since $t \mapsto \det(\mathrm{d}_{t\eta}\exp_p)$ is smooth for all $\eta \in U_p$. The above proves that the set of points of constant order is open and closed in U_p. □

Remark 28. The notion of order finiteness that we introduce in the section is related to those of ample and equiregular geodesics introduced in [3]. Indeed, if $\gamma(t) := \exp_p(t\lambda_0')$ is ample and equiregular for all λ_0' in a neighbourhood of λ_0 in $\mathrm{T}_p^*(M)$, then $\lambda_0 \in \mathrm{Conj}^R(p)$ must have finite order. In general, the set of $\mathcal{A} \subseteq \mathrm{T}^*(M)$ of (p, λ_0) such that the corresponding normal geodesic is ample and equiregular is non-empty and dense (see [2, Proposition 7.1]).

In the following theorem, we prove that in the neighbourhood of a conjugate covector that has finite order, the conjugate locus can be given a structure of submanifold.

Theorem 29. *If $\lambda_0 \in \mathrm{T}_p^*(M)$ is a regular conjugate covector that has finite order m, then there exists a neighbourhood U of λ_0 such that $\mathrm{Conj}_U(p)$ is a submanifold of codimension one in $\mathrm{T}_p^*(M)$. Moreover, $\mathrm{T}_{\lambda_0}(\mathrm{T}_p^*(M)) = \mathrm{T}_{\lambda_0}(\mathrm{Conj}_U(p)) \oplus \mathrm{Im}(\mathrm{r}_{p,\lambda_0})$.*

Proof. Consider the smooth function

$$\Delta_p^{m-1} : U_p \to \mathbb{R} : \lambda \mapsto \left.\frac{\mathrm{d}^{m-1}}{\mathrm{d}t^{m-1}}\right|_{t=1} \left(\det \mathrm{d}_{t\lambda}\exp_p\right).$$

From Proposition 27, we can find a neighbourhood $U \subseteq U_p$ of λ_0 such that all covectors in $\mathrm{Conj}_U(p)$ are regular and of order m. The inclusion $\mathrm{Conj}_U(p) \subseteq (\Delta_p^{m-1})^{-1}(0)$ thus holds. The derivative of Δ_p^{m-1} along the ray r_{p,λ_0} is non-zero. Indeed, since λ_0 is assumed to have order m,

$$\begin{aligned}
\mathrm{d}_1\mathrm{r}_{p,\lambda_0}(\Delta_p^{m-1}) &= \left.\frac{\mathrm{d}}{\mathrm{d}s}\right|_{s=1} \left(\left.\frac{\mathrm{d}^{m-1}}{\mathrm{d}t^{m-1}}\right|_{t=1} \left(\det \mathrm{d}_{ts\lambda_0}\exp_p\right)\right) \\
&= (m-1)\left.\frac{\mathrm{d}^{m-1}}{\mathrm{d}t^{m-1}}\right|_{t=1} \left(\det \mathrm{d}_{t\lambda_0}\exp_p\right) + \left.\frac{\mathrm{d}^{m}}{\mathrm{d}t^{m}}\right|_{t=1} \left(\det \mathrm{d}_{t\lambda_0}\exp_p\right) \\
&= \left.\frac{\mathrm{d}^{m}}{\mathrm{d}t^{m}}\right|_{t=1} \left(\det \mathrm{d}_{t\lambda_0}\exp_p\right) \neq 0.
\end{aligned}$$

Therefore a possibly smaller neighbourhood U of λ_0 in $\mathrm{T}_p^*(M)$, can be chosen to be convex and on which the radial derivative of Δ_p^{m-1} is non-zero. The function Δ_p^{m-1} is in particular smooth on U and has a non-zero differential.

Let us now show that $(\Delta_p^{m-1})^{-1}(0) \subseteq \mathrm{Conj}_U(p)$. Suppose $\lambda_0' \in U$ and $\Delta_p^{m-1}(\lambda_0') = 0$. The neighbourhood U has the property that each ray intersecting it has exactly one conjugate covector. Thus, on the ray passing through λ_0', there must be a (unique) conjugate covector λ_0''. Since U is convex, the line joining λ_0' to λ_0'' is also in U. The radial derivative of Δ_p^{m-1} along that line is non zero while $\Delta_p^{m-1}(\lambda_0') = 0$, by hypothesis, and $\Delta_p^{m-1}(\lambda_0'') = 0$, since all the conjugate vectors in U have the same finite order. Thus, by the mean value theorem, we must have $\lambda_0' = \lambda_0''$, and $\lambda_0' \in \mathrm{Conj}_U(p)$.

