Università degli Studi di Pisa

Dipartimento di Matematica

Corso di Laurea Magistrale in Matematica

New regularity results for sub-Riemannian geodesics

Tesi di Laurea Magistrale

Candidato
Alessandro Pigati

Relatore
Prof. Luigi Ambrosio

Relatore
Prof. Davide Vittone

Controrelatore
Prof. Valentino Magnani

Anno Accademico 2015/2016

Contents

CHAPTER 1

Introduction

This thesis deals mainly with a famous open problem in sub-Riemannian geometry and geometric control theory, namely the regularity of *geodesics* in sub-Riemannian manifolds, which are also called *length minimizers*. Roughly speaking, a sub-Riemannian manifold is a smooth manifold where, at each point, a vector space of admissible directions is assigned, as well as a positive definite inner product on this space, in a smooth fashion. The resulting distribution is required to satisfy the so-called *Hörmander condition*, which is diametrically opposed to the integrability condition. *Horizontal curves* are absolutely continuous curves whose speed is admissible at a.e. time. Defining their length in the usual way, a horizontal curve is then said to be a *length minimizer* if it minimizes the length among all horizontal curves connecting its endpoints.

The regularity problem arose soon after the paper [Str86] by Strichartz was published, thirty years ago. In this paper, it was claimed that all constant-speed length minimizers correspond to solutions to a suitable Hamiltonian system in T^*M, as in the case of Riemannian geometry, and therefore they are always smooth. The proof of this false assertion depended on a flawed application of the celebrated Pontryagin Maximum Principle: basically, the author forgot to treat the so-called *abnormal* case. Later, the author admitted that his paper contained an irreparable mistake.

Some years later, in [Mon94], Montgomery gave the first example of a length minimizer which does not come from the aforementioned Hamiltonian framework. After that, many other examples were found, showing that the nature of length minimizers is much more subtle. It is still an open problem whether constant-speed geodesics, which a priori are only $W^{1,\infty}$-regular, are always smooth (i.e. C^∞-regular) in any sub-Riemannian manifold, or even whether they are always C^1-regular. This problem is open in the model case of *Carnot groups*, as well. This special class of sub-Riemannian manifolds consists of Lie groups whose Lie algebra is stratified. They provide an infinitesimal model for any sub-Riemannian manifold, near any given point (provided the point satisfies a technical condition, which holds generically).

Some partial results are known: if the assigned distribution has *step* at most 2, then all geodesics are smooth (and, in fact, the claim by Strichartz is true in this case). In the context of Carnot groups, the regularity problem has recently been solved also when the step is at most 3 (independently by Tan-Yang in [TY13] and by Le Donne-Leonardi-Monti-Vittone in [LLMV13]). In [Sus14] Sussmann proved that, in presence of analytic data (and in particular in Carnot groups), all geodesics are *analytic* on a dense open set of times, although it is not known whether this set has full measure. Finally, one year ago, the first general regularity result was obtained by Hakavuori-Le Donne (and will appear in [HL16]): in this paper, which builds on the ideas contained in [LM08], it is proved that geodesics cannot have *corner-like singularities*.

The thesis is organized as follows.

- In Chapter 2, besides defining sub-Riemannian manifolds, we prove some basic metric and topological facts, such as the classical Chow-Rashevsky theorem, which asserts that they are connected by horizontal curves. We also introduce the notion of *control* and we derive the first order optimality conditions for geodesics, which allow to classify them as *normal* and *abnormal ones*. Finally, we prove that the solutions of the Hamiltonian system considered by Strichartz are locally length minimizers.

- In Chapter 3 we focus on the special case of Carnot groups and we prove several useful properties of them which are used later, in Chapter 5. We also prove the smoothness of all geodesics when the step is at most 3, following [LLMV13].
- In Chapter 4 we develop a well-known *second order theory* for C^2-regular maps, which is based on a generalization of the Morse index, and we apply it to obtain further optimality conditions for geodesics, known as the *Goh conditions* (which suffice to prove the regularity when the step is at most 2). Then, following [AS95] (and correcting a mistake in this paper), we prove a theorem by Liu-Sussmann which shows the local minimality of a very general class of curves. We apply this result to obtain explicit examples of strictly abnormal geodesics, such as the example by Golé-Karidi (which appeared in [GK95]) in the context of Carnot groups.
- In Chapter 5 we revisit the proof of the aforementioned result by Hakavuori-Le Donne, giving a presentation which is somewhat more transparent than the original one. Then, we obtain a quantitative refinement, which allows us to exclude a wider class of singularities when the manifold is a Carnot group of rank 2. In order to obtain this improvement, we exploit the notion of *excess* and the *compactness* of unit-speed length minimizers, as well as a careful choice of the scales at which a suitable correction technique is applied. Using similar methods, we are able to prove the lack of minimality for an interesting family of horizontal curves.
- Finally, the Appendix is mainly devoted to a revisitation of the classical Cauchy-Lipschitz theory for ODEs, in the generality needed in this thesis, and to the proof of some useful facts concerning flows of time-dependent vector fields. We also present the global Baker-Campbell-Hausdorff formula for nilpotent groups.

CHAPTER 2

Sub-Riemannian manifolds and first order theory

2.1. Sub-Riemannian manifolds

The aim of this section is to introduce the general setting where our regularity problem takes place, namely *sub-Riemannian manifolds*. Later on in the thesis we will specialize in the study of Carnot groups, which represent a model case and can be viewed as a sort of first-order approximation of any sub-Riemannian structure at some given point (see Section 3.4).

Informally speaking, a sub-Riemannian manifold is a smooth manifold where we can move, at an infinitesimal scale, only in a prescribed set of directions, depending on the particular point where we are located. Moreover, each such direction is given a norm, as on a Riemannian manifold: this in turn will enable us to define the length of a path between two points. Let us now give the precise definitions.

Definition 2.1. Given a smooth manifold M, a smooth *distribution* $\mathcal{D}$ with rank $r > 0$ is a smooth subbundle of the tangent bundle, i.e. a map $x \mapsto \mathcal{D}_x$ which assigns to each point $x \in M$ an r-dimensional vector subspace of T_xM, in a way such that locally we can write $\mathcal{D}_x = \langle X_1(x), \dots, X_r(x) \rangle$ for some smooth vector fields $X_1, \dots, X_r$. A *metric* g on $\mathcal{D}$ is a smooth assignment of a positive definite scalar product g_x on each vector space $\mathcal{D}_x$ (here *smooth* means that, whenever $\mathcal{D} = \langle X_1, \dots, X_r \rangle$ on an open subset $U \subseteq M$, the map $x \mapsto g_x(X_i, X_j)$ is smooth on U, for any i, j).

Definition 2.2. We denote by $\Gamma(TM)$ the space of all smooth vector fields on M. Given a smooth distribution $\mathcal{D}$, we define inductively $\mathrm{Lie}^k(\mathcal{D}) \subseteq \Gamma(TM)$ for $k \geq 1$ as follows:

$$\mathrm{Lie}^1(\mathcal{D}) := \{X \in \Gamma(TM) : \forall x \in M \ X(x) \in \mathcal{D}_x\},$$

$$\mathrm{Lie}^{k+1}(\mathcal{D}) := \left\{ X + \sum_{i=1}^{N} [Y_i, Z_i] \mid Y \in \mathrm{Lie}^1(\mathcal{D}), X, Z \in \mathrm{Lie}^k(\mathcal{D}) \right\}.$$

Of course the number N of terms in the last sum is allowed to vary freely. $\mathrm{Lie}^{k+1}(\mathcal{D})$ can be equivalently defined as the real vector subspace of $\Gamma(TM)$ generated by $\mathrm{Lie}^k(\mathcal{D})$ and $\big\{[Y, Z] \mid Y \in \mathrm{Lie}^1(\mathcal{D}), Z \in \mathrm{Lie}^k(\mathcal{D})\big\}$. We also set

$$\mathrm{Lie}^\infty(\mathcal{D}) := \bigcup_{k=1}^{\infty} \mathrm{Lie}^k(\mathcal{D}).$$

Remark 2.3. The spaces $\mathrm{Lie}^k(\mathcal{D})$ form an increasing sequence of vector subspaces of $\Gamma(TM)$. Moreover, it is easy to prove inductively that $\mathrm{Lie}^k(\mathcal{D})$ is in fact a $C^\infty(M)$-submodule of $\Gamma(TM)$, i.e. it is also closed under multiplication by a smooth function. This is clear when $k = 1$ and in general it follows from the fact that $a[Y, Z] = [aY, Z] + Z(a)Y$.

Remark 2.4. If there are $X_1, \dots, X_r$ such that $\mathcal{D} = \langle X_1, \dots, X_r \rangle$, it is immediate to verify inductively (using the identity $[aY, bZ] = ab[Y, Z] + aY(b)Z - bZ(a)Y$) that $\mathrm{Lie}^k(\mathcal{D})$ is generated, as a $C^\infty(M)$-module, by all possible j-fold iterated Lie brackets

$$[X_{i_1}, [\cdots, [X_{i_{j-1}}, X_{i_j}] \cdots]]$$

of these vector fields, as j varies from 1 to k.

Definition 2.5. For any $1 \le k \le \infty$ we set

$$\mathrm{Lie}^k(\mathcal{D}, x) := \Big\{ X(x) \mid X \in \mathrm{Lie}^k(\mathcal{D}) \Big\}.$$

Remark 2.6. It is clear that $\mathrm{Lie}^1(\mathcal{D}, x) = \mathcal{D}_x$. Furthermore, the definition of $\mathrm{Lie}^k(\mathcal{D}, x)$ is local, in the following sense. If $U \subseteq M$ is an open subset, U is endowed with the restricted distribution $\mathcal{D}|_U$ and we can form the $C^\infty(U)$-submodules $\mathrm{Lie}^k\left(\mathcal{D}|_U\right)$ of $\Gamma(TU)$. Then, for any $x \in U$ and any $k \ge 1$, we have

$$\mathrm{Lie}^k(\mathcal{D}, x) = \mathrm{Lie}^k\left(\mathcal{D}|_U, x\right).$$

Indeed, the inclusion $\mathrm{Lie}^k(\mathcal{D}, x) \subseteq \mathrm{Lie}^k\left(\mathcal{D}|_U, x\right)$ is clear. Moreover, one can immediately check, by induction on k, that any $X \in \mathrm{Lie}^k\left(\mathcal{D}|_U\right)$ coincides with a suitable $X' \in \mathrm{Lie}^k(\mathcal{D})$ on a fixed neighbourhood $V \Subset U$ of x (for the base case $k = 1$ it suffices to let $X' := \eta X$, for any $\psi \in C_c^\infty(U)$ such that $\psi \equiv 1$ on V, extending X' by zero outside U).

Definition 2.7. A smooth distribution $\mathcal{D}$ is said to be *bracket-generating* (or *totally non-holonomic*, or to satisfy the *Hörmander condition*) if for any $x \in M$ we have $\mathrm{Lie}^\infty(\mathcal{D}, x) = T_xM$. Notice that, as $\left(\mathrm{Lie}^k(\mathcal{D}, x)\right)_{k \ge 1}$ is an increasing sequence of vector subspaces whose union is $\mathrm{Lie}^\infty(\mathcal{D}, x)$, an equivalent condition is that, for any $x \in M$, there exists a finite $k \ge 1$ such that $\mathrm{Lie}^k(\mathcal{D}, x) = T_xM$.

Definition 2.8. A *sub-Riemannian manifold* M is a smooth, connected n-dimensional manifold, equipped with a bracket-generating distribution $\mathcal{D}$ and with a smooth metric g defined on $\mathcal{D}$. In the sequel, for vectors $v, w \in \mathcal{D}$ we will often use the notation $\langle v, w\rangle$ instead of $g(v, w)$, as well as the shorthand $|v| := g(v, v)^{1/2}$. r is called the *rank* of the sub-Riemannian structure, while the *step* is the least $s \le \infty$ such that $\mathrm{Lie}^s(\mathcal{D}, x) = T_xM$ for any $x \in M$.

Remark 2.9. When $r = n$, the bracket-generating condition is trivially satisfied. This special case corresponds to Riemannian geometry, where a metric is given on the whole tangent bundle TM.

Definition 2.10. A curve $\gamma \in H^1([0, T], M)$ is said to be *horizontal* (or *admissible*) if $\dot\gamma(t) \in \mathcal{D}_{\gamma(t)}$ for a.e. $t \in [0, T]$. In such case, its *length* is $L(\gamma) := \int_0^T |\dot\gamma(t)|\, dt$, while its *energy* is $E(\gamma) := \frac{1}{2}\int_0^T |\dot\gamma(t)|^2\, dt$. The *Carnot-Carathéodory* distance between two points $x, y \in M$ is

$$d_{CC}(x, y) := \inf_{\gamma \in \Omega_{x,y}} L(\gamma),$$

where $\Omega_{x,y}$ is the set of horizontal curves $\gamma \in H^1([0, 1], M)$ with $\gamma(0) = x$ and $\gamma(1) = y$. We will often simply write d in place of d_{CC}.

Since the length of a horizontal curve is clearly invariant under linear reparametrizations, we would obtain an equivalent definition of d_{CC} by replacing $[0, 1]$ with $[0, T]$ in the definition of $\Omega_{x,y}$, with variable T.

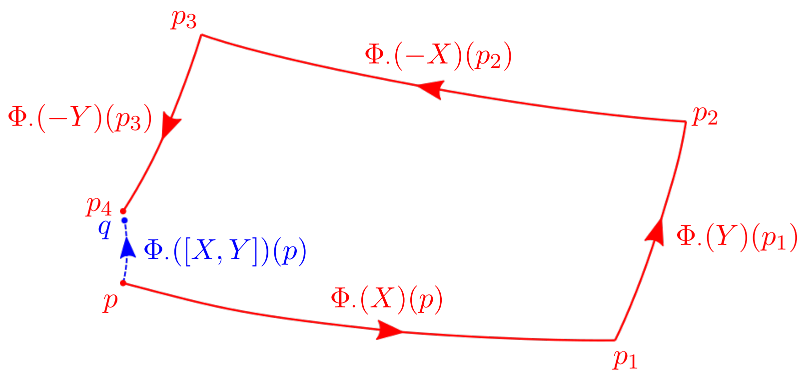
In the last definition, $H^1([0, T], M)$ is the space of continuous curves which in local coordinates belong to H^1. For a precise definition of this space, see Definition B.1 in the appendix. We notice that the same definition of length makes sense more generally for horizontal curves in $AC([0,1],M) = W^{1,1}([0,1],M)$.

DEFINITION 2.11. We use the notation $\mathbb{B}_r(x) := \{y \in M : d_{CC}(x,y) < r\}$ for the Carnot-Carathéodory ball with center x and radius $r > 0$, in order to distinguish it from the usual Euclidean ball $B_r(x)$ when $M = \mathbb{R}^n$.

Let us motivate heuristically the requirement that $\mathcal{D}$ satisfies the bracket-generating condition. Take two vector fields $X, Y \in \mathrm{Lie}^1(\mathcal{D})$. Let us denote by $\Phi_t(X)$ the flow map at time t associated to X, so that $t \mapsto \Phi_t(X)(p)$ is an integral curve for X and is horizontal. The Lie bracket has this geometrical meaning: suppose we move for a short time t along X, then along Y, then back along $-X$ and finally along $-Y$. The outcome of this maneuver is a null displacement at the first order in t, but at the second order a Lie bracket $[X,Y]$ appears. More precisely, as we will show below (see Proposition 2.37), we have

$$\Phi_t(-Y) \circ \Phi_t(-X) \circ \Phi_t(Y) \circ \Phi_t(X)(p) = \Phi_{t^2}([X,Y])(p) + o(t^2),$$

in any fixed local coordinate system (so that the last sum has a meaning).



In the picture we have set $p_1 := \Phi_t(X)(p)$, $p_2 := \Phi_t(Y)(p_1)$, $p_3 := \Phi_t(-X)(p_2)$, $p_4 := \Phi_t(-Y)(p_3)$ and $q := \Phi_{t^2}([X,Y])(p)$. Thus, we can approximate displacements in the direction $[X,Y]$ and iteration of this operation should allow us to move in any direction (but at the expense of using very long horizontal paths). So the bracket-generating condition can be viewed as an infinitesimal condition which should guarantee that any two points of M can be connected by a horizontal path. This is indeed the case and will be proved in Section 2.4.

In fact, we do not yet know that the Carnot-Carathéodory is finite, nor that $d(x,y) = 0$ implies $x = y$. These facts will be proved in the next sections.

2.2. Some basic metric properties

It is often useful to work with constant-speed paths, instead of arbitrary AC or H^1 ones. We now show that it is always possible to reparametrize an AC curve so that the new curve has constant speed.

REMARK 2.12. If $\gamma \in AC([0,T],M)$ is horizontal, $h \in AC([0,\tau])$ is increasing and $h(0) = 0$, $h(\tau) = T$, then $\gamma \circ h \in AC([0,\tau],M)$ is horizontal as well and satisfies $L(\gamma \circ h) = L(\gamma)$: the fact that $\gamma \circ h$ is AC and horizontal follows from Lemma A.8, while

$$L(\gamma \circ h) = \int_0^\tau |\dot\gamma \circ h|\, \dot h \, d\mathcal{L}^1 = \int_0^T |\dot\gamma| \, d\mathcal{L}^1$$

by Lemma A.7.

We will prove that every horizontal curve has this form, for some h and some horizontal γ with $|\dot{\gamma}| = 1$. Before doing that, it is convenient to extend (non-canonically) the metric g to the whole of TM.

LEMMA 2.13. There exists a smooth Riemannian metric g' on TM such that $g = g'|_{\mathcal{D}}$, i.e. such that $g(v, w) = g'(v, w)$ whenever $v, w \in \mathcal{D}$.

Proof. To begin with, we endow M with an arbitrary Riemannian metric h. This enables us to define a complementary distribution $\mathcal{D}' := \mathcal{D}^\perp$, where the orthogonal is taken pointwise in TM with respect to h.

We claim that $\mathcal{D}'$ is smooth as well: locally we can write $\mathcal{D} = \langle X_1, \ldots, X_r \rangle$ and, applying the Gram-Schmidt algorithm, we can assume that $X_1, \ldots, X_r$ are orthonormal with respect to h. Since we are arguing locally, we can also find $X_{r+1}, \ldots, X_n$ such that $(X_1, \ldots, X_n)$ is pointwise a basis of the tangent space. Now

$$\mathcal{D}' = \left\langle X_i - \sum_{j=1}^{r} h(X_i, X_j) X_j \mid i = r+1, \ldots, n \right\rangle,$$

proving the smoothness of $\mathcal{D}'$.

We set $g' := g \oplus (h|_{\mathcal{D}'})$, i.e. given $v, w \in TM$ we decompose them as $v = v_1 + v_2$, $w = w_1 + w_2$, with $v_1, w_1 \in \mathcal{D}$ and $v_2, w_2 \in \mathcal{D}'$, and we define

$$g'(v, w) := g(v_1, w_1) + h(v_2, w_2).$$

g' is the required smooth Riemannian metric (notice that the map $v \mapsto v_1$ is smooth, since locally, if $X_1, \ldots, X_r \in \mathcal{D}$ are orthonormal with respect to h as before, $v_1 = \sum_{j=1}^{r} h(v, X_j) X_j$; hence $v \mapsto v_2 = v - v_1$ is smooth as well). $\square$

We will use the notation $\langle v, w \rangle := g'(v, w)$ and $|v| := g'(v, v)^{1/2}$ for any $v, w \in TM$, extending the previous one.