Finally, by the implicit function theorem, $(\Delta_p^{m-1})^{-1}(0) = \mathrm{Conj}_U(p)$ can be given the desired manifold structure. □

Remark 30. Under the hypothesis of Theorem 29, we have in particular that the set $\Pi_p = \{(t, \eta) \in]1-\varepsilon, 1+\varepsilon[\times U \mid t\eta \text{ is conjugate}\} \subseteq \mathcal{U}_p \subseteq \mathbb{R} \times \mathrm{T}_p^*(M)$ is also a submanifold for $\varepsilon > 0$ sufficiently small.

3.4. Non-local injectivity of the sub-Riemannian exponential map

Putting everything together, we finally prove in this section the relationship between the injectivity of the sub-Riemannian exponential map and its singularities.

We will need the following lemma, which is a consequence of Sard's theorem. It appeared in [22, Lemma II], and we provide here the proof for the sake of completeness.

Lemma 31. *Let X and Y be smooth manifolds with $\dim X = \dim Y$, and assume that X is compact. If f is a C^1 map from X to Y, then almost all points of Y have a finite preimage under f.*

Proof. Let us denote by $\mathrm{Crit}(f)$ the set of critical points of f, i.e.

$$\mathrm{Crit}(f) := \{x \in X \mid \mathrm{d}_x f \text{ has a non-trivial kernel}\}.$$

Let $y \in Y$ such that $f^{-1}(y)$ is infinite. Since X is assumed to be compact, we can deduce that $f^{-1}(y)$ has at least one accumulation point x. We write $(x_n) \subseteq f^{-1}(y)$ for a sequence, distinct from x, converging to x.

Let φ be a coordinate chart around x in X and ψ a coordinate chart around y in Y. We write $f_{\varphi\psi} = \psi \circ f \circ \varphi^{-1}$ for the expression of the map f in these coordinates. When n is large enough, we can be sure that $x_n \in \mathrm{Dom}(\varphi)$ and then consider the map

$$f(t) : [0,1] \to \mathbb{R}^{\dim Y} : t \mapsto f_{\varphi\psi}(\varphi(x) + t(\varphi(x_n) - \varphi(x))).$$

By the mean value theorem, there exists $t_n \in]0,1[$ such that

$$f_{\varphi\psi}(\varphi(x_n)) = f_{\varphi\psi}(\varphi(x)) + \mathrm{d}_{\varphi(x)+t_n(\varphi(x_n)-\varphi(x))} f_{\varphi\psi}[(\varphi(x_n) - \varphi(x))],$$

and therefore, by letting n tend to $+\infty$,

$$\mathrm{d}_{\varphi(x)} f_{\varphi\psi}[v] = 0, \text{ where } v = \lim_{n\to+\infty} \frac{\varphi(x_n) - \varphi(x)}{\|\varphi(x_n) - \varphi(x)\|} \neq 0.$$

Consequently, $\mathrm{d}_x f$ has a non-trivial kernel and x is a critical point, i.e. $y = f(x) \in f(\mathrm{Crit}(f))$. By Sard's theorem, the image of $\mathrm{Crit}(f)$ under f has measure zero in Y and the proof is complete. □

The study of the regularity of the conjugate locus in the previous section allows us to prove that the continuous one-form η_p is indeed exact.

Proposition 32. *If the family of extremals F_p simply covers $(1,q) \in \mathbb{R} \times M$, where $q := \exp_p(\lambda_0)$ and $\lambda_0 \in U_p \subseteq \mathrm{T}^*_p(M)$ is a regular conjugate vector of finite order, then Hilbert integral I_p is invariant on some neighbourhood of $(1,q)$ in $\mathbb{R} \times M$.*

Proof. By Proposition 19, the Hilbert integral I^*_p is path-independent on a small enough convex neighbourhood $\mathcal{U}$ of $(1,\lambda_0)$ in $\mathbb{R} \times \mathrm{T}^*_p(M)$. We would like to argue that as a consequence the same property holds for I_p.