PROPOSITION 2.14. If $\delta \in AC([0, \tau], M)$, then we can write $\delta = \gamma \circ h$ for some $\gamma \in AC([0, T], M)$ with $|\dot{\gamma}| = 1$ a.e. Here $T := L(\delta)$, $h \in AC([0, \tau])$ is increasing and $h(0) = 0$, $h(\tau) = T$. Moreover, such γ is unique and δ is horizontal iff γ is horizontal.

Proof. For the existence part we can localize and assume $M = \mathbb{R}^n$. In order to avoid ambiguity, we will denote by $|\cdot|_{g'}$ the norm associated to g' and with $|\cdot|_e$ the usual Euclidean one. We remark that $|\cdot|_{g'}$ is not a norm on $\mathbb{R}^n$: $|v|_{g'}$ makes sense only when $v \in T\mathbb{R}^n$, whereas the Euclidean norm is defined as usual on $\mathbb{R}^n$ and also on $T\mathbb{R}^n$ by means of the canonical identification $T_x\mathbb{R}^n \simeq \mathbb{R}^n$, for any $x \in \mathbb{R}^n$.

Define

$$h(t) := \int_0^t \left|\dot{\delta}\right|_{g'}(t')\, dt$$

for $0 \le t \le \tau$. h is AC, it is increasing and satisfies $h(0) = 0$, $h(\tau) = T$. Define also the pseudo-inverse

$$k(s) := \min\{t : h(t) = s\}$$

for $0 \le s \le T$ and set $\gamma(s) := \delta \circ k(s)$. Clearly $\delta = \gamma \circ h$. Moreover, for any $0 \le s \le s' \le T$,

$$\left|\gamma(s') - \gamma(s)\right|_e = \left|\delta(k(s')) - \delta(k(s))\right|_e \le \int_{k(s)}^{k(s')} \left|\dot{\delta}\right|_e (t)\, dt \le C \int_{k(s)}^{k(s')} \left|\dot{\delta}\right|_{g'} (t)\, dt = C(s' - s),$$

since there exists some $C > 0$ such that, for any x in the compact set $\gamma([0,T])$ and any $v \in T_x\mathbb{R}^n$, it holds $|v|_e \leq C\,|v|_{g'}$. Thus $\gamma \in W^{1,\infty}([0,T],\mathbb{R}^n)$ and in particular it is AC. So the chain rule (Lemma A.8) applies:

$$\dot{\delta} = (\dot{\gamma} \circ h)\dot{h} \text{ a.e.}$$

We use this convention: $\dot{\gamma}$, $\dot{\delta}$ and $\dot{h}$ denote classical derivatives and, whenever $\dot{h}$ vanishes, the right-hand side of the above formula is meant to vanish as well (even when $\dot{\gamma} \circ h$ is undefined).

Let us call $B \subseteq [0,\tau]$ the Borel set of all t where $\dot{\delta}$, $\dot{\gamma} \circ h(t)$ and $\dot{h}(t)$ are all defined, $\dot{h}(t) = \left|\dot{\delta}\right|_{g'}(t)$, $\dot{h}(t) \neq 0$ and the above formula holds. We know that $\mathcal{L}^1(h([0,\tau] \setminus B)) = 0$ (see Lemmas A.5 and A.6). For any $t \in B$ we have

$$\left|\dot{\delta}\right|_{g'}(t) = |\dot{\gamma}|_{g'}(h(t))\dot{h}(t) = |\dot{\gamma}|_{g'}(h(t))\left|\dot{\delta}\right|_{g'}(t),$$

so that $|\dot{\gamma}|_{g'}(h(t)) = 1$. Since $[0,T] = h([0,\tau])$, we have $[0,T] \setminus h([0,\tau] \setminus B) \subseteq h(B)$, so $|\dot{\gamma}|_{g'}(s) = 1$ on a subset of $[0,T]$ with full measure.

Assume now that γ is horizontal: from the chain rule and the fact that

$$\mathcal{L}^1\left(\left\{t : \dot{\gamma}(h(t)) \notin \mathcal{D}_{h(t)}\right\} \setminus \left\{t : \dot{h}(t) = 0\right\}\right) = 0$$

(see Lemma A.6) we deduce that δ is horizontal as well. Conversely, assume that δ is horizontal: then, letting $B' := \left\{t \in B : \dot{\delta}(t) \in \mathcal{D}_{\delta(t)}\right\}$, for any $t \in B'$ we have

$$\dot{\gamma}(h(t)) = \left(\dot{h}(t)\right)^{-1}\dot{\delta}(t) \in \mathcal{D}_{\delta(t)} = \mathcal{D}_{\gamma(h(t))},$$

so again $\dot{\gamma}(s) \in \mathcal{D}_{\gamma(s)}$ on the subset $h(B') \subseteq [0,T]$, which has full measure.

To prove uniqueness, assume $\delta = \gamma' \circ h'$ for some $\gamma \in AC([0,T'],M)$ with $|\dot{\gamma}'|_{g'} = 1$ a.e. and for some increasing $h' \in AC$ mapping $[0,\tau]$ to $[0,T']$. For any $t \in [0,\tau]$ we have

$$h'(t) = \int_0^t \dot{h}'(t')\,dt' = \int_0^t |\dot{\gamma}'|_{g'}(h'(t'))\dot{h}'(t')\,dt' = \int_0^t \left|\dot{\delta}\right|_{g'}(t')\,dt' = h(t),$$

by the chain rule again. So $T = T'$, $h \equiv h'$ and, from $\gamma \circ h = \gamma' \circ h$ and the surjectivity of h, we deduce $\gamma \equiv \gamma'$. □

REMARK 2.15. It follows that d_{CC} can be defined equivalently as $\inf L(\gamma)$, letting γ vary in $AC([0,1],M)$, in $W^{1,\infty}([0,1],M)$ or even among constant-speed curves, keeping of course the requirement that γ is horizontal and $\gamma(0) = x$, $\gamma(1) = y$. In addition, a very useful characterization of d_{CC} is

$$d_{CC}(x,y) = \inf_{\gamma \in \Omega_{x,y}} (2E(\gamma))^{1/2},$$

$E(\gamma)$ denoting the energy of γ. This fact comes simply from the Cauchy-Schwarz inequality

$$L(\gamma) = \int_0^1 |\dot{\gamma}|(t)\,dt \leq \left(\int_0^1 |\dot{\gamma}|^2(t)\,dt\right)^{1/2} = (2E(\gamma))^{1/2}$$

and the fact that equality holds iff γ has constant speed. Again this is true also replacing H^1 curves with one of the above classes.

We now prove that (M, d_{CC}) is an extended metric space, which means that d_{CC} satisfies all the properties of a distance except that a priori it could take on the value $+\infty$ (which will be ruled out by Theorem 2.34).

PROPOSITION 2.16. For any $x, y, z \in M$ we have $d_{CC}(x,y) = d_{CC}(y,x)$ and $d_{CC}(x,z) \leq d_{CC}(x,y) + d_{CC}(y,z)$. Moreover, $d(x,y) = 0$ iff $x = y$.

Proof. The first property follows from the fact that, if $\gamma : [0,1] \to M$ is a horizontal path from x to y, then $\gamma(1-\cdot)$ is a horizontal path from y to x with the same length. To prove the second property it suffices to notice that, given two horizontal paths $\gamma, \delta : [0,1] \to M$, γ connecting x to y and δ connecting y to z, then their concatenation

$$\eta(t) := \begin{cases} \gamma(2t) & \text{if } t \in \left[0, \frac{1}{2}\right] \\ \delta(2t-1) & \text{if } t \in \left[\frac{1}{2}, 1\right] \end{cases}$$

is a horizontal path from x to z and $L(\eta) = L(\gamma) + L(\delta)$.

It is evident that $d(x,x) = 0$. Now assume that $d(x,y) = 0$ but $x \neq y$. The auxiliary Riemannian metric g' that we built at the beginning of this section defines a distance $d_{g'}$ exactly as g defined d_{CC} (we simply replace g by g' and $\mathcal{D}$ by TM, so that when minimizing $L_{g'}(\gamma)$ all the curves from x to y are admissible). Since clearly $d_{CC}(x,y) \geq d_{g'}(x,y)$, to reach a contradiction it suffices to show that $d_{g'}(x,y) > 0$. For the sake of completeness, we include a proof of this fact (i.e. that the distance induced by a Riemannian metric separates points), although it is well-known and is quite unrelated to the topics treated in this thesis.

Take a smooth chart $\phi : U \to \mathbb{R}^n$, U being an open neighbourhood of y, such that $\phi(y) = 0$. We fix a positive R such that $x \notin \phi^{-1}(\overline{B}_R(0)) =: K$. We also define $r : \mathbb{R}^n \to \mathbb{R}$ by $r(z) := |z|$. Since $\overline{B}_R(0)$ is compact, there exists some $c > 0$ such that

$$\left(\phi^{-1}\right)^* g'(v,v) \geq c^2 \left|v\right|_e^2$$

for all $v \in \bigcup_{x \in \overline{B}_R} T_x\mathbb{R}^n$, where $\left(\phi^{-1}\right)^* g'$ is the pullback of g' by $\phi^{-1} : \mathbb{R}^n \to U$ and $|v|_e$ is the standard euclidean norm. Consider now a curve $\gamma \in H^1([0,1], M)$ from x to y and call

$$s' := \min\left\{t : \gamma(t) \in \phi^{-1}\left(\overline{B}_{R/2}\right)\right\}, \;\; s := \sup\left\{t < s' : \gamma(t) \notin K\right\}.$$

Since $\left\{t : \gamma(t) \in \phi^{-1}(B_{R/2})\right\}$ is open, we deduce $\phi \circ \gamma(s') \in \partial B_{R/2}$. It is clear that $\gamma(s) \in K$ and $\phi \circ \gamma(s) \in \partial B_R$. Now we have

$$\begin{aligned} \frac{R}{2} &= \int_s^{s'} \frac{d}{dt}\left(r \circ \phi \circ \gamma\right)(t)\, dt \leq \int_s^{s'} \left|dr\left(d\phi(\dot\gamma(t))\right)\right| dt \leq \int_s^{s'} \left|d\phi(\dot\gamma(t))\right|_e dt \\ &\leq c^{-1} \int_s^{s'} \left(\left(\phi^{-1}\right)^* g'\left(d\phi(\dot\gamma(t)), d\phi(\dot\gamma(t))\right)\right)^{\frac{1}{2}} dt = c^{-1} \int_s^{s'} \left(g'(\dot\gamma(t), \dot\gamma(t))\right)^{\frac{1}{2}} \leq c^{-1} L(\gamma), \end{aligned}$$

where we used the fact that r is 1-Lipschitz. So $L(\gamma) \geq \frac{cR}{2}$, which is a positive constant independent of γ, proving that $d_{g'}(x,y) > 0$. ☐

REMARK 2.17. The same proof shows that the topology induced by d_{CC} is finer than the original one: indeed, we proved that, if $U \subseteq M$ is open (in the original topology) and $y \in U$, then any point $x \in M \setminus U$ must lie outside some ball $\mathbb{B}_\epsilon(y)$, for some sufficiently small ϵ independent of x, proving that $\mathbb{B}_\epsilon(y) \subseteq U$.

In fact, the two topologies coincide: this will be a byproduct of the proof of the Chow-Rashevsky theorem given below. In the remaining part of this section, we will take this fact for granted, as well as the finiteness of d_{CC}, in order to simplify some proofs (this is legitimate since the results that we are going to prove here will not be used in the rest of the thesis).

On any metric space one has an intrinsic notion of length of a continuous curve, whose definition is built using solely the given distance. Taking into account how d_{CC} was defined on M, it is natural to expect a relation between the length of a horizontal path and its intrinsic length. Corollary 2.24 will address this point.

DEFINITION 2.18. Given a continuous curve $\gamma : [0,T] \to M$, let us define the *intrinsic length* of γ as

$$L_i(\gamma) := \sup \sum_{i=0}^{N-1} d\left(\gamma(t_i), \gamma(t_{i+1})\right),$$

where the supremum is taken over all choices of times $0 = t_0 \le \cdots \le t_N = T$ (letting N vary as well). The *intrinsic distance* between two points x and y is $d_i(x,y) := \inf L_i(\gamma)$, as γ varies among all continuous curves connecting x to y.

REMARK 2.19. Of course, nothing changes if we require that the sequence of times is strictly increasing in the definition of L_i. Notice that L_i is additive, i.e. $L_i(\gamma) = L_i\left(\gamma|_{[0,t]}\right) + L_i\left(\gamma|_{[t,T]}\right)$ for any $t \in (0,T)$. We also remark that, for any $x, y \in M$, we have $d_i(x,y) \ge d(x,y)$ (this happens in every metric space): for any continuous curve $\gamma : [0,T] \to M$ connecting x to y the choice $N := 1$, $t_0 := 0$, $t_1 := T$ gives $L_i(\gamma) \ge d(x,y)$. The next proposition shows that, in our special case, the converse inequality holds as well.

REMARK 2.20. L_i is invariant under continuous reparametrization, i.e. if $\gamma = \delta \circ h$, with h increasing and continuous, then $L_i(\gamma) = L_i(\delta)$: the inequality $L_i(\gamma) \le L_i(\delta)$ is trivial, while the converse one follows from the surjectivity of h.

PROPOSITION 2.21. (M,d) is a *length metric space*, i.e. for any $x, y \in M$ we have $d(x,y) = d_i(x,y)$.

Proof. By Remark 2.19, it suffices to show that $d_i(x,y) \le d(x,y)$. Take any horizontal path $\gamma \in \Omega_{x,y}$. For any choice $0 = t_0 \le \cdots \le t_N = 1$ we have

$$d\left(\gamma(t_i), \gamma(t_{i+1})\right) \le L\left(\gamma|_{[t_i,t_{i+1}]}\right),$$

since $\gamma|_{[t_i,t_{i+1}]}$ is a horizontal path connecting $\gamma(t_i)$ to $\gamma(t_{i+1})$. So

$$\sum_{i=0}^{N-1} d\left(\gamma(t_i), \gamma(t_{i+1})\right) \le \sum_{i=0}^{N-1} L\left(\gamma|_{[t_i,t_{i+1}]}\right) = L(\gamma),$$

which gives $L_i(\gamma) \le L(\gamma)$. Thus $d_i(x,y) \le L(\gamma)$ and the thesis follows by taking the infimum over γ. □

Before proceeding further, let us make a simple but useful remark.

REMARK 2.22 (Localization). Let $U \subseteq M$ be any open subset. If one replaces M with U, the set of horizontal paths between two given points $x, y \in U$ can become smaller (since we are excluding those which intersect $M \setminus U$). So we have the inequality $d(x,y) \le d^U(x,y)$, where d^U denotes the Carnot-Carathéodory distance on the manifold U.

However, whenever $\mathbb{B}_r(x) \subseteq U$, we have $d(x,y) = d^U(x,y)$ for any $y \in \mathbb{B}_r(x)$, so in fact $\mathbb{B}_r(x) = \mathbb{B}_r^U(x)$ ($\mathbb{B}_r^U(x)$ denoting the ball with respect to d^U): to see this, fix any $y \in \mathbb{B}_r(x)$ and $0 < \epsilon < r - d(x,y)$. There exists a horizontal path from x to y (in M) with length less than $d(x,y) + \epsilon < r$, so any point on this path belongs to $\mathbb{B}_r(x) \subseteq U$. So this path is contained in U and we deduce $d^U(x,y) \le d(x,y) + \epsilon$. Now it suffices to let $\epsilon \to 0$ to obtain $d^U(x,y) \le d(x,y)$.

We also remark that, if $\mathbb{B}_r(x) \subseteq U$, then $d(y,z) = d^U(y,z)$ for any $y, z \in \mathbb{B}_{r/3}(x)$: indeed, $d(y,z) < \frac{2}{3}r$ and any horizontal path connecting y to z with length less than $\frac{2}{3}r$ is contained in $\mathbb{B}_r(x) \subseteq U$. We conclude that $d^U(y,z) \leq d(y,z)$ in the same way as before.

THEOREM 2.23. If $\gamma : [0,T] \to M$ is continuous and $L_i(\gamma) < +\infty$, then γ is a continuous reparametrization of a horizontal unit-speed $\widehat{\gamma}$, with $L(\widehat{\gamma}) = L_i(\gamma)$.

Proof. Step 1. To begin with, we will replace γ by its arc length reparametrization $\widehat{\gamma}$, with respect to L_i. To this end, we define $h(t) := L_i\left(\gamma|_{[0,t]}\right)$ (for $t \in [0,T]$).

Let us prove that h is continuous: fix any $0 \leq t < T$ and any $\epsilon > 0$. We can find $t = t_0 < \cdots < t_N = T$ such that $\sum_{i=0}^{N-1} d(\gamma(t_i), \gamma(t_{i+1})) \geq L_i\left(\gamma|_{[t,T]}\right) - \epsilon$. So, if $t' \in (t, t_1)$ satisfies $d(\gamma(t'), \gamma(t_1)) \geq d(\gamma(t), \gamma(t_1)) - \epsilon$, we deduce

$$\begin{aligned} L_i\left(\gamma|_{[t',T]}\right) &\geq d(\gamma(t'), \gamma(t_1)) + \sum_{i=1}^{N-1} d(\gamma(t_i), \gamma(t_{i+1})) \\ &\geq \sum_{i=0}^{N-1} d(\gamma(t_i), \gamma(t_{i+1})) - \epsilon \geq L_i\left(\gamma|_{[t,T]}\right) - 2\epsilon \end{aligned}$$

and finally

$$h(t') = h(t) + L_i\left(\gamma|_{[t,t']}\right) = h(t) + L_i\left(\gamma|_{[t,T]}\right) - L_i\left(\gamma|_{[t',T]}\right) \leq h(t) + 2\epsilon,$$

so h is right-continuous. An analogous argument shows left continuity.

Now we define $k(s) := \min\{t : h(t) = s\}$ (for $s \in [0, L_i(\gamma)]$). Then we set $\widehat{\gamma} := \gamma \circ k$, so that $\gamma = \widehat{\gamma} \circ h$ (since γ has to be constant on $[k \circ h(t), t]$). We have

$$d(\widehat{\gamma}(s), \widehat{\gamma}(t)) = d(\gamma(k(s)), \gamma(k(t))) \leq L_i\left(\gamma|_{[k(s),k(t)]}\right) = t - s,$$

proving that $\widehat{\gamma}$ is 1-Lipschitz with respect to d. In particular, $\widehat{\gamma}$ is a continuous curve from x to y.

Step 2. We have to show that $\widehat{\gamma}$ is a unit-speed horizontal path. It suffices to obtain that $\widehat{\gamma}$ is horizontal and $\left|\dot{\widehat{\gamma}}\right| \leq 1$ a.e., because then the proof of Proposition 2.21 shows that $L(\widehat{\gamma}) \geq L_i(\widehat{\gamma}) = L_i(\gamma)$, so we must have $\left|\dot{\widehat{\gamma}}\right|_{g'} = 1$ a.e., as well.

Let us localize the problem. As $\gamma([0,T]) = \widehat{\gamma}([0, L_i(\gamma)])$ is compact, we can cover it with finitely many open sets U_j diffeomorphic to $\mathbb{R}^n$. We can also find some $r > 0$ such that $\mathbb{B}_r(\widehat{\gamma}(t))$ is compactly contained in some U_j (depending on t) for all $t \in [0, L_i(\gamma)]$: here we use the equality between the two topologies. Finally, since $\widehat{\gamma}$ is 1-Lipschitz, we can find a subdivision $0 = s_0 < \cdots < s_k = L_i(\gamma)$ such that

$$\widehat{\gamma}([s_{i-1}, s_i]) \subseteq \mathbb{B}_{r/3}(\widehat{\gamma}(s_{i-1}))$$

for all $1 \leq i \leq k$. By Remark 2.22, we have $d(x,y) = d^{U_j}(x,y)$ whenever $x, y \in \mathbb{B}_{r/3}(\widehat{\gamma}(t_{i-1}))$ and $\mathbb{B}_r(\widehat{\gamma}(t_{i-1})) \subseteq U_j$, so for any $t_{i-1} \leq s < t \leq t_i$ the intrinsic length of $\widehat{\gamma}|_{[s,t]}$ with respect to d^{U_j} is still $t - s$. Restricting our attention to $\gamma|_{[t_{i-1},t_i]}$, we reduce to the case $M = \mathbb{R}^n$. We still denote by $\widehat{\gamma}$ the new curve and (up to translating the time) we call $[0, \delta]$ its parametrization domain.

Step 3. For any $j \geq 0$ we define the set of times $\pi_j := \left\{\frac{i}{2^j}\delta \mid 0 \leq i \leq 2^j\right\}$ and we denote $t_i^j := \frac{i}{2^j}\delta$. Since $d\left(\widehat{\gamma}(t_i^j), \widehat{\gamma}(t_{i+1}^j)\right) = t_{i+1}^j - t_i^j$, we can find a constant-speed horizontal path $\gamma_i^j : [t_i^j, t_{i+1}^j] \to M$ joining $\widehat{\gamma}(t_i^j)$ to $\widehat{\gamma}(t_{i+1}^j)$, with $L(\gamma_i^j) \le \frac{\delta}{2^j}\left(1+\frac{1}{j}\right)$. We can also assume $\gamma_i^j\left([t_i^j, t_{i+1}^j]\right) \subseteq \mathbb{B}_r(\widehat{\gamma}(0)) \Subset \mathbb{R}^n$ (this ball is to be meant with respect to the Carnot-Carathéodory distance). Let us call $\gamma_j : [0,\delta] \to \mathbb{R}^n$ the curve obtained by concatenating $\gamma_0^j, \dots, \gamma_{2^j-1}^j$.

We remark that $\left|\dot\gamma_j\right|_{g'} \le 1 + \frac{1}{j}$ a.e., so that (since γ_j takes values in a compact set independent of j) $\left|\dot\gamma_j\right|_e \le C$ a.e., for some C independent of j. For any time having the form $t = \frac{i}{2^{j_0}}\delta$ we have $\gamma_j(t) = \widehat{\gamma}(t)$ whenever $j \ge j_0$. Thus, (γ_j) being an equicontinuous sequence of curves converging on a dense subset to $\widehat{\gamma}$, we deduce $\gamma_j \to \widehat{\gamma}$ uniformly on $[0,\delta]$. In particular, $\widehat{\gamma}$ is Lipschitz (as a map with values in $\mathbb{R}^n$) and $\dot\gamma_j \xrightharpoonup{*} \dot{\widehat{\gamma}}$ (in the usual duality with $L^1([0,\delta],\mathbb{R}^n)$).

Now we show that $\int_a^b \left|\dot{\widehat{\gamma}}\right|_{g'}^2 dt \le b-a$ for any $0 \le a < b \le \delta$. This is an easy consequence of weak convergence: indeed, identifying $T_x\mathbb{R}^n \simeq \mathbb{R}^n$ for every x (so that $\dot\gamma_j(t)$ will be considered as a vector in $\mathbb{R}^n$, rather than an element of $T_{\gamma_j(t)}\mathbb{R}^n$) and denoting by $g'(x) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ the auxiliary metric g' at x,

$$
\begin{aligned}
b-a &= \lim_{j\to\infty} \int_a^b g'(\gamma_j(t))\,(\dot\gamma_j(t), \dot\gamma_j(t))\, dt = \lim_{j\to\infty}\int_a^b g'(\widehat{\gamma}(t))\,(\dot\gamma_j(t),\dot\gamma_j(t))\, dt \\
&\ge \int_a^b \left|\dot{\widehat{\gamma}}\right|_{g'}^2 dt + 2\liminf_{j\to\infty}\int_a^b g'(\widehat{\gamma}(t))\left(\dot{\widehat{\gamma}}(t), \dot\gamma_j(t) - \dot{\widehat{\gamma}}(t)\right) dt = \int_a^b \left|\dot{\widehat{\gamma}}\right|_{g'}^2 dt.
\end{aligned}
$$

The second equality follows from the continuity of g', the uniform convergence $\gamma_j \to \widehat{\gamma}$ and the estimate $\left|\dot\gamma_j\right|_e \le C$, while the inequality is a consequence of

$$
\begin{aligned}
g'(\widehat{\gamma}(t))\,(\dot\gamma_j(t),\dot\gamma_j(t)) =& g'(\widehat{\gamma}(t))\left(\dot{\widehat{\gamma}}(t), \dot{\widehat{\gamma}}(t)\right) + 2g'(\widehat{\gamma}(t))\left(\dot{\widehat{\gamma}}(t), \dot\gamma_j(t) - \dot{\widehat{\gamma}}(t)\right) \\
&+ g'(\widehat{\gamma}(t))\left(\dot\gamma_j(t) - \dot{\widehat{\gamma}}(t), \dot\gamma_j(t) - \dot{\widehat{\gamma}}(t)\right).
\end{aligned}
$$

Finally, the $\liminf$ vanishes thanks to the weak convergence. Since a and b are arbitrary, we obtain $\left|\dot{\widehat{\gamma}}\right|_{g'} \le 1$ a.e.

Step 4. The fact that $\widehat{\gamma}$ is horizontal can be seen in two ways. For instance, one can find locally $n-r$ differential forms $\omega_{r+1}, \dots, \omega_n$ such that $v \in T_x\mathbb{R}^n$ lies in $\mathcal{D}$ iff $\omega_i(v) = 0$ for every $r+1 \le i \le n$ (locally there exist n linearly independent vector fields $X_1, \dots, X_n$ such that $\mathcal{D} = \langle X_1, \dots, X_r\rangle$ and it suffices to take the dual basis $\omega_1, \dots, \omega_n$ pointwise). Now, for any fixed $i = r+1, \dots, n$, we have $\int_a^b \omega_i(\gamma_j(t))\,(\dot\gamma_j(t))\, dt = 0$ and, by weak convergence again, $\int_a^b \omega_i(\widehat{\gamma}(t))\left(\dot{\widehat{\gamma}}(t)\right) dt = 0$ (for any sufficiently small interval $[a,b]$), so that $\omega_i(\widehat{\gamma}(t))\left(\dot{\widehat{\gamma}}(t)\right) = 0$ a.e.

Alternatively, recalling the proof of Lemma 2.13, we can define the *penalized metrics* $g'_m := g \oplus m\,(h|_{\mathcal{D}'})$, for any $m \ge 1$. Replacing g' with g'_m we obtain

$$
\int_0^\delta \left|\dot{\widehat{\gamma}}\right|_{g'_m}^2 dt \le \delta.
$$

But, for any $v \in T_x\mathbb{R}^n$, we have

$$
g'_m(v,v) \uparrow g'_\infty(v,v) := \begin{cases} g(v,v) & \text{if } v \in \mathcal{D} \\ +\infty & \text{otherwise} \end{cases}
$$

(in fact, g'_∞ should be regarded as the natural extension of g to the whole tangent space). By monotone convergence we deduce

$$\int_0^\delta \left|\dot{\widehat{\gamma}}\right|^2_{g'_\infty} dt \le \delta < +\infty,$$

so we cannot have $\dot{\widehat{\gamma}} \notin \mathcal{D}$ on a subset with positive measure. □

COROLLARY 2.24. Let $\gamma \in AC([0,T],M)$ and assume that $L_i(\gamma) < +\infty$. Then γ is horizontal and $L(\gamma) = L_i(\gamma)$.

Proof. By Theorem 2.23, $\gamma = \widehat{\gamma} \circ h$ for some horizontal unit-speed $\widehat{\gamma} : [0, L_i(\gamma)] \to M$ and some continuous increasing $h : [0,T] \to [0, L_i(\gamma)]$. Let us prove that γ is horizontal: we can assume that $M = \mathbb{R}^n$. By Theorem 2.23 we know that $\gamma = \widehat{\gamma} \circ h$ for some horizontal unit-speed $\widehat{\gamma}$. Let $w : [0, L_i(\gamma)] \to \mathbb{R}^n$ be any Borel function such that $c \le |w| \le C$ everywhere (for some suitable $c, C > 0$) and $w = \dot{\widehat{\gamma}}$ a.e. By Lemma A.4, denoting by μ the unique positive measure on $[0,T]$ such that $h(t) = \mu([0,t])$,

$$\gamma(t) = \gamma(0) + \int_0^{h(t)} w(s)\,ds = \gamma(0) + \int_0^t w \circ h\,d\mu$$

(to be precise, in the last equality one should integrate on $h^{-1}([0,h(t)])$ instead of $[0,t]$, but the difference between these two intervals is μ-negligible as $h \equiv h(t)$ there). We deduce that

$$(w \circ h)\mu = \dot{\gamma}\mathcal{L}^1.$$

Comparing the corresponding total variations and using $c \le |w \circ h| \le C$, it follows that $\mu \ll \mathcal{L}^1$. So h is AC and the horizontality of γ follows from Remark 2.12. Moreover, $L(\gamma) = L(\widehat{\gamma}) = L_i(\gamma)$. □

2.3. Admissible controls and the endpoint map

Throughout this section we will always assume that $\mathcal{D} = \langle X_1, \ldots, X_r \rangle$ for suitable globally defined vector fields. We now show a convenient way to parametrize horizontal paths in $H^1([0,1],M)$ starting at a given point x. Notice that we are fixing only the starting point; on the contrary, in general the set $\Omega_{x,y}$ does not have any reasonable structure, since the additional constraint that the final point is y can become singular (which is also the reason why the first order conditions derived in this chapter will not suffice to obtain the regularity of length minimizers).

DEFINITION 2.25. The Hilbert space $L^2([0,1],\mathbb{R}^r)$ is called the set of *controls*. A generic control will be usually denoted by u. Let $\mathcal{U}_x \subseteq L^2([0,1],\mathbb{R}^r)$ be the set of the controls u such that the Cauchy problem

$$\begin{cases} \dot{\gamma}(t) = \sum_{i=1}^r u_i(t) X_i(\gamma(t)) \\ \gamma(0) = x \end{cases}$$

has a (unique) solution $\gamma \in H^1([0,1],M)$ satisfying the above ordinary differential equation for a.e. t. $\mathcal{U}_x$ is called the set of *admissible controls* and sometimes it will be simply denoted by $\mathcal{U}$. We will use the compact notation $\langle u(t), X \rangle (x') := \sum_{i=1}^r u_i(t) X_i(x')$ (for any $x' \in M$).

In Appendix B we show that this Cauchy problem has at most one solution and that $\mathcal{U}$ is an open set.

Remark 2.26. Given a horizontal curve $\delta \in H^1([0,1],M)$ with $\delta(0)=x$, we can write $\dot\delta(t)=\sum_{i=1}^r u_i(t)X_i(\delta(t))$ for some $u \in L^2([0,1],\mathbb{R}^r)$, which is uniquely determined a.e. Notice that δ is the solution of the Cauchy problem with the control u, so $u \in \mathcal{U}_x$. u will be called the control *associated to* δ. Conversely, given $u \in \mathcal{U}_x$, the solution γ will be called the *trajectory associated to* u. It is now clear that there is a correspondence between $\mathcal{U}_x$ and the set of horizontal curves (in $H^1([0,1],M)$) starting at x, given by the above associations, which are inverse to each other.

Definition 2.27. More generally, for any $T>0$, we denote by $\mathcal{U}_{x,T}$ (or simply by $\mathcal{U}_T$) the open subset of $L^2([0,T],\mathbb{R}^r)$ consisting of the controls u such that the above Cauchy problem has a solution in $H^1([0,T],M)$.

Definition 2.28. Given $u \in L^2([0,T],\mathbb{R}^r)$, for any $\lambda>0$ we define the *rescaled* control $u_\lambda \in L^2([0,\lambda T],\mathbb{R}^r)$ by

$$u_\lambda(t) := \lambda^{-1}u(\lambda^{-1}t).$$

The *reversed* control is $\check{u} \in L^2([0,T],\mathbb{R}^r)$, given by

$$\check{u}(t) := -u(T-t).$$

Given $u' \in L^2([0,T'],\mathbb{R}^r)$, we also define the *join* of u and u' to be the control $u*u' \in L^2([0,T+T'],\mathbb{R}^r)$ defined as

$$u*u'(t) := \begin{cases} u(t) & t\in[0,T] \\ u'(t-T) & t\in(T,T+T']. \end{cases}$$

Remark 2.29. Again, there is a correspondence between $\mathcal{U}_{x,T}$ and the set of horizontal curves defined on $[0,T]$ and starting at x. Notice that, if $\gamma \in H^1([0,T],M)$ is a horizontal curve with $\gamma(0)=x$, then $\gamma(T\cdot)$ is a horizontal curve defined on $[0,1]$. This gives a bijection between the horizontal curves on $[0,T]$ and the ones on $[0,1]$ (with starting point x); the corresponding bijection between $\mathcal{U}_{x,T}$ and $\mathcal{U}_x$ is given by $u \mapsto u_{T^{-1}}$.

Definition 2.30. We define the *endpoint map*

$$\mathrm{End}_T : \{(x,u) : u \in \mathcal{U}_{x,T}\} \subseteq M \times L^2([0,T],\mathbb{R}^r), \quad (x,u) \mapsto \gamma(T),$$

where $\gamma \in H^1([0,T],M)$ is the unique solution to the Cauchy problem. When $T=1$ we will use the simpler notation $\mathrm{End}(x,u)$. Sometimes, when the starting point x is fixed, End_T (or End) will just have $\mathcal{U}_T=\mathcal{U}_{x,T}$ (or $\mathcal{U}=\mathcal{U}_x$) as its domain.

In Appendix B the regularity properties of End_T are studied. In particular, Proposition B.10 and Corollary B.16 (applied to the rescaled controls, so as to reduce to the case $T=1$) tell us that $\{(x,u) : u \in \mathcal{U}_{x,T}\}$ is an open subset of $M \times L^2([0,T],\mathbb{R}^r)$ and that End_T is C^∞ on it.

We now list some basic properties of the endpoint map.

Proposition 2.31. Let $u \in \mathcal{U}_{x,T}$, $u' \in \mathcal{U}_{y,T'}$, where $y := \mathrm{End}_T(x,u)$, and $\lambda>0$. Then:

(1) $u_\lambda \in \mathcal{U}_{x,\lambda T}$, $\mathrm{End}_{\lambda T}(x,u_\lambda)=\mathrm{End}_T(x,u)$ and $\operatorname{im} d\,\mathrm{End}_{\lambda T}(x,\cdot)_{u_\lambda} = \operatorname{im} d\,\mathrm{End}_T(x,\cdot)_u$;

(2) $\check{u} \in \mathcal{U}_{y,T}$, $\mathrm{End}_T(y,\check{u})=x$ and $\operatorname{rk} d\,\mathrm{End}_T(x,\cdot)_u = \operatorname{rk} d\,\mathrm{End}_T(y,\cdot)_{\check{u}}$;

(3) $u*u' \in \mathcal{U}_{x,T+T'}$, $\mathrm{End}_{T+T'}(x,u*u')=\mathrm{End}_{T'}(y,u')$ and

$$\operatorname{rk} d\,\mathrm{End}_{T+T'}(x,\cdot)_{u*u'} \geq \max\left\{\operatorname{rk} d\,\mathrm{End}_T(x,\cdot)_u, \operatorname{rk} d\,\mathrm{End}_{T'}(y,\cdot)_{u'}\right\}.$$

Proof. Calling γ and δ the trajectories associated to u and u' (starting at x and y respectively), notice that $\gamma(\lambda^{-1}\cdot)$, $\gamma(T-\cdot)$ and the curve obtained by concatenating γ with δ are all horizontal curves in H^1. The associated controls are those written in the three items of the statement and the assertions concerning their admissibility and their endpoints are now clear. Moreover, from $\mathrm{End}_T(x,u) = \mathrm{End}_{\lambda T}(x,u_\lambda)$ we get

$$d\,\mathrm{End}_T(x,\cdot)_u[v] = d\,\mathrm{End}_{\lambda T}(x,\cdot)_{u_\lambda}[v_\lambda],$$

proving that $\operatorname{im} d\,\mathrm{End}_{\lambda T}(x,\cdot)_{u_\lambda} = \operatorname{im} d\,\mathrm{End}_T(x,\cdot)_u$ (since $v \mapsto v_\lambda$ is a bijection between $L^2([0,T],\mathbb{R}^r)$ and $L^2([0,\lambda T],\mathbb{R}^r)$). For the second item, differentiation of the identity

$$\mathrm{End}_T(\mathrm{End}_T(x,u),\breve{u}) = x \tag{2.1}$$

with respect to u gives, using the smoothness of End_T and the chain rule,

$$d\,\mathrm{End}_T(\cdot,\breve{u})_y[d\,\mathrm{End}_T(x,\cdot)_u[v]] + d\,\mathrm{End}_T(y,\cdot)_{\breve{u}}[\breve{v}] = 0.$$

This gives $\operatorname{rk} d\,\mathrm{End}_T(x,\cdot)_u = \operatorname{rk} d\,\mathrm{End}_T(y,\cdot)_{\breve{u}}$, since $d\,\mathrm{End}_T(\cdot,\breve{u})_y$ is invertible (which can be seen differentiating (2.1) with respect to x). Similarly,

$$\begin{aligned}
d\,\mathrm{End}_{T+T'}(x,\cdot)_{u*u'}[v*v'] &= \frac{d}{ds}\,\mathrm{End}_{T+T'}(x,(u+sv)*(u'+sv'))\Big|_{s=0} \\
&= \frac{d}{ds}\,\mathrm{End}_{T'}(\mathrm{End}_T(x,u+sv),u'+sv')\Big|_{s=0} \\
&= d\,\mathrm{End}_{T'}(\cdot,u')_y[d\,\mathrm{End}_T(x,\cdot)_u[v]] + d\,\mathrm{End}_{T'}(y,\cdot)_{u'}[v']
\end{aligned}$$

and the last assertion follows from the fact that $d\,\mathrm{End}_{T'}(\cdot,u')_y$ is invertible, which is seen by differentiating the following identity with respect to y:

$$\mathrm{End}_{T'}(\mathrm{End}_{T'}(y,u'),(u')\breve{}\,) = y.$$

□

Let us call Φ_t the flow associated to some fixed control $\overline{u} \in \mathcal{U}_{x,T}$, i.e. $\Phi_t(x') := \mathrm{End}_t(x',\overline{u})$ for $t \in [0,T]$. By the same argument used at the end of the preceding proof, we know that $\Phi_T|_V$ is a diffeomorphism onto its image for a suitable neighbourhood V of x (depending on T). It will be convenient to introduce a modified version of the endpoint map, namely

$$\widehat{\mathrm{End}}_T(u) := \Phi_T|_V{}^{-1} \circ \mathrm{End}_T(x,u).$$

Notice that $\widehat{\mathrm{End}}_T(\underline{u}) = x$ and that $\widehat{\mathrm{End}}_T$ is defined on some neighbourhood of $\overline{u}$.