Let $\Gamma(s) = (t(s), q(s))$ be a smooth closed curve in $F_p(\mathcal{U})$. If the corresponding curve $\Gamma^*(s) = (t(s), \lambda_0(s)) := F_p^{-1}(\Gamma(s))$ were to be smooth, we would be able to conclude this by Proposition 23. However, F_p is only a homeomorphism and thus we do not have enough regularity in general to evaluate the line integral I^*_p along Γ^*. This issue is addressed by introducing a coordinate system φ containing Γ, and by approximating in C^0 the loop $\varphi(\Gamma)$ by a sequence of parametrised polygons (p_n) in $\mathbb{R}^{n+1}$. These polygons may be chosen such that they intersect the conjugate locus only a finite number of times. Indeed, if $x \notin \Pi_p$, one can deduce from Lemma 31 that almost all rays passing through x in $\mathbb{R}^{n+1}$ intersect Π_p a finite number of times. Finally we may conclude by Proposition 23, letting n tends to $+\infty$, that

$$I_p[\Gamma] = \lim_{n\to+\infty} I_p[\varphi^{-1}(p_n)] = \lim_{n\to+\infty} I_p^*[\varphi^{-1}(p_n)] = 0,$$

that is to say, I_p is path independent. □

The integral I_p is the analogy of what is called *Hilbert invariant integral* in the classical calculus of variations. The fact that the family of extremals F_p induces the invariance of the Hilbert integral I_p corresponds to F_p defining what is commonly called a *(central) field of extremals*. The strategy now is to deduce that subsegment of normal geodesics are length-minimising if they are part of a field of extremals.

Proposition 33. *Let $\varepsilon > 0$ and $\gamma : [0, 1+\varepsilon[\to M$ be a strongly normal extremal starting at $p \in M$ and with initial covector $\lambda_0 \in U_p$. Assume that λ_0 is a first conjugate covector that is regular and has finite order. If the family of extremals F_p simply covers $(1, \exp_p(\lambda_0))$, then any subsegment of γ is a local length-minimiser among all the admissible trajectories with the same endpoints.*

Proof. The assumption that λ_0 is strongly normal and is a first conjugate covector implies, by adapting the proof of Proposition 12 with the conclusions of Theorem 9, that the neighbourhood on which F_p is injective can be chosen such that it contains the ray r_{p,λ_0}.

Consider an admissible trajectory $c : [t_0, t_1] \to M$ with control u', different from $\gamma|_{[t_0,t_1]}$, but with the same endpoints ($t_0, t_1 \in]0, 1+\varepsilon[$). We denote by $\Gamma'(t) = (t, c(t))$ the corresponding curves in $F_p(\mathcal{U})$. We also write $\lambda_0'(t)$ for the curve of initial covectors in $\mathrm{T}_p^*(M)$ such that $\Gamma'(t) = F_p(t, \lambda_0'(t))$, as well as the lift $\lambda'(t) := \mathrm{e}^{t\vec{H}}(p, \lambda_0'(t))$, as usual.

By definition of the sub-Riemannian Hamiltonian, we have that

$$H(\lambda'(t)) \geq \langle \lambda'(t), \dot{c}(t)\rangle - \frac{1}{2}|u'(t)|^2, \text{ for almost every } t \in [t_0, t_1],$$

as well as

$$H(\lambda(t)) = \langle \lambda(t), \dot{\gamma}(t)\rangle - \frac{1}{2}|u(t)|^2, \text{ for all } t \in [t_0, t_1],$$

since $\lambda(t)$ is the lift of the normal geodesic γ.

Therefore, we deduce that

$$\frac{1}{2}\mathrm{L}(\gamma|_{[t_0,t_1]})^2 = \frac{1}{2}\int_{t_0}^{t_1} |u(t)|^2 \mathrm{d}t = \int_{t_0}^{t_1} \left(\langle \lambda(t), \dot{\gamma}(t)\rangle - H(\lambda(t))\right) \mathrm{d}t = I_p[\Gamma] = I_p[c]$$

since the Hilbert integral I_p is path-independent by Proposition 32, and

$$\frac{1}{2}\mathrm{L}(\gamma|_{[t_0,t_1]})^2 = \int_{t_0}^{t_1} \left(\langle \lambda'(t), \dot{c}(t)\rangle - H(\overline{\lambda}(t))\right) \mathrm{d}t \leq \frac{1}{2}\int_{t_0}^{t_1} |u'(t)|^2 \mathrm{d}t = \frac{1}{2}\mathrm{L}(c)^2. \quad \square$$

We can now conclude recalling from Theorem 9 that a normal extremal that does not contain any abnormal subsegment cannot be length minimising past a conjugate point. The proof of Theorem 1 will be completed by the following Theorem.