LEMMA 2.32. The differential of $\widehat{\mathrm{End}}_T$ at $\overline{u}$ is given by

$$d(\widehat{\mathrm{End}}_T)_{\overline{u}}[v] = \int_0^T \Phi_t^* \left\langle v(t), X\right\rangle (x)\,dt.$$

Proof. It suffices to compute $d\,\widehat{\mathrm{End}}_{\overline{u}}[v]$. Assume for the moment that $\overline{u}, v \in C^\infty$. From Proposition C.11 we have (in local coordinates near x)

$$\widehat{\mathrm{End}}_T(\overline{u}+sv) = s\int_0^T \Phi_t^* \left\langle v(t), X\right\rangle \left(\widehat{\mathrm{End}}_t(\overline{u}+sv)\right)\,dt.$$

So, noticing that $\widehat{\mathrm{End}}_t(\overline{u}) = x$ for any t,

$$\frac{\partial}{\partial s}\,\widehat{\mathrm{End}}_T(\overline{u}+sv)\Big|_{s=0} = \int_0^T \Phi_t^* \left\langle v(t), X\right\rangle (x)\,dt.$$

Since $\widehat{\mathrm{End}}_T(\overline{u}+v)$ is smooth as $\overline{u}$ and v vary (notice that $\overline{u}$ appears also in the definition of $\widehat{\mathrm{End}}_T$, due to the presence of Φ_T) it follows that the same formula holds for any $\overline{u} \in \mathcal{U}$ and any $v \in L^2([0,1],\mathbb{R}^r)$, provided we show that the right-hand side is continuous in $(\overline{u}, v)$ (in the L^2 topology). The right-hand side equals

$$\sum_{i=1}^{r} \int_0^T v_i(t) d\left(\Phi_t\right)_x^{-1} \left[X_i\left(\Phi_t(x)\right)\right] dt$$

and it suffices to know that, whenever $u_n \to \overline{u}$ in L^2, we have $\Phi_t^n(x) \to \Phi_t(x)$ uniformly on $[0,T]$ and that the same convergence holds (in local coordinates) for the first spatial derivatives of Φ_t^n, where Φ_t^n denotes the flow associated to u_n. This follows from Proposition B.12 and the compactness of $[0,T] \times \{x\}$. ☐

Corollary 2.33. If $\overline{u} \in \mathcal{U}_{x,T}$, then $X_i(y) \in \operatorname{im} d\operatorname{End}_T(x,\cdot)_{\overline{u}}$ for any $i = 1, \dots, r$, where $y := \operatorname{End}_T(x, \overline{u})$.

Proof. For any $0 < \epsilon < T$ we define $v_{i,\epsilon} \in L^2([0,T],\mathbb{R}^r)$ by

$$v_{i,\epsilon}(t) := \begin{cases} 0 & t \in [0, T-\epsilon] \\ \frac{e_i}{\epsilon} & t \in (T-\epsilon, T]. \end{cases}$$

For fixed i, since $t \mapsto \Phi_t^* X_i(x)$ is continuous, we have

$$d(\widehat{\operatorname{End}}_T)_{\overline{u}}[v_{i,\epsilon}] = \int_0^T \Phi_t^* \left\langle v_{i,\epsilon}(t), X \right\rangle (x)\, dt = \frac{1}{\epsilon}\int_{T-\epsilon}^T (\Phi_t^* X_i)(x)\, dt \to (\Phi_T^* X_i)(x)$$

as $\epsilon \downarrow 0$. Since $\operatorname{im} d(\widehat{\operatorname{End}}_T)_{\overline{u}}$ is closed in T_xM (as it is a finite-dimensional subspace), we deduce $d(\Phi_T)_x^{-1}[X_i(y)] = (\Phi_T^* X_i)(x) \in \operatorname{im} d(\widehat{\operatorname{End}}_T)_{\overline{u}}$. This gives the thesis, since $d\operatorname{End}_T(x,\cdot)_{\overline{u}} = d\left(\Phi_T \circ \widehat{\operatorname{End}}_T\right)_{\overline{u}} = d(\Phi_T)_x \circ d(\widehat{\operatorname{End}}_T)_{\overline{u}}$. ☐

2.4. Chow-Rashevsky theorem

In this section we prove the fundamental fact, proved independently by Chow (1939) and Rashevsky (1938), that any two points in a sub-Riemannian manifold can be joined by a horizontal path.

Theorem 2.34. If M is connected (as a smooth manifold) and is endowed with a bracket-generating distribution $\mathcal{D}$, then for any couple of points $x, y \in M$ there exists a constant-speed curve $\gamma : [0,1] \to M$ satisfying $\gamma(0) = x$, $\gamma(1) = y$ and $\gamma'(t) \in \mathcal{D}$ for a.e. $t \in [0,1]$.

This result follows easily from the local openness of the endpoint map End, which we state as a lemma (to be proved later).

Lemma 2.35. Assume that $\mathcal{D} = \langle X_1, \dots, X_r \rangle$ and fix $\overline{x} \in M$. For any $\overline{u} \in \mathcal{U}$ ($\mathcal{U}$ being the set of admissible controls for the endpoint map $\operatorname{End}(\overline{x},\cdot)$), $\operatorname{End}(\overline{x},\cdot) : \mathcal{U} \to M$ is locally open at $\overline{u}$ (i.e., for any positive r such that $B_r(\overline{u}) \subseteq \mathcal{U}$, $\operatorname{End}(\overline{x}, B_r(\overline{u}))$ is a neighbourhood of $\operatorname{End}(\overline{x},\overline{u})$). As a consequence, $\operatorname{End}(\overline{x},\cdot)$ is an open map.

Before proving the lemma, let us see how Chow-Rashevsky theorem is deduced.

Proof of Theorem 2.34. In view of Proposition 2.14, it suffices to find a horizontal curve in H^1 joining x to y. We define a relation on points of M by saying that $a \sim b$ iff a and b can be joined with a horizontal curve in H^1. This is clearly an equivalence relation. Since M is connected, we are reduced to showing that the equivalence classes are open, as this implies that there is only an equivalence class. So, fixing any $\overline{x} \in M$, we have to prove that all the points in a neighbourhood of $\overline{x}$ can be joined to $\overline{x}$. By localizing we can assume $\mathcal{D} = \langle X_1, \ldots, X_r \rangle$. The thesis follows immediately from Lemma 2.35, choosing $\overline{u} := 0$ and noticing that $\mathrm{End}(\overline{x}, 0) = \overline{x}$. $\square$

Proof of Lemma 2.35. The proof is divided into three steps. In the sequel we will omit the dependence of End on the starting point $\overline{x}$.

Step 1. Let us recall that End is smooth (see Corollary B.16). The goal of this and the following step is to find an arbitrarily small $u \in \mathcal{U}$ such that $d\,\mathrm{End}_u$ is surjective. To this aim, we can localize and assume $M = \mathbb{R}^n$, as well. Let

$$k := \lim_{\epsilon \to 0} \max \left\{ \mathrm{rk}\, d\,\mathrm{End}_u \mid u \in \mathcal{U}, \|u\|_2 < \epsilon \right\}$$

(the requirement $u \in \mathcal{U}$ becomes redundant when ϵ is sufficiently small). We remark that the maximum in the above formula is decreasing as $\epsilon \to 0$, so the limit exists and we can find some $\overline{\epsilon}$ such that the maximum equals k for any $0 < \epsilon \leq \overline{\epsilon}$. Fix any $u \in \mathcal{U}$ with $\|u\|_2 < \epsilon \leq \overline{\epsilon}$ and $\mathrm{rk}\, d\,\mathrm{End}_u = k$. We can find $v_1, \ldots, v_k \in L^2([0,1])$ such that $d\,\mathrm{End}_u[v_1], \ldots, d\,\mathrm{End}_u[v_k]$ form a basis of $\mathrm{im}\, d\,\mathrm{End}_u$. So the map $F : B_r(0) \subseteq \mathbb{R}^k \to \mathbb{R}^n$ given by

$$F(t_1, \ldots, t_k) := \mathrm{End}\left(u + \sum_{i=1}^{k} t_i v_i \right)$$

is a smooth embedding if r is small enough. Let us call $N := F(B_r)$. Moreover, shrinking r if necessary, we can assume $\|u + \sum t_i v_i\|_2 < \epsilon$ for any $t \in B_r$. Hence, by the choice of ϵ, we have $\mathrm{rk}\, d\,\mathrm{End}_{u'} \leq k$, for any u' having the form $u' = u + \sum t_i u_i$ (with $t \in B_r$). But, as $\mathrm{im}\, d\,\mathrm{End}_{u'} \supseteq \mathrm{im}\, dF_t$, we have $\mathrm{rk}\, d\,\mathrm{End}_{u'} \geq \mathrm{rk}\, dF_t = k$, thus we obtain $\mathrm{rk}\, d\,\mathrm{End}_{u'} = \mathrm{rk}\, dF_t$ and $\mathrm{im}\, d\,\mathrm{End}_{u'} = \mathrm{im}\, dF_t$.

Step 2. From the last equality and Corollary 2.33 we get $X_i(p) \in T_p N$ for any $p \in N$ and any $i = 1, \ldots, r$. Before proceeding further, we need two simple facts about flows of smooth vector fields.

PROPOSITION 2.36. If X is a smooth vector field on $\mathbb{R}^n$ such that $X(p) \in T_p N$ for any $p \in N$, then $\Phi_s(p, X) \in N$ for any small enough s (depending on $p \in N$).

Proof. Since N is embedded, X restricts to a smooth vector field $X|_N$ on N. The thesis follows from the fact that any integral curve for $X|_N$ (in N) is also an integral curve for X in $\mathbb{R}^n$. $\square$

PROPOSITION 2.37. If X and Y are smooth vector fields on $\mathbb{R}^n$, then for any x_0 we have

$$[X, Y](x_0) = \lim_{t \to 0} \frac{\Phi_{-t}(Y) \circ \Phi_{-t}(X) \circ \Phi_t(Y) \circ \Phi_t(X)(x_0) - x_0}{t^2}.$$

Proof. Since $\frac{d}{dt}\Phi_t(X)(x) = X(\Phi_t(X)(x))$, we have

$$\begin{aligned} \Phi_t(X)(x) &= x + \int_0^t X(\Phi_\tau(X)(x))\, d\tau = x + tX(x) + \int_0^t \int_0^\tau dX[X](\Phi_s(X)(x))\, ds\, d\tau \\ &= x + tX(x) + \frac{t^2}{2} dX[X](x) + o(t^2), \end{aligned}$$

where the error is uniform as x varies on a compact neighbourhood of x_0 (we use the notation $dX[X](p)$ instead of $dX_p[X(p)]$ for brevity). Thus

$$\begin{aligned}\Phi_t(Y)\circ\Phi_t(X)(x_0) &= \Phi_t(X)(x_0) + tY\left(\Phi_t(X)(x_0)\right) + \frac{t^2}{2}dY[Y]\left(\Phi_t(X)(x_0)\right) + o(t^2)\\ &= x_0 + t(X+Y)(x_0) + \frac{t^2}{2}\left(dX[X] + 2dY[X] + dY[Y]\right)(x_0) + o(t^2),\end{aligned}$$

$$\begin{aligned}\Phi_{-t}(X)\circ\Phi_t(Y)\circ\Phi_t(X)(x_0) = x_0 + tY(x_0) + \frac{t^2}{2}(dX[X] + 2dY[X] + dY[Y] - 2dX[X+Y]\\ +dX[X])(x_0) + o(t^2) = x_0 + tY(x_0) + \frac{t^2}{2}\left(2dY[X] - 2dX[Y] + dY[Y]\right)(x_0) + o(t^2),\end{aligned}$$

$$\begin{aligned}\Phi_{-t}(Y)\circ\Phi_{-t}(X)\circ\Phi_t(Y)\circ\Phi_t(X)(x_0) = x_0 + \frac{t^2}{2}(2dY[X] - 2dX[Y] + dY[Y]\\ -2dY[Y] + dY[Y])(x_0) + o(t^2) = x_0 + t^2[X,Y](x_0) + o(t^2).\end{aligned}$$

□

From Proposition 2.36 we deduce that

$$\sigma(s) := \Phi_{-\sqrt{s}}(X_j)\circ\Phi_{-\sqrt{s}}(X_i)\circ\Phi_{\sqrt{s}}(X_j)\circ\Phi_{\sqrt{s}}(X_i)(p) \in N$$

if s is small enough, for any fixed $p \in N$. But by Proposition 2.37 $\sigma'(0) = [X_i, X_j](p)$, so $[X_i, X_j](p) \in T_pN$ (notice that, to reach this conclusion, we do not even need to check that σ' is continuous at 0). Iterating this argument with X_i and X_j replaced by $[X_i, X_j]$ and X_k and so on, by the bracket-generating condition we finally obtain $T_pN = \mathbb{R}^n$, so $k = n$.

Step 3. Let us go back to the original statement. We have obtained that $d\,\mathrm{End}_u$ is surjective for some arbitrarily small u. Now let us define

$$\overline{v} := u_\delta * (\check{u})_\delta * (\overline{u})_{1-2\delta},$$

(see Definition 2.28). We remark that $\mathrm{End}(\overline{v}) = \mathrm{End}(\overline{u})$. Moreover, by Proposition 2.31, $d(\mathrm{End}_\delta)_{u_\delta}$ is still surjective and $\mathrm{rk}\, d\,\mathrm{End}_{\overline{v}} \geq \mathrm{rk}\, d(\mathrm{End}_\delta)_{u_\delta} = n$, so $d\,\mathrm{End}_{\overline{v}}$ is surjective as well. Thus End is locally open at $\overline{v}$, i.e. the image of any neighbourhood of $\overline{v}$ is a neighbourhood of $\mathrm{End}(\overline{u})$. In order to conclude the proof of the lemma, we just have to show that $\overline{v}$ can be made arbitrarily close to $\overline{u}$. In fact, we have

$$\begin{aligned}\|\overline{v}-\overline{u}\|_2^2 = &\int_0^\delta \left|\frac{1}{\delta}u\left(\frac{t}{\delta}\right) - \overline{u}(t)\right|^2 dt + \int_\delta^{2\delta}\left|\frac{1}{\delta}u\left(2-\frac{t}{\delta}\right) - \overline{u}(t)\right|^2 dt\\ &+ \int_{2\delta}^1 \left|\frac{1}{1-2\delta}\overline{u}\left(\frac{t-2\delta}{1-2\delta}\right) - \overline{u}(t)\right|^2 dt\\ \leq &\frac{4}{\delta}\|u\|_2^2 + 2\int_0^{2\delta}|\overline{u}|^2(t)\,dt + \int_{2\delta}^1\left|\frac{1}{1-2\delta}\overline{u}\left(\frac{t-2\delta}{1-2\delta}\right) - \overline{u}(t)\right|^2 dt.\end{aligned}$$

We can choose δ such that the last two terms are arbitrarily small (the fact that the last one is infinitesimal is clear if $\overline{u}$ is replaced by any function in $C_c\left((0,1),\mathbb{R}^r\right)$ and it can be deduced for $\overline{u}$, as well, by a standard approximation argument). Now that δ is fixed, we recall that $\|u\|_2 < \epsilon$, so we are done upon selecting some ϵ such that $\frac{4}{\delta}\epsilon^2$ is small as we want. □

Corollary 2.38. The topology induced by d_{CC} on M coincides with the original one.

Proof. Call τ the original topology and τ' the one induced by the Carnot-Carathéodory distance. We have already proved that $\tau \subseteq \tau'$ (see Remark 2.17). We are left to show that, given any $x \in M$ and $\epsilon > 0$, we can find an open set $U \in \tau$ such that $x \in U$ and $U \subseteq \mathbb{B}_\epsilon(x)$. Shrinking ϵ if necessary, we can localize and assume $\mathcal{D} = \langle X_1, \ldots, X_r \rangle$. We can also assume that $X_1, \ldots, X_r$ are orthonormal. By Lemma 2.35 the set $U := \mathrm{End}(x, B_\epsilon(0) \cap \mathcal{U})$ belongs to τ. To conclude, notice that $U \subseteq \mathbb{B}_\epsilon(x)$: indeed, if $u \in B_\epsilon(0) \cap \mathcal{U}$, calling $\gamma : [0,1] \to \mathbb{R}^n$ the curve associated to this control, we have

$$L(\gamma) = \int_0^1 |u|(t)\, dt \leq \|u\|_2 < \epsilon,$$

so that $\mathrm{End}(x, u) = \gamma(1)$ belongs to $\mathbb{B}_\epsilon(x)$. $\square$

2.5. Ball-box comparison

The fact that the topology induced by d_{CC} coincides with the original one can also be deduced from a classical result proved by Nagel, Stein and Wainger in [NSW85], which we now state for the sake of completeness (although it will not be used in this thesis).

Assume $M = \mathbb{R}^n$ and $\mathcal{D} = \langle X_1, \ldots, X_r \rangle$ for simplicity. For any $x \in \mathbb{R}^n$ let

$$s(x) := \min \left\{ k : \mathrm{Lie}^k(\mathcal{D}, x) = T_x\mathbb{R}^n \right\}.$$

By the bracket-generating condition, $s(x)$ is finite for any x. Now fix any compact subset $K \subseteq \mathbb{R}^n$. Since, for any x, all points y close to x still satisfy $\mathrm{Lie}^{s(x)}(\mathcal{D}, y) = T_y\mathbb{R}^n$, we have $s(y) \leq s(x)$ near x, i.e. s is upper semicontinuous. Hence, $s(\cdot)$ has a maximum $\overline{s}$ on K.

Let $Y_1, \ldots, Y_q$ be an enumeration of all possible iterated commutators of the form

$$[[\cdots [X_{j_1}, X_{j_2}], \cdots], X_{j_m}],$$

with $j_1, \ldots, j_m = 1, \ldots, r$ and $1 \leq m \leq \overline{s}$. We call $d(i)$ the length of the iterated commutator which gives Y_i (i.e. the integer m). For any $x \in \mathbb{R}^n$ and any $I = (i_1, \ldots, i_n) \in \{1, \ldots, q\}^n$, we let

$$d(I) := d(i_1) + \cdots + d(i_n), \quad \lambda_I(x) := \det\left(Y_{i_1}(x), \ldots, Y_{i_n}(x)\right),$$

where $Y_{i_1}(x), \ldots, Y_{i_n}(x)$ are identified with vectors in $\mathbb{R}^n$. We also define, for any $x, h \in \mathbb{R}^n$,

$$\exp_x^I(h) := \Phi_1(x, h_1 Y_{i_1} + \cdots + h_n Y_{i_n})$$

(see Definition C.2 for the notation) and

$$\mathrm{Box}_r^I(x) := \left\{ \exp_x^I(h) \mid h \in \mathbb{R}^n, \max_{1 \leq k \leq n} |h_k|^{1/d(i_k)} < r \right\},$$

which is well-defined and is also an open neighbourhood of x if $\lambda_I(x) \neq 0$ and h is sufficiently small.

THEOREM 2.39 (Ball-box comparison). There exist $\overline{r} > 0$ and $C > 1$ (both depending on K) such that, for any $x \in K$ and any $0 < r < \overline{r}$, we have

$$\mathrm{Box}_{r/C}^I(x) \subseteq \mathbb{B}_r(x) \subseteq \mathrm{Box}_{Cr}^I(x)$$

whenever $|\lambda_I(x)|\, r^{d(I)} > \frac{1}{2} \max_J |\lambda_J(x)|\, r^{d(J)}$. Moreover, there exists some $C' > 0$ such that $d_{CC}(x, y) \leq C' |x - y|^{1/\overline{s}}$ for any $x, y \in K$.

For the proof, we refer the reader to the original paper [NSW85].

Fixing any $x \in \mathbb{R}^n$ and using this theorem with $K := \{x\}$, from the double inclusion we deduce that the two topologies on $M = \mathbb{R}^n$ coincide. Notice that, as in the previous section, in order to obtain this corollary the assumptions on M can be easily removed: using Remark 2.17, in the general case it suffices to show that, given $x \in M$ and $r > 0$, $\mathbb{B}_r(x)$ is a neighbourhood of x in the original topology. To show this, we can assume that $\mathbb{B}_r(x) \subseteq U$, for some U diffeomorphic to $\mathbb{R}^n$ and such that here $\mathcal{D}$ has a global frame. Since $\mathbb{B}_r^U(x) \subseteq \mathbb{B}_r(x)$ (see Remark 2.22), we are reduced to showing the thesis on $U \cong \mathbb{R}^n$, which is the already treated special case.

2.6. Local existence of length minimizers, lack of uniqueness

We now define length minimizers and show that they locally exist.

Definition 2.40. A horizontal curve $\gamma \in AC([0,T],M)$ is a *length minimizer* or a *geodesic* if, for any horizontal curve $\delta([0,T],M)$ with $\delta(0) = \gamma(0)$ and $\delta(T') = \delta(T)$, we have $L(\gamma) \leq L(\delta)$. Moreover, if γ has constant speed, we say that it is a *strict length minimizer* if it is a length minimizer and if, whenever equality occurs in the preceding inequality for some constant speed δ, we have $\delta \equiv \gamma$.

Remark 2.41. Equivalently, a horizontal curve $\gamma \in AC([0,T],M)$ is a length minimizer if

$$L(\gamma) = d(\gamma(0), \gamma(T)).$$

The condition for strict length minimality can be reformulated by asking that, whenever equality occurs for some (not necessarily constant-speed) δ, then $\delta = \gamma \circ h$ for some AC increasing $h : [0,T] \to [0,T]$.

Proposition 2.42 (Local existence of geodesics). Any point $x \in M$ has an open neighbourhood $U \subseteq M$ such that, for any $y \in U$, there exists a constant-speed $\gamma \in \Omega_{x,y}$ which is a length minimizer. We can choose U in such a way that $\gamma([0,1]) \subseteq M$ for any length minimizer connecting x to y.

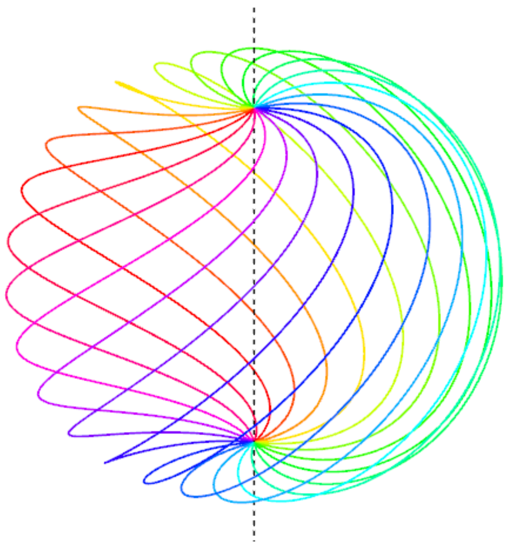
Proof. Assume first that $M = \mathbb{R}^n$ and $x = 0$. By Corollary 2.38 (or simply by Remark 2.17) we can find some $r > 0$ satisfying $U := \mathbb{B}_r(x) \subseteq B_1$. Now fix any $y \in U$ and pick a sequence of curves $\gamma_k \in \Omega_{x,y}$ such that $L(\gamma_k) \to d(x,y)$. We can assume that these curves have constant speed, so that $|\dot{\gamma}_k|_g = L(\gamma_k) \leq C$ for a suitable constant $C < +\infty$. Since $\gamma_k([0,1]) \subseteq \overline{B}_1$, which is a compact set independent of k, we deduce that $|\dot{\gamma}_k|_e \leq C'$ for some $C' < +\infty$. Thus the curves γ_k have a common Lipschitz constant, so (up to subsequences) we can assume $\gamma_k \to \gamma$ in $C^0([0,1],\mathbb{R}^n)$, as well as $\dot{\gamma}_k \rightharpoonup \dot{\gamma}$ in $L^2([0,1],\mathbb{R}^n)$. The fact that γ is horizontal can be seen by the same argument used in Step 4 of the proof of Theorem 2.23. Finally, $\gamma(0) = x$, $\gamma(1) = y$ and

$$L(\gamma)^2 \leq 2E(\gamma) \leq \liminf_{k\to\infty} 2E(\gamma_k) = \lim_{k\to\infty} L(\gamma_k)^2 = d(x,y)^2.$$

We used the lower semicontinuity of the energy, which was already proved in Step 3 of the aforementioned proof. This proves that γ is the required geodesic.

In the general case, let V be a neighbourhood of x such that there exists a diffeomorphism $\phi : V \to \mathbb{R}^n$, with $\phi(x) = 0$. Again, we can find some $r > 0$ satisfying $U := \mathbb{B}_r(x) \subseteq \phi^{-1}(B_1)$ (B_1 denoting the usual Euclidean ball). In order to prove that U has the desired property, Remark 2.22 allows us to replace M with V, and hence with $\mathbb{R}^n$. We have thus reduced to the already treated case. The second part of the thesis follows from the fact that U was chosen to be a ball. $\square$

EXAMPLE 2.43. Contrary to the special case of Riemannian geometry, in sub-Riemannian manifolds we do not have local uniqueness of length minimizers: in Section 3.6 we will compute the geodesics in the Heisenberg group $\mathbb{H}$, which is the simplest example of a sub-Riemannian manifold which is not Riemannian (i.e. with $r < n$). As we will see, $\mathbb{H}$ can be identified with $\mathbb{R}^3$ by using the exponential coordinates and we will show that *any* nonzero point z on the x_3-axis (in particular, as close as we like to 0) is connected to the origin by a one-parameter family of geodesics, which are obtained from each other by a rotation around the x_3-axis. The situation is illustrated in the following picture, where the upper point is z and the lower one is e.



2.7. First order minimality conditions, normal and abnormal geodesics

In this section we derive the first order necessary conditions for a given horizontal path $\gamma \in H^1([0,1], M)$ to be an energy minimizer, which is equivalent to being a constant-speed length minimizer, as we noticed in Remark 2.15. In Riemannian geometry one can simply perform a first variation and obtain that geodesics have to satisfy a suitable Euler-Lagrange equation, which is an ordinary differential equation. Smoothness of geodesics is then easily deduced by a simple bootstrap argument.

In sub-Riemannian geometry we cannot directly adapt this method, due to the horizontality constraint. We will obtain some necessary conditions using the method of Lagrange multipliers. The possibility of applying this method relies on the fact that we still have a natural way of parametrizing horizontal curves starting from $x := \gamma(0)$, which is provided by the set of admissible controls $\mathcal{U} = \mathcal{U}_{x,1}$ (assuming, as we will do throughout the rest of this chapter, that there is a fixed global orthonormal frame $\mathcal{D} = \langle X_1, \dots, X_r \rangle$). While minimizing the energy of the curve corresponding to $u \in \mathcal{U}$, i.e. $\frac{1}{2}\|u\|^2$, we have to impose the constraint that $\mathrm{End}(x,u) = y$, where $y := \gamma(1)$. This constraint can be degenerate: in this case, Lagrange multipliers could fall in the so-called abnormal case and become insufficient to deduce the regularity of γ.

In what follows, γ is fixed and the corresponding control is called $\overline{u}$. The starting point x will be omitted when dealing with the endpoint map.

DEFINITION 2.44. The *extended endpoint map* is defined as

$$\mathrm{extEnd} : \mathcal{U} \to M \times \mathbb{R}, \quad \mathrm{extEnd}(u) := (\mathrm{End}(u), E(u)),$$

where $E(u) := \frac{1}{2} \|u\|_2^2$ is the energy of the horizontal curve associated to u.

DEFINITION 2.45. A map $f : X \to Y$ between two topological spaces X, Y is *locally open at x* if, for any neighbourhood U of x, $f(U)$ is a neighbourhood of $f(x)$.

The key observation is that, since γ minimizes the energy in $\Omega_{x,y}$, extEnd cannot be locally open at $\overline{u}$: indeed, if this were not the case, we could find some $u \in \mathcal{U}$ (arbitrarily close to $\overline{u}$) with $\mathrm{End}(u) = \mathrm{End}(\overline{u}) = y$ and $E(u) < E(\overline{u})$. Let us use the same notation introduced before the statement of Lemma 2.32. We will simply write $\widehat{\mathrm{End}}$ (instead of $\widehat{\mathrm{End}}_1$), as well as

$$\widehat{\mathrm{extEnd}} := \left(\widehat{\mathrm{End}}, E\right).$$

LEMMA 2.46. The differential of $\widehat{\mathrm{extEnd}}$ at $\overline{u}$ is given by

$$d\,\widehat{\mathrm{extEnd}}_{\overline{u}}[v] = \left(\int_0^1 \Phi_t^* \langle v(t), X\rangle (x)\, dt, \int_0^1 \langle \overline{u}(t), v(t)\rangle\, dt\right).$$

Proof. This immediately follows from Lemma 2.32, combined with the fact that

$$E(\overline{u} + v) = E(\overline{u}) + \int_0^1 \langle \overline{u}(t), v(t)\rangle\, dt + O(\|v\|_2^2)$$

(which gives $dE_{\overline{u}}[v] = \int_0^1 \langle \overline{u}(t), v(t)\rangle\, dt$). □

Notice that $\widehat{\mathrm{extEnd}}$ cannot be locally open at $\overline{u}$, as well. Thus its differential cannot be surjective: this means that there exists a nonzero covector $(\overline{\lambda}, \overline{\mu}) \in T_x^*M \times \mathbb{R} = T_{\widehat{\mathrm{extEnd}}(\overline{u}}(M \times \mathbb{R})$ which vanishes on $\mathrm{im}\, d\,\widehat{\mathrm{extEnd}}_{\overline{u}}$, i.e.

$$\int_0^1 \overline{\lambda}\Phi_t^* \langle v(t), X\rangle (x)\, dt + \overline{\nu} \int_0^1 \langle \overline{u}(t), v(t)\rangle\, dt = 0$$

for all $v \in L^2([0,1], \mathbb{R}^r)$. Rescaling $\overline{\lambda}$ if necessary, we can assume that $\overline{\mu} \in \{0, 1\}$. Let us remark that, for any vector field Y,

$$\left\langle \overline{\lambda}, \Phi_t^* Y(x)\right\rangle = \left\langle \left(\Phi_t^{-1}\right)^* \overline{\lambda}, Y\left(\Phi_t(x)\right)\right\rangle = \left\langle \left(\Phi_t^{-1}\right)^* \overline{\lambda}, Y(\gamma(t))\right\rangle.$$

Hence, defining $\lambda(t) := \left(\Phi_t^{-1}\right)^* \overline{\lambda} \in T_{\gamma(t)}^* M$, we deduce

$$\int_0^1 \lambda(t) \langle v(t), X\rangle (\gamma(t))\, dt + \overline{\nu} \int_0^1 \langle \overline{u}(t), v(t)\rangle\, dt = 0.$$

Since v is arbitrary, we arrive at

$$\langle \lambda(t), X_i(\gamma(t))\rangle + \overline{\nu}\overline{u}_i(t) = 0$$

a.e., for all $i = 1, \dots, r$.

DEFINITION 2.47. The curve $\lambda : [0,1] \to T^*M$, associated to a given covector $\overline{\lambda} \in T_x^*M$ as above, is called *dual curve*.

DEFINITION 2.48. In general, we say that a curve $\gamma \in H^1([0,1], M)$ with control $\overline{u}$ is an *extremal* if $\overline{u}$ is a critical point for extEnd (with starting point $x := \gamma(0)$), or equivalently if $\overline{u}$ is a critical point for $\widehat{\mathrm{extEnd}}$. We say that γ is a *normal extremal* if there exists a (necessarily nonzero) covector $(\overline{\lambda}, 1) \in \left(\operatorname{im} d\,\widehat{\mathrm{extEnd}}_{\overline{u}}\right)^{\perp}$, i.e. vanishing on $\operatorname{im} d\,\widehat{\mathrm{extEnd}}_{\overline{u}}$; we say that γ is an *abnormal extremal* if there exists a nonzero covector $(\overline{\lambda}, 0) \in \left(\operatorname{im} d\,\widehat{\mathrm{extEnd}}_{\overline{u}}\right)^{\perp}$. We emphasize that this definition makes sense only in presence of a fixed global frame $X_1, \dots, X_r$ for $\mathcal{D}$.

In the normal case we could well have $\overline{\lambda} = 0$, while in the abnormal case $\overline{\lambda}$ is forced to be nonzero, so that $\lambda(t) \neq 0$ for all $t \in [0,1]$. It should be clear that γ is an extremal iff it is a normal or abnormal extremal.

DEFINITION 2.49. Given a dual curve λ, if there exists $\overline{\nu} \in \mathbb{R}$ such that

$$\left(\overline{\lambda}, \overline{\nu}\right) \in \left(\operatorname{im} d\,\widehat{\mathrm{extEnd}}_{\overline{u}}\right)^{\perp} \setminus \{0\},$$

we say that the couple (γ, λ) is a *biextremal* (here $\overline{\lambda} := \lambda(0)$). If $\overline{\nu} = 1$ we say that (γ, λ) is a *normal biextremal*, while if $\overline{\nu} = 0$ we say that it is an *abnormal biextremal*.

We have proved the following result.

THEOREM 2.50 (first order necessary conditions). Given a constant-speed length minimizer (or more generally an extremal) $\gamma : [0,1] \to M$ with control $\overline{u}$, there exists a nonzero multiplier $(\overline{\lambda}, \overline{\nu})$ (with $\overline{\nu} \in \{0,1\}$ and $\overline{\lambda} \in T^*_{\gamma(0)}M$) such that, calling λ the corresponding dual curve, it holds

$$\langle \lambda(t), X_i(\gamma(t)) \rangle + \overline{\nu}\,\overline{u}_i(t) = 0 \tag{2.2}$$

for a.e. t and all $i = 1, \dots, r$.

COROLLARY 2.51. In the Riemannian case, where $r = n$, abnormal extremals do not exist. Thus, any length minimizer is necessarily a normal extremal.

Proof. Assume by contradiction that the statement of Theorem 2.50 holds with $\overline{\nu} = 0$. Then we have $\lambda(t) \in \mathcal{D}^{\perp}_{\gamma(t)} = \{0\}$ for any t. In particular, $(\overline{\lambda}, \overline{\nu}) = 0$, which contradicts the nontriviality condition for the multiplier. □

REMARK 2.52. By Theorem 2.50, a biextremal (γ, λ) cannot be both normal and abnormal, unless $\overline{u} \equiv 0$, which corresponds to the case of the constant curve $\gamma(t) \equiv x$. Conversely (assuming $r < n$), choosing $\overline{\lambda} \in \mathcal{D}^{\perp}_x \setminus \{0\}$, the constant couple $(\gamma, \lambda) := (x, \overline{\lambda})$ is a normal and abnormal biextremal. Nonetheless, there are nontrivial examples of extremals γ which are both normal and abnormal (with respect to different covectors $\overline{\lambda}$).

REMARK 2.53. The same argument can be followed backwards: we obtain that, if $\overline{\nu} \in \mathbb{R}$ and $\lambda : [0,1] \to T^*M$ are such that $\lambda(t) \in T^*_{\gamma(t)}M$, $\lambda(t) = \left(\Phi_t^{-1}\right)^* \lambda(0)$, $(\overline{\nu}, \lambda(\cdot))$ solve (2.2) and $(\overline{\nu}, \lambda(0)) \neq 0$, then γ is an extremal (and (γ, λ) is a biextremal). Notice that $(\overline{\nu}, \lambda(0)) \neq 0$ is equivalent to asking that $\overline{\nu} \neq 0$ or $\lambda \not\equiv 0$ (since $\lambda \not\equiv 0$ iff $\lambda(0) \neq 0$ iff $\lambda(t) \neq 0$ for all t).

In order to deduce information about the regularity of γ, it is useful to express the dual curve as the solution of a suitable ordinary differential equation.

LEMMA 2.54. We have $\lambda \in H^1([0,1], T^*M)$. Moreover, choosing local coordinates (x,p) for T^*M, the p-component of λ solves the differential equation

$$\dot{p}(t) = -p(t) \langle \overline{u}(t), dX(\gamma(t)) \rangle . \tag{2.3}$$

Here $\langle \overline{u}(t), dX(\gamma(t)) \rangle$ is shorthand for $\sum_i \overline{u}_i(t) dX_i(\gamma(t))$, dX_i is identified with the $n \times n$ Jacobian matrix of X_i and $p(t)$ is viewed as a row vector. Conversely, any curve $\lambda \in H^1([0,1], T^*M)$ which lifts γ and solves locally this differential equation is the dual curve associated to $\lambda(0)$.