Theorem 34. *Let M be a sub-Riemannian manifold and $p \in M$. If $\lambda_0 \in U_p \subseteq \mathrm{T}_p^*(M)$ is a strongly normal and regular conjugate covector of finite order, then the exponential map $\exp_p : U_p \to M$ is not injective in any neighbourhood of λ_0.*

Proof. Without loss of generality, we may assume that λ_0 is a first conjugate covector along the geodesic it generates. Since U_p is open, the geodesic $\gamma(t) := \exp_p(t\lambda_0)$ is well defined for all $t \in [0, 1+\varepsilon[$ for some $\varepsilon > 0$. If $\exp_p$ were to be injective in some neighbourhood of λ_0, then the family of extremals F_p would simply cover $(1, \exp_p(\lambda_0))$ by Proposition 12. The corresponding Hilbert integral I_p would be path independent by Proposition 32. Now, Proposition 33 would imply that the normal geodesic $\gamma : [0, 1+\varepsilon[\to M$ is a local length minimiser among all the admissible trajectories with the same endpoints, which would be in contradiction with Theorem 9 since λ_0 is assumed to be conjugate. □

In the classical theory of calculus of variations (see for example [17]), the fact that an extremal is part of a field of extremals is a sufficient condition for minimality is usually deduced from studying the Weierstrass E-function (see [24]). That argument is replaced in optimal control problems with the maximum principle. Let us mention two instances where we have found related uses of these techniques. In [19], fields of extremals are constructed in order to formulate sufficient conditions of optimality generalising Weierstrass' condition. The fact that normal extremal trajectories are locally minimising is proved by constructing a (non-central) field of extremals in [1, Section 4.7].

Let us finally discuss how the result Theorem 34 could be extended to the whole conjugate locus. If the regular conjugate covectors of finite order is dense in the regular conjugate locus, then by density the sub-Riemannian exponential map will also fail to be injective in any neighbourhood of a regular conjugate covector of infinite order. This is trivially the case if the sub-Riemannian manifold is analytic or if any geodesic $\gamma(t) := \exp_p(t\lambda_0)$ is ample and equiregular for all $(p, \lambda_0) \in \mathrm{T}^*(M)$. Furthermore, if the sub-Riemannian manifold is ideal then the property (R3) from Theorem 24 implies that regular conjugate covectors is dense in the conjugate locus, as in [26, Theorem 3.1].

4. Applications and final remarks

4.1. Conjugate points in metric geometry

The non-local injectivity of the sub-Riemannian exponential map implies a cotangent version of Shankar–Sormani's equivalence for synthetic notions of conjugate points in length space (see [23]) along a strongly normal geodesic with an initial covector that is regular conjugate and that has finite order.

Definition 35. Let X be a geodesic space and denote by $\mathcal{L}$ its length structure. If γ_n, γ are minimising geodesics parametrised on $[0, 1]$, we say that γ_n converges to γ if they converge for the metric

$$\mathrm{d}_{\mathrm{Geo}(X)}(\gamma_1, \gamma_2) := \sup_{t \in [0,1]} |\gamma_1(t) - \gamma_2(t)| + |\mathrm{L}(\gamma_1) - \mathrm{L}(\gamma_2)|.$$

If the geodesic space in question is a sub-Riemannian manifold M, then it is easy to see that a sequence of normal geodesics $\gamma_n(t) := \exp_{p_n}(t\lambda_0^n)$ converges to $\gamma(t) := \exp_p(t\lambda_0)$ if and only if p_n converges to p in M and λ_0^n converges to λ_0 in $\mathrm{T}_p^*(M)$.

Definition 36. Let X be a geodesic space, and a geodesic $\gamma : [0, 1] \to X$ joining two points p and q of X. We say that

(i) q is *one-sided conjugate to* p along γ if there exists a sequence of points (q_n) converging to q such that for every n, there are two distinct geodesics γ_n^1 and γ_n^2 joining p to q_n and both converging to γ;
(ii) p and q are *symmetrically conjugate along* γ if there exist sequences of points (p_n) converging to p and (q_n) converging to q such that for every n, there are two distinct geodesics γ_n^1 and γ_n^2 joining p_n to q_n and both converging to γ;

(iii) p and q are *unreachable conjugate along* γ if there are sequences of points (q_n) converging to q and (p_n) converging to p such that if γ_n is geodesic joining q_n to p_n for all n, then the sequence (γ_n) cannot converge to γ;

(iv) p and q are *ultimate conjugate points along* γ if they are symmetrically conjugate or unreachable conjugate along γ.