Proof. Let $[a,b] \subseteq [0,1]$ be any interval such that $\gamma([a,b])$ lies in the domain of the local chart. Let us show that the thesis holds on $[a,b]$. We can assume that $a = 0$: indeed, defining $\overline{u}' \in L^2([0,1-a], \mathbb{R}^r)$ by $\overline{u}'(t) := \overline{u}(t+a)$ and denoting by Φ'_t the associated flow, we remark that $\overline{u}'$ is the control associated to $\gamma(a+\cdot)$ and that, since $\Phi_{t+a} = \Phi'_t \circ \Phi_a$,

$$\lambda(t+a) = (\Phi_{t+a}^{-1})^* \overline{\lambda} = ((\Phi'_t)^{-1})^* (\Phi_a^{-1})^* \overline{\lambda} = ((\Phi'_t)^{-1})^* \lambda(a),$$

so we can replace γ, λ, $\overline{u}$ and b with $\gamma(a+\cdot)$, $\lambda(a+\cdot)$, $\overline{u}'$ and $b-a$ (respectively) in what follows.

Let us denote by $p(t)$ the p-component of λ and by $\overline{p}$ the same for $\overline{\lambda}$. We have

$$p(t) = \overline{p}\, (d\Phi_t(\gamma(0)))^{-1} = \overline{p} J(t)^{-1},$$

where $J(t) := d\Phi_t(\gamma(0))$. Thus, as is shown in the proof of Proposition B.12, $p(t)$ is H^1-regular on $[0,b]$ and solves

$$\dot{p}(t) = -\overline{p} J(t)^{-1} \dot{J}(t) J(t)^{-1} = -\overline{p} J(t)^{-1} \langle \overline{u}(t), dX(\gamma(t)) \rangle = -p(t) \langle \overline{u}(t), dX(\gamma(t)) \rangle$$

(due to the well-known fact that $f(A) := A^{-1}$ has differential $df_A[B] = -A^{-1} B A^{-1}$). The converse statement follows from the fact that a curve solving locally the differential equation is uniquely determined by $\lambda(0)$. $\square$

THEOREM 2.55 (smoothness of normal extremals). Any normal extremal $\gamma \in H^1([0,1], M)$ is smooth, i.e. $\gamma \in C^\infty([0,1], M)$.

Proof. Let λ be a dual curve such that (γ, λ) is a normal biextremal. Let us prove, by induction on k, that $\gamma, \lambda \in C^k$ for any integer $k \geq 0$. The base case is clear: γ and λ are continuous, as they belong to H^1. Now assume that $\gamma, \lambda \in C^k$: since $\overline{\nu} = 1$, equation (2.2) gives

$$\overline{u}_i(t) = -\langle \lambda(t), X_i(\gamma(t)) \rangle \in C^k.$$

Thus, from $\dot{\gamma}(t) = \sum_{i=1}^r \overline{u}_i(t) X_i(\gamma(t)) \in C^k$, we deduce $\gamma \in C^{k+1}$. Similarly, (2.3) gives $\lambda \in C^{k+1}$. $\square$

As a consequence, the regularity problem for constant-speed length minimizers is reduced to the case of geodesics which are not normal extremals. This motivates the following definition.

DEFINITION 2.56. An extremal γ is said to be *strictly abnormal* if it is not normal, i.e. if $\left(\operatorname{im} d\,\widehat{\operatorname{extEnd}}_{\overline{u}}\right)^\perp$ does not contain any covector of the form $(\overline{\lambda}, 1)$.

REMARK 2.57. Of course, a strictly abnormal extremal is abnormal. Moreover, if γ is strictly abnormal, we have $\left(\operatorname{im} d\,\widehat{\operatorname{extEnd}}_{\overline{u}}\right)^\perp = Z \times \{0\}$ for some vector subspace Z of T_x^*M, so $\operatorname{im} d\,\widehat{\operatorname{extEnd}}_{\overline{u}} = W \times \mathbb{R}$ with $W = Z^\perp$. Notice that, in the case of a (strictly) abnormal extremal, Theorem 2.50 tells us only that there exists some dual curve λ satisfying $\lambda(t) \in \mathcal{D}^{\perp}_{\gamma(t)}$ for any t.

Let us now explore more in depth the structure of normal biextremals.

As $\gamma(t)$ is simply the projection of $\lambda(t) \in T^*M$ on the base space M, the couple (γ, λ) can be identified with the curve λ taking values in T^*M, with the caveat that, when working in local coordinates, $\lambda(t)$ needs to be represented with $2n$ coordinates.

In the case of a normal biextremal, λ evolves according to the Hamiltonian $H(\lambda) := -\frac{1}{2}\sum_{i=1}^r \langle \lambda, X_i \rangle^2$, with respect to the canonical symplectic structure of T^*M (we recall that the canonical symplectic form is $\omega = -d\theta$, θ being the tautological one-form, whose expression in local coordinates (x,p) is $\theta = \sum_{i=1}^n p_i dx_i$). This means that, in local coordinates (x,p), λ solves the system

$$\begin{cases} \dot{x} = \frac{\partial H}{\partial p}(x,p) \\ \dot{p} = -\frac{\partial H}{\partial x}(x,p), \end{cases}$$

or equivalently

$$\begin{cases} \dot{x}(t) = -\sum_i \langle p(t), X_i(\gamma(t)) \rangle X_i(\gamma(t)) \\ \dot{p}(t) = \sum_i \langle p(t), X_i(\gamma(t)) \rangle \langle p(t), dX_i(\gamma(t)) \rangle, \end{cases}$$

which follows from $\dot{\gamma}(t) = \sum_i \overline{u}_i(t) X_i(\gamma(t))$ and (2.3), keeping in mind that $\overline{u}_i(t) = -\langle \lambda(t), X_i(\gamma(t)) \rangle$.

Conversely, if $\lambda : [0,1] \to T^*M$ solves (locally) this Hamiltonian system, then the couple (γ, λ) is a normal biextremal, where $\gamma := \pi \circ \lambda$ is the projection on M. Indeed, by the first equation, the control associated to γ is given by $\overline{u}_i(t) = -\langle \lambda(t), X_i(\gamma(t)) \rangle$. Hence, the second equation of the system tells us that λ is a dual curve for γ (by Lemma 2.54). Remark 2.53, applied with $\overline{\nu} := 1$, shows that (γ, λ) is a normal biextremal.

As a corollary, given any normal biextremal (γ, λ) with control $\overline{u}$, we have the following identities:

$$H(\lambda(t)) = -\frac{1}{2}\sum_{i=1}^r \overline{u}_i^2(t) = -\frac{1}{2}\langle \dot{\gamma} \rangle^2(t), \tag{2.4}$$

$$-\langle \lambda(t), \dot{\gamma}(t) \rangle = -\sum_{i=1}^r \overline{u}_i(t) \langle \lambda(t), X_i(\gamma(t)) \rangle = \sum_{i=1}^r \overline{u}_i^2(t) = \langle \dot{\gamma} \rangle^2(t), \tag{2.5}$$

which both follow from the fact that $\overline{u}_i(t) = -\langle \lambda(t), X_i(\gamma(t)) \rangle$.

Moreover, the Hamiltonian is constant along λ: indeed, in local coordinates, writing $\lambda(t) = (x(t), p(t))$ we have

$$\frac{d}{dt}H = \frac{\partial H}{\partial x}\dot{x}(t) + \frac{\partial H}{\partial p}\dot{p}(t) = \frac{\partial H}{\partial x} \cdot \frac{\partial H}{\partial x} - \frac{\partial H}{\partial p} \cdot \frac{\partial H}{\partial x} = 0,$$

where we omitted the dependence of H and its derivatives on $(x(t), p(t))$, for simplicity. Thus, by (2.4), any normal extremal is automatically constant-speed.

2.8. Minimality of short normal extremals

In what follows, we will always use the redundant notation (γ, λ) for a biextremal, so that in local coordinates λ can be safely identified with a covector in $(\mathbb{R}^n)^*$.

THEOREM 2.58. Any normal extremal $\gamma : [a,b] \to M$ is locally length minimizing, which means that, for any $s \in [a,b]$, there exists some $\epsilon' > 0$ such that $\gamma|_{[s-\epsilon',s+\epsilon']\cap[a,b]}$ is a length minimizer. As a consequence, there exists some $\delta > 0$ such that, whenever $s, s' \in [a,b]$ and $0 < s' - s \leq \delta$, $\gamma|_{[s,s']}$ is a length minimizer.

Proof. The existence of δ follows from the first part by a compactness argument: assume by contradiction that there exist $s_n, s'_n \in [a,b]$ with $0 < s'_n - s_n \to 0$ and such that $\gamma|_{[s,s']}$ is not a length minimizer. Up to subsequences $s_n \to s$ for some $s \in [a,b]$ and we have $s'_n \to s$, as well. Choosing $\epsilon' > 0$ as in the first part of the thesis, for large n we have $[s_n, s'_n] \subseteq [s-\epsilon', s+\epsilon'] \cap [a,b] =: I$. Since $\gamma|_I$ is a length minimizer, $\gamma|_{[s_n,s'_n]}$ has to be a length minimizer as well, contradiction.

In order to prove the first part, let λ be a dual curve such that (γ, λ) is a normal biextremal. We can assume that $s = 0$ and that γ has unit speed (since, if $|\dot\gamma| \equiv v > 0$, $\gamma(v^{-1}\cdot)$ is unit-speed and $(\gamma(v^{-1}\cdot), v^{-1}\lambda(v^{-1}\cdot))$ still solves the Hamiltonian system). The core of the proof will consist in finding a suitable calibration $-\Lambda$, which will be a smooth, closed (and, in fact, exact) one-form Λ defined on a suitable neighbourhood V of $\gamma(0)$, such that

$$\langle -\Lambda(x), v\rangle \leq |v| \tag{2.6}$$

for any $v \in \mathcal{D}_x$, with equality when $x = \gamma(t)$ and $v = \dot\gamma(t)$ for some (small) t.

In order to build the calibration, we can work in local coordinates. We will go back to M in the last step of the proof.

Step 1. Up to translating and rotating the coordinates, we can assume that $\gamma(0) = 0$ and $\lambda(0) = e^*_n$ ($e^*_1, \ldots, e^*_n$ being the usual dual basis of $(\mathbb{R}^n)^*$). Let $H' := \mathbb{R}^{n-1} \times 0$. We remark that $\dot\gamma(0) \notin T_0H'$, since $\langle \lambda(0), \dot\gamma(0)\rangle = 1$ (by (2.5)). For a sufficiently small neighbourhood U' of 0 in $\mathbb{R}^{n-1}$, we can find a (unique) smooth $\xi : U' \to (\mathbb{R}^n)^*$ such that $\xi(0) = \lambda(0)$ and $H((x',0), \xi(x')) = -\frac{1}{2}$, $\xi(x') \in \langle e^*_n\rangle$ for any $x' \in U'$.

Possibly shrinking U', we can assume that the Hamiltonian system has a solution defined on $(-\epsilon, \epsilon)$, for every initial condition

$$(x,p)(0) = ((x',0), \xi(x')),$$

as x' varies in U'. Let us call $(\Gamma(x',t), \Lambda(x',t))$ the solution at time t and notice that $\Gamma(0,t) = \gamma(t)$, $\Lambda(0,t) = \lambda(t)$ (when $t \geq 0$). Shrinking U' and ϵ, we can assume that $\Gamma|_{U'\times(-\epsilon,\epsilon)}$ is a diffeomorphism onto its image V ($d\Gamma_{(0,0)}$ is invertible because, as remarked earlier, $\dot\gamma(0) = d\Gamma_{(0,0)}[(0,1)] \notin T_0H' = d\Gamma_{(0,0)}[\mathbb{R}^{n-1} \times \{0\}]$). Thus x' and t can be viewed as smooth functions defined on V, as well as Λ, by composition.

The fact that H is preserved along the Hamiltonian flow tells us that, for any $x \in V$, $H(x, \Lambda(x',t)) = -\frac{1}{2}$, or equivalently

$$\sum_{i=1}^{r} \langle \Lambda(x',t), X_i(x)\rangle^2 = 1. \tag{2.7}$$

Step 2. We now show that $-\Lambda$ is the desired calibration. In order to obtain (2.6), let $v \in \mathcal{D}_x$. Writing $x = \Gamma(x',t)$ and $v = \sum_{i=1}^r h_i X_i(x)$, we compute

$$\langle -\Lambda, v\rangle = -\sum_{i=1}^{r} h_i \langle \Lambda(x',t), X_i(x)\rangle \leq \left(\sum_{i=1}^{r} h_i^2\right)^{1/2} = |v|,$$

thanks to (2.7) and the Cauchy-Schwarz inequality. If $|v| = 1$, equality holds exactly when

$$v = -\sum_{i=1}^{r} \left\langle \Lambda(x', t), X_i(x) \right\rangle X_i(x) = \dot{\Gamma}(x', t)$$

(here $\dot{\Gamma} = \frac{d}{dt}\Gamma$); in particular, it holds when $x = \gamma(t)$ and $v = \dot{\gamma}(t)$, for some $t \in [0, \epsilon)$.

In order to obtain the exactness of $-\Lambda$, we show that in fact $-\Lambda = dt$. We define

$$Y(x) := \dot{\Gamma}(x', t),$$

i.e. $Y(x)$ is the speed (at x) of the extremal passing through x. The fact that (Γ, Λ) solve the Hamiltonian system gives

$$Y(x) = -\sum_i \left\langle \Lambda(x', t), X_i(x) \right\rangle X_i(x), \tag{2.8}$$

$$\begin{aligned} \dot{\Lambda}(x', t) &= \sum_i \left\langle \Lambda(x', t), X_i(x) \right\rangle \left\langle \Lambda(x', t), dX_i(x) \right\rangle \\ &= -\left\langle \Lambda(x', t), dY(x) \right\rangle - \sum_i d\left(\left\langle \Lambda(x', t), X_i(x) \right\rangle\right) \left\langle \Lambda(x', t), X_i(x) \right\rangle \\ &= -\left\langle \Lambda(x', t), dY(x) \right\rangle + dH(x, \Lambda(x', t)) = -\left\langle \Lambda(x', t), dY(x) \right\rangle, \end{aligned} \tag{2.9}$$

thanks to the constancy of $H(x, \Lambda(x', t))$.

Now $-\Lambda$ and dt clearly agree on $d\Gamma_{(x',t)}[(0, 1)] = Y(x)$: their common value is 1 (because $\langle \Lambda, Y(x) \rangle = 2H(x, \Lambda(x', t)) = -1$, by (2.8)). It suffices to show that, given any $w \in \mathbb{R}^{n-1}$, they agree on $d\Gamma_{(x',t)}[w, 0]$, as well. But dt always vanishes on this vector, while $-\Lambda$ vanishes on it if $t = 0$ (recall that $\Lambda(x', 0) = \xi(x')$ was chosen to be a multiple of e_n^*). Finally,

$$\frac{d}{dt} \left\langle \Lambda(x', t), d\Gamma_{(x',t)}[w, 0] \right\rangle = 0,$$

since

$$\begin{aligned} \left\langle \frac{d}{dt}\Lambda(x', t), d\Gamma_{(x',t)}[w, 0] \right\rangle &= -\left\langle \Lambda(x', t), dY(x) \circ d\Gamma_{(x',t)}[w, 0] \right\rangle, \\ \left\langle \Lambda(x', t), \frac{d}{dt} d\Gamma_{(x',t)}[w, 0] \right\rangle &= \left\langle \Lambda(x', t), dY(x) \circ d\Gamma_{(x',t)}[w, 0] \right\rangle, \end{aligned}$$

thanks to (2.9). So $\left\langle -\Lambda(x', t), d\Gamma_{(x',t)}[w, 0] \right\rangle$, being constant in t, has to vanish identically.

Step 3. We are ready to conclude the proof. In order to avoid ambiguities we write $-\Lambda = df$ (instead of $-\Lambda = dt$). Let us choose any Carnot-Carathéodory ball $\mathbb{B}_r(\gamma(0)) \subseteq V$ and choose any $\epsilon' < \frac{r}{3}$. Let us write $[-\alpha, \beta] := [-\epsilon', \epsilon'] \cap [a, b]$ (for suitable $\alpha, \beta \geq 0$, not both vanishing). $\gamma|_{[-\alpha,\beta]}$ is a length minimizer: let $\delta : [-\alpha, \beta] \to M$ be a horizontal curve with $L(\delta) \leq L\left(\gamma|_{[-\alpha,\beta]}\right) = \alpha + \beta$, $\delta(-\alpha) = \gamma(-\alpha)$ and $\delta(\beta) = \gamma(\beta)$. As $d(\delta(t), \delta(-\alpha)) \leq L(\delta) \leq 2\epsilon'$ for any $t \in [-\alpha, \beta]$, we get

$$d(\delta(t), \gamma(0)) \leq 2\epsilon' + d(\gamma(-\alpha), \gamma(0)) \leq 3\epsilon' < r,$$

so that $\delta([-\alpha, \beta]) \subseteq \mathbb{B}_r(\gamma(0)) \subseteq V$. Finally,

$$\begin{aligned} \alpha + \beta &= \int_{-\alpha}^{\beta} \langle -\Lambda, \dot{\gamma}(t) \rangle \, dt = f(\gamma(\beta)) - f(\gamma(-\alpha)) = f(\delta(\beta)) - f(\delta(-\alpha)) \\ &= \int_{-\alpha}^{\beta} \left\langle -\Lambda, \dot{\delta}(t) \right\rangle \, dt \leq \int_{-\alpha}^{\beta} \left|\dot{\delta}\right|(t) \, dt = L(\delta). \end{aligned} \tag{2.10}$$

□

REMARK 2.59. The proof shows that $\gamma|_{[-\alpha,\beta]}$ is in fact a *strict* length minimizer, i.e., if $\delta : [-\alpha, \beta] \to M$ is a constant-speed horizontal curve with the same endpoints, then $L(\delta) < L(\gamma)$ unless $\delta \equiv \gamma$: indeed, if $L(\delta) = L(\gamma)$, δ has to be unit-speed and, since equality holds in (2.10),

$$\left\langle -\Lambda, \dot{\delta}(\tau) \right\rangle = \left| \dot{\delta} \right| (\tau) = 1$$

for a.e. τ. As we remarked in Step 2 of the preceding proof, this implies $\dot{\delta}(\tau) = \dot{\Gamma}(x', t)$ for a.e. τ (writing $\delta(\tau) = \Gamma(x', t)$). Now, thinking x' as a function $x' : V \to U'$ as in the above proof, we have

$$\frac{d}{d\tau}(x' \circ \delta)(\tau) = dx'_{\delta(\tau)}[\dot{\delta}(\tau)] = dx'_{\Gamma(x',t)}[\dot{\Gamma}(x', t)] = 0$$

for a.e. τ, which implies

$$x'(\delta(\tau)) \equiv x'(\delta(-\alpha)) = x'(\gamma(-\alpha)) = 0$$

for any τ. Similarly, thinking t as a map $t : V \to (-\epsilon, \epsilon)$, we get $\frac{d}{d\tau}(t \circ \delta)(\tau) = 1$, which combined with $t(\delta(-\alpha)) = t(\gamma(-\alpha)) = -\alpha$ gives $t(\delta(\tau)) = \tau$ for all τ. Thus $\delta(\tau) = \Gamma(0, \tau) = \gamma(\tau)$, i.e. $\delta \equiv \gamma$.

CHAPTER 3

Carnot groups

In this chapter we restrict our attention to a special class of sub-Riemannian manifolds, namely Carnot groups. Their importance in sub-Riemannian geometry is due to the fact that they serve the same role of tangent spaces to sub-Riemannian manifolds, much as Euclidean spaces in Riemannian geometry: the precise statement of this principle is given in Section 3.4.