The relationship in sub-Riemannian geometry between these different definitions of conjugate points is given by the following theorem.

Theorem 37. *Let M be an ideal sub-Riemannian manifold, $\gamma : [0,1] \to M$ be a normal geodesic such that its initial covector λ_0 is regular conjugate and has finite order, and denote $p := \gamma(0)$ and $q := \gamma(1)$. Then, the following statements are equivalent:*

(i) *q is conjugate to p along γ;*
(ii) *q is one-sided conjugate to p along γ;*
(iii) *q is symmetrically conjugate to p along γ.*

Furthermore, if p and q are unreachable conjugate points along γ, then q is also conjugate to p along γ.

Remark 38. The ideal assumption in Theorem 37 is there to ensure that the sequences of geodesics in Definition 36 consist of normal extremals.

Proof. If (q_n) is a sequence of points converging to q and if (γ_n^1), (γ_n^2) are two sequences of geodesics joining p to q_n and converging to γ, then their initial covectors η_n^1 and η_n^2 will converge to λ_0 in $\mathrm{T}_p^*(M)$. By Theorem 34, the normal extremals γ_n^1 and γ_n^2 must coincide for n large enough. This proves that (i) implies (ii). It is easy to see from Definition 36 that (ii) immediately implies (iii).

Consider the function

$$E : \mathrm{T}^*(M) \to M \times M : (x, \eta) \mapsto (x, \exp_p(\eta)),$$

which has a differential at (p, λ_0) that is a linear isomorphism if q is not conjugate to p along γ. By the inverse function theorem, there exists a neighbourhood of (p, λ_0) on which the function E is a diffeomorphism. Suppose now that q is one-sided conjugate to p along γ, so that there exists a sequence of points (q_n) converging to q and for all n two distinct normal extremals γ_n^1 and γ_n^2, with initial covectors η_n^1 and η_n^2 respectively, joining p to q_n and converging to γ. The proof of (ii) implies (i) follows by observing that $E(p, \eta_n^1) = E(p, \eta_n^2)$ for n large enough while $\eta_n^1 \neq \eta_n^2$ since γ_n^1 and γ_n^2 are distinct. By the same argument, it can be seen that (iii) implies (i). The last statement about unreachable conjugate points is the same as in [23]. □

4.2. Structure of the sub-Riemannian cut locus

Theorem 1 is also related to the structure of the sub-Riemannian cut locus. Let M be an ideal sub-Riemannian manifold and define the cut time of $(p, \lambda_0) \in \mathrm{T}_p^*(M)$ as

$$t_{\mathrm{cut}}(p, \lambda_0) := \sup\{t > 0 \mid \exp_p(\cdot \lambda_0)|_{[0,t]} \text{ is a length-minimising geodesic}\}.$$

The (cotangent) cut locus is then

$$\mathrm{Cut}(p) := \{\lambda_0 \in \mathrm{T}_p^*(M) \mid t_{\mathrm{cut}}(p, \lambda_0) = 1\}.$$

and we denote by $\mathrm{Cut}^1(p)$ the subset of $\mathrm{Cut}(p)$ consisting of those covectors λ_0 for which there exists another $\lambda_0' \in \mathrm{Cut}(p)$ such that $\exp_p(\lambda_0) = \exp_p(\lambda_0')$ and $\lambda_0 \neq \lambda_0'$. Note that by [1, Theorem 8.72], if

$\lambda_0 \in \mathrm{Cut}(p) \setminus \mathrm{Cut}^1(p)$, then $\lambda_0 \in \mathrm{Conj}(p)$. In this specific case, we are able to prove the local non-injectivity property of the exponential map without extra assumptions and without invoking the regularity of the conjugate locus.

Theorem 39. *Let M be an ideal sub-Riemannian manifold, and $p \in M$. If $\lambda_0 \in \mathrm{Cut}(p) \setminus \mathrm{Cut}^1(p)$, then the exponential map $\exp_p$ fails to be injective in any neighbourhood of λ_0.*

Proof. Let $\gamma(t) := \exp_p(t\lambda_0)$ be the unique length minimising geodesic between p and $q := \exp_p(\lambda_0)$. Let u be its minimal control and $\lambda_1 \in \mathrm{T}_p^*(M)$ its Lagrange multiplier. By [1, Corollary 8.74], we find a sequence of points $q_k \in M$ converging to q such that for each k there are two distinct length minimisers γ_k^1 and γ_k^2 joining p and q_k. We denote by $\lambda_{1,k}^1$ and $\lambda_{1,k}^2$ the normal Lagrange multipliers of γ_k^1 and γ_k^2 respectively, as well as u_k^1 and u_k^2 for their respective minimal control. We write $\lambda_{0,k}^1$ and $\lambda_{0,k}^2$ for their initial covectors. They satisfy $\exp_p(\lambda_{0,k}^1) = \exp_p(\lambda_{0,k}^2)$ and we would like to prove that $\lambda_{0,k}^1$ and $\lambda_{0,k}^2$ both converge to λ_0 as $k \to +\infty$.

Modulo extraction of a subsequence, we may assume by compactness of length minimisers ([1, Proposition 8.67]) that u_k^1 and u_k^2 converge in the strong L^2 topology, as well as γ_k^1 and γ_k^2 converge uniformly to a geodesic joining p and q. This is unique by our assumption. So, γ_k^1 and γ_k^2 converge uniformly to γ and u_k^1 and u_k^2 converge to u.

Choose any metric $|\cdot|$ on $\mathrm{T}_p^*(M)$ that we only use to prove estimates. We want to show that $\lambda_{1,k}^1$ is convergent. Assume that it is not, this would mean that there exists a subsequence of $|\lambda_{1,k}^1|$ that diverges to $+\infty$. Now, $\eta_{1,k}^1 := \lambda_{1,k}^1/|\lambda_{1,k}^1|$ of course converges to some η_1^1. The Lagrange multiplier rule ([1, Section 8.3]) for γ_k^1 implies that

$$\frac{\lambda_{1,k}^1}{|\lambda_{1,k}^1|} D_{u_k} E_p = \frac{u_k}{|\lambda_{1,k}^1|}$$

and thus, by taking the limit, $\eta_1^1 D_u E_p = 0$. Here, we have written E_p for the endpoint map (see [1, Section 8.1]). We obtain that η_1^1 is an abnormal Lagrange multiplier but this is impossible since we assumed γ to be non-abnormal.

Therefore, $\lambda_{1,k}^1$ must be convergent and by taking again the limit in the Lagrange multiplier rule

$$\lambda_{1,k}^1 D_{u_k} E_p = u_k$$

we obtain that its limit must a (normal) Lagrange multiplier for γ. This is necessarily unique since γ is not abnormal and thus $\lambda_{1,k}^1 \to \lambda_1$, as well as $\lambda_{1,k}^2 \to \lambda_1$ by the same argument. This also implies that $\lambda_{0,k}^1 \to \lambda_0$ and $\lambda_{0,k}^2 \to \lambda_0$ since initial and final covectors are linked by the Hamiltonian flow. □

Corollary 40. *Let M be an ideal sub-Riemannian manifold, and $p \in M$. The set $\mathrm{Cut}^1(p)$ is dense in $\mathrm{Cut}(p)$.*

Proof. Let $\lambda_0 \in \mathrm{Cut}(p)\setminus\mathrm{Cut}^1(p)$ and consider a sequence of decreasing open neighbourhoods $(U_n) \subseteq \mathrm{T}_p^*(M)$ such that $\cap_n U_n = \{\lambda_0\}$. Then, by Theorem 34, the exponential map $\exp_p$ fails to be injective on each U_n and so for every n, the intersection $\mathrm{Cut}^1(p) \cap U_n$ is non empty. We therefore find a sequence of covectors $\lambda_n \in \mathrm{Cut}^1(p)$ such that $\lambda_n \to \lambda_0$. □

Corollary 40 can be used to prove that the map $\mathrm{d}(p,\cdot)^2$ is smooth in a neighbourhood of q if and only if there is a unique length-minimiser γ joining p to q and q is not conjugate to p along γ. In the non-ideal case, one usually bypass this type of argument to study the regularity of the sub-Riemannian squared distance (see [1, Chapter 11]).