3.1. Definition and basic properties

Some basic facts about Lie groups are listed at the beginning of Appendix D.

Definition 3.1. A *stratified group* is a simply connected Lie group $\mathbb{G}$ whose Lie algebra $\mathfrak{g}$ admits a *stratification*, i.e. a decomposition

$$\mathfrak{g} = V_1 \oplus \cdots \oplus V_s$$

such that $V_{i+1} = [V_1, V_i]$ for $i = 1, \dots, s-1$ and $[V_1, V_s] = \{0\}$. We define $r := \dim V_1$ and, for any $X \in \mathfrak{g}$, we denote by $\overline{\pi}_i(X)$ its projection on V_i (so that $X = \overline{\pi}_1(X) + \cdots + \overline{\pi}_s(X)$).

Remark 3.2. It is a well-known fact in the theory of Lie groups that simply connected Lie groups are determined, up to isomorphism, by (the isomorphism class of) their Lie algebra. Moreover, a deep result, known as *Ado's theorem*, guarantees that any abstract Lie algebra $\mathfrak{g}$ (over $\mathbb{R}$) can be realized as the Lie algebra of a simply connected Lie group $\mathbb{G}$.

In the case of nilpotent Lie algebras, we can find a formula for the group law of $\mathbb{G}$ using the global chart provided by the exponential map, together with the Baker-Campbell-Hausdorff formula: see Proposition D.7. Put in another way, we can exhibit *a posteriori* such a group $\mathbb{G}$, by choosing $\mathbb{G} := \mathfrak{g}$ and defining a group operation by means of the Baker-Campbell-Hausdorff formula, as is done in the proof of Proposition D.7 (but we know that such operation is associative just because we already know that a Lie group which realizes $\mathfrak{g}$ exists, so in fact we cannot avoid the appeal to Ado's theorem).

Proposition 3.3. The Lie algebra $\mathfrak{g}$ of a stratified Lie group $\mathbb{G}$ is graded, meaning that $[V_i, V_j] \subseteq V_{i+j}$ for all $i, j > 0$, with the convention that $V_k := \{0\}$ if $k > s$. In particular, $\mathfrak{g}$ is nilpotent.

Proof. We can assume $i, j \leq s$. The proof is by induction on i. When $i = 1$ the thesis holds by definition. Assuming that it holds for some i, let us prove it for $i + 1$. Let $X \in V_{i+1}$ and $Y \in V_j$: we have to show that $[X, Y] \in V_{i+j+1}$. Since $X \mapsto [X, Y]$ is linear, we can assume that $X = [Z, W]$ for some $Z \in V_1$ and $W \in V_i$. Jacobi's identity gives

$$[X, Y] = [[Z, W], Y] = [Z, [W, Y]] - [W, [Z, Y]].$$

By the inductive hypothesis we have $[W, Y] \in V_{i+j}$, so $[Z, [W, Y]] \in [V_1, V_{i+j}] = V_{i+j+1}$. Similarly, $[Z, Y] \in [V_1, V_j] = V_{j+1}$ and, by the inductive hypothesis again, $[W, [Z, Y]] \in V_{i+j+1}$.

From the property $[V_i, V_j] \subseteq V_{i+j}$ we get

$$\mathrm{ad}(Y_1)\cdots\mathrm{ad}(Y_s)Y_{s+1} \in V_{i_1+\cdots+i_{s+1}} = 0$$

whenever $Y_1 \in V_{i_1}, \dots, Y_{s+1} \in V_{i_{s+1}}$, since $i_1 + \cdots + i_{s+1} \geq s+1$. By the multilinearity of the left-hand side, we deduce that this holds for all $Y_1, \dots, Y_{s+1} \in \mathfrak{g}$. In particular,

$$\mathrm{ad}(Y_1)\cdots\mathrm{ad}(Y_s) = 0 \tag{3.1}$$

for all $Y_1, \dots, Y_s \in \mathfrak{g}$, proving the nilpotency of $\mathfrak{g}$. □

So any stratified group is a nilpotent Lie group (which means that its Lie algebra is nilpotent). By Proposition D.7 in the appendix, the exponential map $\exp : \mathfrak{g} \to \mathbb{G}$ is a diffeomorphism.

Associated to a fixed stratification $\mathfrak{g} = V_1 \oplus \cdots \oplus V_s$ is a left-invariant smooth distribution $\mathcal{D}$, given by

$$\mathcal{D}_g := dL_g[V_1],$$

where L_g denotes left multiplication by g (i.e. $L_g(x) := gx$). The fact that $\mathcal{D}$ is left-invariant means that $dL_g(\mathcal{D}_h) = \mathcal{D}_{gh}$, which immediately follows from $L_{gh} = L_g \circ L_h$.

Definition 3.4. A *Carnot group* is a stratified group $\mathbb{G}$ (with a fixed stratification), together with a positive definite inner product $\overline{g}$ on V_1. $\mathbb{G}$ is a sub-Riemannian manifold once we define the sub-Riemannian metric $g_x := (L_{x^{-1}})^*\overline{g}$ on $\mathcal{D}$, or more precisely

$$g_x(v, w) := \overline{g}(dL_{x^{-1}}[v], dL_{x^{-1}}[w])$$

for any $v, w \in \mathcal{D}_x$ and any $x \in \mathbb{G}$. Here, of course, $dL_{x^{-1}}$ should be interpreted as $d(L_{x^{-1}})_x$.

Remark 3.5. $\mathcal{D}$ satisfies the bracket-generating condition. Moreover, choosing a basis $X_1, \dots, X_r$ of V_1 and calling X_i^L the corresponding left-invariant vector fields, we have $\mathcal{D} = \langle X_1^L, \dots, X_s^L \rangle$, so Remark 2.4 shows that s is exactly the step of $\mathcal{D}$, in agreement with the notation introduced in Definition 2.8.

Remark 3.6. The metric g defined above is left-invariant: given $v, w \in \mathcal{D}_x$, we have $g(dL_a[v], dL_a[v]) = g(v, w)$ (again we write dL_a instead of $d(L_a)_x$), since

$$\begin{aligned} g(dL_a[v], dL_a[v]) &= \overline{g}(dL_{(ax)^{-1}} \circ dL_a[v], dL_{(ax)^{-1}} \circ dL_a[w]) \\ &= \overline{g}(dL_{x^{-1}} \circ dL_{a^{-1}} \circ dL_a[v], dL_{x^{-1}} \circ dL_{a^{-1}} \circ dL_a[w]) \\ &= g(v, w). \end{aligned}$$

Definition 3.7. For any $r > 0$ we define $d_r : \mathfrak{g} \to \mathfrak{g}$ by the formula

$$d_r(x) := \sum_{i=1}^{s} r^i \overline{\pi}_i(X),$$

which is a Lie algebra automorphism (thanks to the property $[V_i, V_j] \subseteq V_{i+j}$). Since $\mathbb{G}$ is simply connected, there is a unique Lie group homomorphism $\delta_r : \mathbb{G} \to \mathbb{G}$ such that $d(\delta_r)_e = d_r$. Such maps δ_r will be referred to as the *dilations* of $\mathbb{G}$.

Remark 3.8. By uniqueness we have $\delta_1 = \mathrm{id}_{\mathbb{G}}$ and (since $d_{rs} = d_r \circ d_s$) $\delta_{rs} = \delta_r \circ \delta_s$. So $\mathrm{id} = \delta_r \circ \delta_{r^{-1}} = \delta_{r^{-1}} \circ \delta_r$ and in particular all dilations are Lie group automorphisms. Moreover, $d(\delta_r)_x(\mathcal{D}_x) = \mathcal{D}_{\delta_r(x)}$: indeed, for any $v \in \mathcal{D}_x$ we can write $v = d(L_x)_e[w]$ (for a suitable $w \in V_1$) and it suffices to notice that

$$d(\delta_r)_x[v] = d(\delta_r \circ L_x)_e[w] = d(L_{\delta_r(x)} \circ \delta_r)_e[w] = d(L_{\delta_r(x)})_e[d_r(w)] \in \mathcal{D}_{\delta_r(x)}.$$

We used the identity $\delta_r \circ L_x = L_{\delta_r(x)} \circ \delta_r$, which holds since, for any $y \in \mathbb{G}$,

$$\delta_r \circ L_x(y) = \delta_r(xy) = \delta_r(x)\delta_r(y) = L_{\delta_r(x)} \circ \delta_r(y).$$

The computation which was used to show the invariance of $\mathcal{D}$ under δ_r also shows that

$$|d(\delta_r)_x[v]| = |d_r(w)| = r\,|w| = r\,|v|$$

for any $v \in \mathcal{D}_x$ (where $|v| := g(v,v)^{1/2}$, as usual).

PROPOSITION 3.9. Fix a horizontal curve $\gamma \in AC([0,T],\mathbb{G})$. For any $a \in \mathbb{G}$ and any $r > 0$ the curves $L_a \circ \gamma$ and $\delta_r \circ \gamma$ are horizontal and we have

$$L(L_a \circ \gamma) = L(\gamma), \quad L(\delta_r \circ \gamma) = rL(\gamma).$$

As a consequence, for any $x, y \in \mathbb{G}$, the Carnot-Carathéodory distance satisfies

$$d(ax, ay) = d(x,y), \quad d(\delta_r(x), \delta_r(y)) = rd(x,y).$$

Proof. Since $\frac{d}{dt}(L_a \circ \gamma)(t) = dL_a[\dot\gamma(t)] \in \mathcal{D}_{a\gamma(t)}$ for a.e. t (by left-invariance), we obtain that $L_a \circ \gamma$ is horizontal. By Remark 3.6,

$$L(L_a \circ \gamma) = \int_0^T |dL_a[\dot\gamma(t)]|\,dt = \int_0^T |\dot\gamma(t)|\,dt = L(\gamma).$$

Similarly, by Remark 3.8, $\delta_r \circ \gamma$ is horizontal as well, with

$$L(\delta_r \circ \gamma) = \int_0^T |d\delta_r[\dot\gamma(t)]|\,dt = \int_0^T r\,|\dot\gamma(t)|\,dt = rL(\gamma).$$

We deduce that $d(ax, ay) \le d(x,y)$ and $d(\delta_r(x), \delta_r(y)) \le rd(x,y)$. Replacing a with a^{-1} and x, y with ax, ay in the first inequality, we obtain the converse one. Similarly for the second inequality, replacing r with r^{-1} and x, y with $\delta_r(x), \delta_r(y)$. $\square$

COROLLARY 3.10. Fix any $a \in \mathbb{G}$ and any $r > 0$. A horizontal curve $\gamma : [0,T] \to \mathbb{G}$ is a length minimizer iff $L_a \circ \gamma$ is a length minimizer. Similarly, γ is a length minimizer iff $\delta_r \circ \gamma$ is.

Proof. Recall that γ is a length minimizer iff $L(\gamma) = d(\gamma(0), \gamma(T))$. The thesis now follows from the preceding proposition. $\square$

It will be useful to have at our disposal a (non-canonical) positive definite inner product $\overline{g}'$ on the whole of $\mathfrak{g}$, such that $\overline{g}'|_{V_1} = \overline{g}$ and $V_i \perp V_j$ whenever $i \ne j$. Notice that, defining $g'_x := (L_{x^{-1}})^*\,\overline{g}'$, we immediately obtain Lemma 2.13 in the special case of Carnot groups.

We fix from now on an *adapted basis* $X_1, \dots, X_n$ of $\mathfrak{g}$, i.e. a basis which is obtained by joining bases of $V_1, \dots, V_s$ (so that $X_1, \dots, X_r$ is a basis for V_1, $X_{r+1}, \dots, X_{r+\dim V_2}$ is a basis for V_2 and so on). We can (and will) also assume that $X_1, \dots, X_n$ are orthonormal with respect to the inner product $\overline{g}'$. We call X_i^L the corresponding left-invariant vector fields on $\mathbb{G}$.

DEFINITION 3.11. For any $i = 1, \dots, n$ there exists a unique $1 \le j \le s$ such that $X_i \in V_j$. We define $d(i) := j$ and we call $d(i)$ the *degree* of i. Moreover, let r_j denote the maximum index with degree j, so that $r_1 = r$, $r_s = n$ and $d(i) = j$ iff $r_{j-1} < i \le r_j$ (for any $i = 1, \dots, n$ and any $j = 1, \dots, s$, with the convention $r_0 := 0$). Notice that $\dim V_j = r_j - r_{j-1}$.

Definition 3.12. The *structure constants* $(c_{ij}^k)_{i,j,k=1,\dots,n}$ of $\mathfrak{g}$ are defined by the n^2 equations

$$[X_i, X_j] = \sum_{k=1}^{n} c_{ij}^k X_k.$$

Now we show that some properties, which hold only locally in sub-Riemannian manifolds, become global in the special case of Carnot groups.

Proposition 3.13. For any $x \in \mathbb{G}$ and any $r > 0$, the Carnot-Carathéodory ball $\mathbb{B}_r(x)$ has compact closure.

Proof. Recalling Corollary 2.38, we can find some $\epsilon > 0$ such that $\mathbb{B}_\epsilon(e)$ has compact closure in $\mathbb{G}$. Proposition 3.9 now gives

$$\mathbb{B}_r(x) = L_x\left(\mathbb{B}_r(e)\right) = L_x \circ \delta_{\epsilon^{-1}r}\left(\mathbb{B}_\epsilon(e)\right),$$

so $\mathbb{B}_r(x)$ has compact closure, as well. $\square$

Proposition 3.14. For any $x, y \in \mathbb{G}$ there exists a geodesic connecting x to y.

Proof. By Proposition 2.42, we can find some open neighbourhood U of e such that e can be connected to any point $z \in U$ by a length minimizer. For any fixed $x, y \in \mathbb{G}$, we have $\delta_r(x^{-1}y) \in U$ if $r > 0$ is small enough. Let γ be a length minimizer connecting e to $\delta_r(x^{-1}y)$: by Corollary 3.10, $L_x \circ \delta_{r^{-1}} \circ \gamma$ is the desired geodesic. $\square$

Proposition 3.15. For any fixed $\overline{x} \in \mathbb{G}$, denoting by $\mathcal{U} \subseteq L^2([0,1], \mathbb{R}^r)$ the open subset of admissible controls (with respect to $X_1^L, \dots, X_r^L$), we have $\mathcal{U} = L^2([0,1], \mathbb{R}^r)$.

We give a proof which shows that this fact holds in general in any Lie group (provided that we still use left-invariant vector fields).

Proof. Assume by contradiction that, for some control $u \in L^2([0,1], \mathbb{R}^r)$, we have the proper inclusion $I_{max}(\overline{x}, u) \subsetneq [0,1]$ (see Definition B.7). Fix any compact neighbourhood K of e. Writing $I_{max}(\overline{x}, u) = [0, T)$, by repeated application of Corollary B.11 we can find a sequence $t_j \uparrow T$ such that $\gamma(t_{j+1}) \notin \gamma(t_j)K$ for any j (where $\gamma : [0, T) \to \mathbb{G}$ is the trajectory associated to $(\overline{x}, u)$). Now,

$$u^j := u\mathbf{1}_{[t_j, t_{j+1}]} \in \mathcal{U}_{\gamma(t_j)}, \quad \mathrm{End}(\gamma(t_j), u^j) = \gamma(t_{j+1}):$$

indeed, the trajectory associated to $(\gamma(t_j), u^j)$ is

$$\gamma^j(t) := \begin{cases} \gamma(t_j) & t \in [0, t_j] \\ \gamma(t) & t \in [t_j, t_{j+1}] \\ \gamma(t_{j+1}) & t \in [t_{j+1}, 1]. \end{cases}$$

Thus, by left translation, we have

$$u^j \in \mathcal{U}_e, \quad \mathrm{End}(e, u^j) = \gamma(t_j)^{-1}\gamma(t_{j+1}) \notin K.$$

This contradicts the continuity of $\mathrm{End}(e, \cdot)$, which on the contrary gives $\mathrm{End}(e, u^j) \to e$ (as $u^j \to 0$ in $L^2([0,1], \mathbb{R}^r)$). $\square$

3.2. Exponential coordinates

As is shown by Proposition D.7 in the appendix, $\exp : \mathfrak{g} \to \mathbb{G}$ is a diffeomorphism. Moreover, $\mathfrak{g}$ can be identified with $\mathbb{R}^n$ by means of the previously fixed basis of $\mathbb{R}^n$. Thus we obtain a global coordinate chart on $\mathbb{G}$, defined by the correspondence

$$\mathbb{R}^n \ni (x_1, \dots, x_n) \leftrightarrow \exp(x_1 X_1 + \cdots + x_n X_n) \in \mathbb{G}.$$

These global coordinates will be called *exponential coordinates of the first kind*, or simply *exponential coordinates.*

DEFINITION 3.16. Given a monomial $x_1^{\alpha_1} \cdots x_n^{\alpha_n}$, the sum $\sum_{i=1}^n d(i)\alpha_i$ is called its *weighted degree.* A polynomial in $x_1, \dots, x_n$ is said to be *homogeneous* if the monomials which compose it have the same weighted degree. We use similar definitions for polynomials in two variables: the weighted degree of $x_1^{\alpha_1} \cdots x_n^{\alpha_n} y_1^{\beta_1} \cdots y_n^{\beta_n}$ is defined to be $\sum_{i=1}^n d(i)(\alpha_i + \beta_i)$.

REMARK 3.17. From Proposition D.7 we have $\exp(X)\exp(Y) = \exp(P(X,Y))$, with

$$P(X,Y) = X + Y + \sum_{p=1}^{s-1} \frac{(-1)^p}{p+1} \sum_{\substack{0 \le k_1, \dots, k_p < s \\ 0 \le \ell_1, \dots, \ell_p < s \\ k_i + \ell_i \ge 1}} \frac{(\operatorname{ad} X)^{k_1} (\operatorname{ad} Y)^{\ell_1} \cdots (\operatorname{ad} X)^{k_p} (\operatorname{ad} Y)^{\ell_p}}{(k_1 + \cdots + k_p + 1) k_1! \cdots k_p! \, \ell_1! \cdots \ell_p!} X. \tag{3.2}$$

In this double sum we have left out all the terms which are automatically zero by (3.1), in order to emphasize the fact that $P(X,Y)$ is a finite sum. Writing

$$X = \sum_i x_i X_i, \quad Y = \sum_j y_j X_j, \quad P(X,Y) = \sum_k P_k(X,Y) X_k,$$

by the multilinearity of each term in the formula for $P(X,Y)$ we get that $P_k(X,Y)$ is a polynomial in the variables $x_1, \dots, x_n, y_1, \dots, y_n$. Most importantly, $P_k(X,Y)$ is homogeneous with weighted degree $d(k)$: indeed, $P(X,Y)$ is a linear combination of terms of the form

$$\underbrace{[x_{i_{1,1}} X_{i_{1,1}}, [x_{i_{1,2}} X_{i_{1,2}}, \cdots,}_{k_1} \underbrace{[y_{j_{1,1}} X_{j_{1,1}}, [y_{j_{1,2}} X_{j_{1,2}}, \cdots,}_{\ell_1} [\cdots, x_i X_i] \cdots]$$

($k_2 + \ell_2 + \cdots + k_p + \ell_p$ additional blocks of commutators are there but are not displayed). Each such term equals a constant vector in some V_m multiplied by a monomial in $x_1, \dots, x_n, y_1, \dots, y_n$ with weighted degree m (which immediately follows by repeated application of the property $[V_a, V_b] \subseteq V_{a+b}$, recalling that $X_i \in V_{d(i)}$). Thus only terms containing a monomial with weighted degree $d(k)$ can contribute to $P_k(X,Y)$.