4.3. Final remarks

As mentioned in the introduction, a proof of the local non-injectivity of the Riemannian exponential map can be found in [18,22], and [26]. There is also an argument towards this result in [13], and we explain here why it is incomplete. The Ref. [13, Theorem 2.1.12] argues by contradiction, assuming that $\exp_p$ is injective in a neighbourhood of a conjugate vector v (and thus it is a homeomorphism by invariance of domain). Here we use notation of the reference. A curve b that has the same endpoints as the geodesic c along which p is conjugate to $\exp_p(v)$ is constructed so that one may use standard arguments using Jacobi fields to conclude that $\mathrm{L}(b) < \mathrm{L}(c)$. The last (and flawed) step consists in defining the curve $\tilde{b} := \exp_p^{-1}(b)$ in $\mathrm{T}_p(M)$ and to invoke a corollary of Gauss' lemma from [13, Theorem 1.9.2] to conclude that $\mathrm{L}(b) \geq \mathrm{L}(c)$. However, this is exactly where things break down. This very last step can only be performed if $\tilde{b}$ is regular enough, say (piecewise) smooth. However, $\exp_p$ is only a homeomorphism, and not a diffeomorphism. This is in fact the heart of the problem that needs to be overcome (see Remark 21), and the reason why a more complex argument is necessary.

On another note, it seems possible that a conjugate covector has infinite order in the sense of Definition 26, even if the sub-Riemannian structure is ideal. It would be interesting to investigate further this property, and to find relevant examples. For instance, is it possible to construct an example for which $\mathrm{Conj}(p)$ contains only conjugate covectors of infinite order? In particular, the proof of Theorem 34 relies on Theorem 24, which says that $\mathrm{Conj}(p)$ is a smooth hypersurface near a regular conjugate covector of finite order. In the case of infinite order, we do not know if $\mathrm{Conj}(p)$ can still be endowed with such a manifold structure.

Let us remark again that Warner's proof of the Riemannian analogous to Theorem 34 is different to what we have done here. At the core of Warner's proof is Whitney's singularity theory, used to find the normal forms of the exponential map near a conjugate vector, which precisely works there because the Riemannian Jacobi equation is a second order differential equation. It is unclear if it is possible to have such a general description of the normal forms of the sub-Riemannian exponential map, again because of the difference between the multiplicity of a conjugate vector and what we have called the order of a conjugate vector. Nevertheless, we have successfully pursued this approach for some specific examples: for the Heisenberg group in [8], and for the α-Grushin plane, $\mathrm{SU}(2)$ and $\mathrm{SL}(2)$ in [7]. We expect this approach to work in other cases, possibly with techniques similar to what has been done for the 3D contact case in [1, Chapters 17 to 19] (and the references therein), and in [4], where the authors classify the singularities of the sub-Riemannian exponential for low-dimensional generic structures.

Finally, we have not dealt with abnormal geodesics. For example, it would be interesting to understand if the metric notions of conjugate points make sense along abnormal geodesics, and if they are still equivalent.

Acknowledgements

We would like to thank Andrei Agrachev and Luca Rizzi for their interest and comments. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 945655).

References

[1] Andrei Agrachev, Davide Barilari, Ugo Boscain, A Comprehensive Introduction to Sub-Riemannian Geometry, in: Cambridge Studies in Advanced Mathematics, vol. 181, Cambridge University Press, Cambridge, 2020, p. xviii+745, From the Hamiltonian viewpoint, With an appendix by Igor Zelenko. 3971262.

[2] Andrei A. Agrachev, Davide Barilari, Elisa Paoli, Volume geodesic distortion and Ricci curvature for Hamiltonian dynamics, Ann. Inst. Fourier (Grenoble) 69 (3) (2019) 1187–1228, URL: http://aif.cedram.org/item?id=AIF_2019__69_3_1187_0. 3986913.

[3] Andrei Agrachev, Davide Barilari, Luca Rizzi, Curvature: A variational approach, Mem. Amer. Math. Soc. 256 (1225) (2018) v+142, http://dx.doi.org/10.1090/memo/1225, 3852258.

[4] Davide Barilari, Ugo Boscain, Grégoire Charlot, Robert W. Neel, On the heat diffusion for generic Riemannian and sub-Riemannian structures, Int. Math. Res. Not. IMRN (15) (2017) 4639–4672, http://dx.doi.org/10.1093/imrn/rnw141, 3685111.

[5] Davide Barilari, Luca Rizzi, On Jacobi fields and a canonical connection in sub-Riemannian geometry, Arch. Math. (Brno) 53 (2) (2017) 77–92, http://dx.doi.org/10.5817/AM2017-2-77, 3672782.

[6] Samuël Borza, Distortion coefficients of the α-Grushin plane, J. Geom. Anal. 32 (3) (2022) http://dx.doi.org/10.1007/s12220-021-00736-8, Paper No. 78, 284363751.

[7] Samuël Borza, Normal forms for the sub-Riemannian exponential map of $\mathbb{G}_\alpha$, SU(2), and SL(2), 2023, arXiv:2302.00524.

[8] Samuël Borza, Wilhelm Klingenberg, Regularity and continuity properties of the sub-Riemannian exponential map, J. Dynam. Control Syst. (2023) http://dx.doi.org/10.1007/s10883-022-09624-y.

[9] Wei-Liang Chow, Über Systeme von linearen partiellen Differentialgleichungen erster Ordnung, Math. Ann. 117 (1939) 98–105, http://dx.doi.org/10.1007/BF01450011, 1880.

[10] Martin Golubitsky, Victor Guillemin, Stable Mappings and their Singularities, in: Graduate Texts in Mathematics, Vol. 14, Springer-Verlag, New York-Heidelberg, 1973, p. x+209, 0341518.

[11] Mikhael Gromov, Carnot-Carathéodory spaces seen from within, in: Sub-Riemannian Geometry, in: Progr. Math., vol. 144, Birkhäuser, Basel, ISBN: 3-7643-5476-3, 1996, pp. 79–323, 1421823.

[12] Lars Hörmander, Hypoelliptic second order differential equations, Acta Math. 119 (1967) 147–171, http://dx.doi.org/10.1007/BF02392081.

[13] Wilhelm P. A. Klingenberg, Riemannian Geometry, second ed., in: De Gruyter Studies in Mathematics, vol. 1, Walter de Gruyter & Co., Berlin, ISBN: 3-11-014593-6, 1995, p. x+409, http://dx.doi.org/10.1515/9783110905120, 1330918.

[14] Thomas Mietton, Luca Rizzi, Branching geodesics in sub-Riemannian geometry, Geom. Funct. Anal. 30 (4) (2020) 1139–1151, http://dx.doi.org/10.1007/s00039-020-00539-z, 4153911.

[15] Richard Montgomery, A Tour of Subriemannian Geometries, their Geodesics and Applications, in: Mathematical Surveys and Monographs, vol. 91, American Mathematical Society, Providence, RI, ISBN: 0-8218-1391-9, 2002, p. xx+259, http://dx.doi.org/10.1090/surv/091, 1867362.

[16] Marston Morse, The order of vanishing of the determinant of a conjugate base, Proc. Natl. Acad. Sci. USA 17 (5) (1931) 319.

[17] Marston Morse, The Calculus of Variations in the Large, in: American Mathematical Society Colloquium Publications, vol. 18, American Mathematical Society, Providence, RI, ISBN: 0-8218-1018-9, 1996, p. xii+368, http://dx.doi.org/10.1090/coll/018.

[18] Marston Morse, Sebastian B. Littauer, A characterization of fields in the calculus of variations, Proc. Natl. Acad. Sci. USA 18 (12) (1932) 724.

[19] Andrzej Nowakowski, Field theories in the modern calculus of variations, Trans. Amer. Math. Soc. 309 (2) (1988) 725–752, http://dx.doi.org/10.2307/2000935, 961610.

[20] N.S. Pontryagin, Vladimir Grigorievich Boltyanski, M.V. Gamkrelidze, Yevgeny Frolovich Mishchenko, The Mathematical Theory of Optimal Processes, Nauka, 1962.

[21] P.K. Rashevsky, About connecting two points of a completely nonholonomic space by admissible curve, Uch. Zapiski Ped. Inst. Libknechta 2 (1938) 83–94.

[22] Leonard J. Savage, On the crossing of extemals at focal points, Bull. Amer. Math. Soc. 49 (1943) 467–469, http://dx.doi.org/10.1090/S0002-9904-1943-07960-8, 9101.

[23] Krishnan Shankar, Christina Sormani, Conjugate points in length spaces, Adv. Math. 220 (3) (2009) 791–830, http://dx.doi.org/10.1016/j.aim.2008.10.002, 2483228.

[24] John F. Shaw, The Weierstrass E-Function, McGill University, 1966, URL: https://central.bac-lac.gc.ca/.item?id=TC-QMM-47612&op=pdf&app=Library.

[25] Robert S. Strichartz, Sub-Riemannian geometry, J. Differ. Geom. 24 (2) (1986) 221–263, URL: http://projecteuclid.org/euclid.jdg/1214440436. 862049.

[26] Frank W. Warner, The conjugate locus of a Riemannian manifold, Amer. J. Math. 87 (1965) 575–604, http://dx.doi.org/10.2307/2373064, 208534.