PROPOSITION 3.18. In exponential coordinates we have

$$X_i^L(x) = \partial_i + \sum_{j : d(j) > d(i)} f_{ij}(x) \partial_j,$$

for suitable polynomials $f_{ij} \in \mathbb{R}[x_1, \dots, x_n]$ which are homogeneous with weighted degree $d(j) - d(i)$.

Proof. We have

$$X_i^L(x) = dL_x \left[\frac{d}{dt} \exp(tX_i) \bigg|_{t=0} \right] = \frac{d}{dt} (x \cdot te_i) \bigg|_{t=0},$$

so, writing $X_i^L(x) = \sum_{j=1}^n f_{ij} \partial_j$, f_{ij} is given by the formula

$$f_{ij}(x) = \frac{d}{dt} P_j \left(\sum_k x_k X_k, tX_i \right) \bigg|_{t=0}.$$

By (3.2), when $d(j) \le d(i)$ we have

$$P_j\left(\sum_k x_k X_k, tX_i\right) = x_j + \delta_{ij} t$$

(since the double sum in (3.2) belongs to $V_{d(i)+1}$), so $f_{ij}(x) = \delta_{ij}$. For $d(j) > d(i)$, the right-hand side is obtained by selecting the terms in $P_j(\sum_k x_k X_k, tX_i)$ (thought as a polynomial in the variables $x_1, \dots, x_n, t$) where t appears with exponent 1 and dividing them by t. Since this polynomial is homogeneous with weighted degree $d(j)$ when t is replaced with y_i, it follows that f_{ij} is homogeneous as well, with weighted degree $d(j) - d(i)$. □

We now turn to a very simple, yet useful estimate, which can be seen as a special case of the ball-box estimates already stated in Theorem 2.39 in the general context of sub-Riemannian manifolds.

PROPOSITION 3.19 (Ball-box estimate). For any $x \in \mathbb{G}$, writing $x = \exp(x_1X_1 + \cdots + x_nX_n)$, we have

$$C^{-1} \max_{i=1,\dots,n} |x_i|^{1/d(i)} \le d(x,e) \le C \max_{i=1,\dots,n} |x_i|^{1/d(i)}$$

for a suitable $C > 1$ depending only on $\mathbb{G}$. As a consequence, calling

$$\mathrm{Box}_r := \left\{x : \max_{i=1,\dots,n} |x_i|^{1/d(i)} < r\right\},$$

we have the inclusions $\mathrm{Box}_{r/C} \subseteq \mathbb{B}_r(0) \subseteq \mathrm{Box}_{Cr}$.

Proof. Notice that, since $\delta_r \circ \exp = \exp \circ d_r$, we have

$$\delta_r(x) = \delta_r \circ \exp\left(\sum_i x_i X_i\right) = \exp \circ d_r \left(\sum_i x_i X_i\right) = \exp\left(\sum_i r^{d(i)} x_i X_i\right). \tag{3.3}$$

So, working in exponential coordinates, the i-th component of $\delta_r(x)$ equals $r^{d(i)}x_i$. Now let

$$f(x) := \max_{i=1,\dots,n} |x_i|^{1/d(i)}.$$

From the preceding discussion we deduce that $f \circ \delta_r(x) = rf(x)$. The distance from e satisfies the same property, i.e. $d(\delta_r(x), e) = rd(x,e)$ (by Proposition 3.9). Thus (since the thesis is trivial when $x = e$) we are left to show that the thesis holds for some C when $f(x) = 1$.

Let $K := f^{-1}(x)$, which is a compact set with respect to the standard topology of $\mathbb{R}^n$ (when viewed in exponential coordinates). By Corollary 2.38, K is compact with respect to the topology induced by d. Thus, $d(\cdot, e)$ has positive maximum and minimum values on K, which is the thesis. □

REMARK 3.20. Defining $S_r := \{x : d(x,e) = r\}$, these spheres are all homeomorphic to each other: in fact we have $S_r = \delta_r(S_1)$. Moreover, in exponential coordinates (i.e. identifying $\mathbb{G}$ with $\mathbb{R}^n$), we can define

$$F : \partial B_1 \to S_1, \quad F(x) := \delta_{d(x,e)^{-1}}(x).$$

For any $y \in S_1$ let $R_y := \{\delta_r(y) \mid r > 0\}$. Each R_y intersects ∂B_1 at exactly one point $G(y)$: indeed, the map $r \mapsto |\delta_r(x)|$ is strictly increasing (see (3.3)) and its limits at 0^+ and $+\infty$ are 0 and $+\infty$. We also have $F(G(y)) = y$, so F is surjective. F is injective as well: if $F(x) = F(x') =: y$, then $x, x' \in R_y$, so that $x = G(y) = x'$.

This proves that F is a homeomorphism. Hence, all the spheres S_r are homeomorphic to S^{n-1}. In the general context of sub-Riemannian manifolds, it is still an open problem whether any sphere which is sufficiently small (depending on its center) is homeomorphic to S^{n-1}.

PROPOSITION 3.21. Fix any nonzero $X \in V_1$ and let $\gamma : \mathbb{R} \to \mathbb{G}$, $\gamma(t) := \exp(tX)$. Then γ is a length minimizer between any couple of its points, i.e. $\gamma|_{[a,b]}$ is a length minimizer for any $a < b$.

Proof. We can clearly assume that $|X| = 1$. Let us fix two reals $a < b$. $\delta : [0, b-a] \to \mathbb{G}$, given by $\delta(t) := \gamma(t+a)$, is a reparametrization of $\gamma|_{[a,b]}$. Moreover,

$$\delta(t) = \exp((t+a)X) = \exp(aX)\exp(tX) = L_{\exp(aX)} \circ \gamma(t),$$

so by Corollary 3.10 it suffices to treat the case $a = 0$. Now assume $X = X_1$ (which we always can, up to choosing another adapted, orthonormal basis for $\mathfrak{g}$) and let $\eta \in \Omega_{e,\gamma(b)}$. Calling u the control associated to η, in exponential coordinates we have $\dot\eta_1 = u_1$ a.e. (since the first component of X_i^L vanishes when $i > 1$ and equals 1 for $i = 1$). Thus,

$$L(\eta) = \int_0^1 |u|(t)\,dt \geq \int_0^1 |u_1|(t)\,dt \geq \int_0^1 \dot\eta_1(t)\,dt = \eta_1(1) - \eta_1(0) = b,$$

while $L\left(\gamma|_{[0,b]}\right) = b$. □

3.3. Horizontal curves in Carnot groups

In Carnot groups, besides using controls, we have another useful way to parametrize horizontal paths starting at e.

Recall that, for any $1 \leq j \leq s$, $\overline{\pi}_j : \mathfrak{g} \to V_j$ denotes the canonical projection (with respect to the direct sum $\mathfrak{g} = V_1 \oplus \cdots \oplus V_s$). When $j = 1$ we will write $\overline{\pi}$ instead of $\overline{\pi}_1$.

LEMMA 3.22. There is a canonical group homomorphism $\pi : \mathbb{G} \to (V_1, +)$, given by

$$\pi : \exp(x_1X_1 + \cdots + x_nX_n) \mapsto x_1X_1 + \cdots + x_rX_r,$$

which depends only on the chosen stratification of $\mathfrak{g}$.

Proof. The smooth map given by the above formula does not depend on the choice of the adapted basis $X_1, \ldots, X_n$: indeed, we have the intrinsic formula $\pi = \overline{\pi} \circ \exp^{-1}$.

We are left to show that π is a group homomorphism. Given $g = \exp(x_1X_1 + \cdots + x_nX_n)$ and $g' = \exp(x_1'X_1 + \cdots + x_n'X_n)$, the Baker-Campbell-Hausdorff formula (3.2) tells us that $\exp^{-1}(gg')$ equals $(x_1 + x_1')X_1 + \cdots + (x_r + x_r')X_r$ plus some terms contained in $V_2 \oplus \ldots \oplus V_s$. So

$$\pi(gg') = \overline{\pi}(\exp^{-1}(gg')) = (x_1 + x_1')X_1 + \cdots + (x_r + x_r')X_r = \pi(g) + \pi(g').$$

□

This lemma can be easily generalized as follows.

DEFINITION 3.23. For any $1 \leq j \leq s$ we let $W_j := V_j \oplus \cdots \oplus V_s$ and $\mathbb{G}_j := \exp(W_j)$.

LEMMA 3.24. For any $1 \le j \le s$, $\mathbb{G}_j$ is a closed subgroup of $\mathbb{G}$. There exists a canonical homomorphism

$$\pi_j : \mathbb{G}_j \to V_j, \quad \exp(x_{r_{j-1}+1}X_{r_{j-1}+1} + \cdots + x_n X_n) \mapsto x_{r_{j-1}+1}X_{r_{j-1}+1} + \cdots + x_{r_j}X_{r_j},$$

which depends only on the chosen stratification of $\mathfrak{g}$.

Proof. The fact that $\mathbb{G}_j$ is a subgroup follows from the Baker-Campbell-Hausdorff formula, while its closedness is clear since exp is a diffeomorphism. Now recall that $X_{r_{j-1}+1}, \ldots, X_{r_j}$ is a basis of V_j. We have the intrinsic formula $\pi_j = \overline{\pi}_j \circ \exp|_{W_j}{}^{-1}$, which shows that π_j does not depend on the choice of the basis for $\mathfrak{g}$.

The assertion that $\pi_j : \mathbb{G}_j \to V_j$ is a homomorphism can be obtained exactly as in the previous proof. □

The next lemma will be used in the proof of Theorem 3.26.

LEMMA 3.25. The homomorphism $\pi : \mathbb{G} \to V_1$ satisfies $d\pi_x(X^L(x)) = X$, for any $x \in \mathbb{G}$ and any $X \in V_1$. Here we are using the canonical identification $T_{\pi(x)}V_1 \cong V_1$.

Proof. We have

$$X^L(x) = dL_x \left[\frac{d}{dt} \exp(tX) \Big|_{t=0} \right] = \frac{d}{dt}(x \exp(tX)) \Big|_{t=0},$$

so that, using the fact that π is a homomorphism,

$$d\pi_x(X^L(x)) = \frac{d}{dt}\pi(x\exp(tX))\Big|_{t=0} = \frac{d}{dt}(\pi(x) + \pi(\exp(tX)))\Big|_{t=0} = \frac{d}{dt}(tX)\Big|_{t=0} = X.$$

□

The following theorem says that there is a correspondence between horizontal curves starting at e and curves in V_1 starting at 0 (given by the composition with π)

THEOREM 3.26. For any $\gamma \in H^1([0,T], \mathbb{G})$, calling u the associated control, for a.e. t we have $\frac{d}{dt}(\pi \circ \gamma)(t) = \sum_{i=1}^r u_i(t)X_i$. Moreover, given any curve $\delta \in H^1([0,T], V_1)$ with $\delta(0) = 0$, there exists a unique horizontal curve $\gamma \in H^1([0,T], \mathbb{G})$ such that $\pi \circ \gamma = \delta$ and $\gamma(0) = e$.

Proof. Using Lemma 3.25 we obtain

$$\frac{d}{dt}(\pi \circ \gamma)(t) = d\pi_{\gamma(t)}[\dot\gamma(t)] = \sum_{i=1}^r u_i(t) d\pi_{\gamma(t)}[X_i^L(x)] = \sum_{i=1}^r u_i(t)X_i$$

for a.e. t. This also shows that, if $\pi \circ \gamma = \delta$, the control u is uniquely determined by δ (since $X_1, \ldots, X_r$ is a basis of V_1), proving the uniqueness part of the second statement.

Conversely, to prove the existence of a lift of δ, let us write $\dot\delta(t) = \sum_{i=1}^r u_i(t)X_i$ for a.e. t, for suitable $u_1, \ldots, u_r \in L^2([0,T])$. By Proposition 3.15, the horizontal curve γ starting at e and associated to the control $u := (u_1, \ldots, u_r)$ exists. The above argument shows that $\frac{d}{dt}(\pi \circ \gamma)(t) = \dot\delta(t)$ for a.e. t, which gives $\pi \circ \gamma \equiv \delta$ (since $\pi \circ \gamma(0) = 0 = \delta(0)$, as well). □

The following generalization can be obtained by a completely analogous proof (or alternatively can be deduced from the result just proved by a left translation).

THEOREM 3.27. Fix any $x \in \mathbb{G}$ and call $\overline{x} := \pi(x)$. Given any curve $\delta \in H^1([0,T],V_1)$ with $\delta(0) = \overline{x}$, there exists a unique horizontal curve $\gamma \in H^1([0,T],\mathbb{G})$ such that $\pi \circ \gamma = \delta$ and $\gamma(0) = x$.

3.4. Carnot groups as tangent spaces for sub-Riemannian manifolds

Carnot groups are model spaces among all sub-Riemannian manifolds, in that they play the same role as Euclidean spaces in the class of Riemannian manifolds. In this section we will limit ourselves to give a precise statement of this fact, whose proof is beyond the scope of this thesis. To begin with, let us give some basic definitions which allow to extend the notion of tangent space to a very general setting.

DEFINITION 3.28. The *Hausdorff distance* between two nonempty subsets A, B of a metric space (X, d) is

$$d_H(A,B) := \max\left\{\sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(x,y)\right\}.$$

Note that, in spite of the name, d_H in general is not a distance on $\mathcal{P}(X) \setminus \{\emptyset\}$ (since we have $d_H(A,\overline{A}) = 0$; we could also have $d_H(A,B) = +\infty$): it becomes a distance only when restricting to the class of compact subsets of X. In the sequel we will not need to guarantee that d_H is a distance, though.

DEFINITION 3.29. The *Gromov-Hausdorff distance* between two (nonempty) metric spaces X, Y is

$$d_{GH}(X,Y) := \inf d_H(i(X), j(Y)),$$

as i, j vary among all isometric embeddings $i : X \to Z$, $j : Y \to Z$ and Z varies among all metric spaces (we remark that there is no set-theoretic issue in this definition, as we can always assume that $Z = i(X) \cup j(Y)$, or even that $Z = i(X) \sqcup j(Y)$).

Again, d_{GH} is not really a distance: $d_{GH}(X,Y) = 0$ does not imply that X and Y are isometric; moreover, we could have $d(X,Y) = +\infty$, as well. However, it can be shown that d_{GH} becomes a distance on the class of nonempty compact metric spaces (considered up to isometries): see [AT04, Proposition 4.5.2].

DEFINITION 3.30. A sequence of pointed metric spaces (X_n, x_n) is said to converge to (X, x) if $d_{GH}(B_R^{X_n}(x_n), B_R^X(x)) \to 0$ for any $R > 0$. A similar definition can be given for the convergence of a family of metric spaces (X_λ, x_λ) (indexed by $\lambda \in \mathbb{R}^+$), as $\lambda \downarrow +\infty$.

DEFINITION 3.31. Given a metric space (X, d) and any $\lambda > 0$, the dilated metric space $(\lambda X, d_{\lambda X})$ is given by $\lambda X := X$ and $d_{\lambda X}(x,y) := \lambda d(x,y)$ for any $x, y \in \lambda X = X$. We say that a pointed metric space (Y, y) is a *tangent space of* X *at* $x \in X$ if

$$(\lambda X, x) \to (Y, y)$$

as $\lambda \uparrow +\infty$, in the sense of the previous definition.

Finally, we give a technical condition, on points of a sub-Riemannian manifold, which will be sufficient to have a Carnot group as a tangent space. As is shown in Remark 3.33, this condition is satisfied generically (in the topological sense).

DEFINITION 3.32. Given a sub-Riemannian manifold M and calling $\mathcal{D}$ its distribution, as usual, we say that $\mathcal{D}$ is *equiregular* at $x \in M$, or that x is an *equiregular point*, if there exists a neighbourhood U of x such that, for any $x' \in U$ and any $k \geq 1$,

$$\dim \mathrm{Lie}^k(\mathcal{D}, x') = \dim \mathrm{Lie}^k(\mathcal{D}, x),$$

(see Definition 2.5).

REMARK 3.33. Equiregular points form an open dense subset $W \subseteq M$: the fact that W is open is clear from the definition. In order to obtain that W is dense, fix a nonempty open set $U \subseteq M$. We have to show that $U \cap W \neq \emptyset$. Let

$$s(x) := \min \left\{ k : \mathrm{Lie}^k(\mathcal{D}, x) = T_x M \right\},$$

as we already did in Section 2.5, and

$$\underline{s} := \min \{ s(x) \mid x \in U \}.$$

Notice that, since $s(\cdot)$ is upper semicontinuous, $\{x \in U : s(x) = \underline{s}\}$ is an open subset of U. Possibly replacing U with this set, we can assume that $s(x) = \underline{s}$ for any $x \in U$. Now set $U_{\underline{s}} := U$ and define inductively, for $1 \leq k < \underline{s}$,

$$d_k := \max \left\{ \dim \mathrm{Lie}^k(\mathcal{D}, x) \mid x \in U_{k+1} \right\}, U_k := \left\{ x \in U_{k+1} : \dim \mathrm{Lie}^k(\mathcal{D}, x) = d_k \right\}.$$

Since $x \mapsto \dim \mathrm{Lie}^k(\mathcal{D}, x)$ is lower semicontinuous, U_{k+1} is open in U_k. By a reverse induction, we obtain that U_k is a nonempty open subset of M, for any $k = 1, \dots, \underline{s}$. U_1 is formed by equiregular points, since here $\dim \mathrm{Lie}^k(\mathcal{D}, \cdot) = n$ if $k \geq \underline{s}$, while $\dim \mathrm{Lie}^k(\mathcal{D}, \cdot) = d_k$ if $1 \leq k < \underline{s}$.

We are now ready to state this fundamental result, which was first obtained by Mitchell in [Mit85].

THEOREM 3.34 (Mitchell). If $\mathcal{D}$ is equiregular at $x \in M$, then there exists a Carnot group $\mathbb{G}$ (depending on x) such that $(\mathbb{G}, e)$ is the unique tangent space to (M, d_{CC}) at x.

A slightly more complicated statement holds for *any* point of M and was proved by Bellaïche in [Bel94].

THEOREM 3.35 (Bellaïche). Given any $x \in M$, (M, d_{CC}) has a unique tangent space at x, which equals $(H\backslash\mathbb{G}, [e])$ for some Carnot group $\mathbb{G}$ and a connected closed subgroup H, having the form $H = \exp(\mathfrak{h})$ for a suitable graded Lie subalgebra $\mathfrak{h}$ of the Lie algebra of $\mathbb{G}$.

Here the set of right cosets $H\backslash\mathbb{G}$ is endowed with the distance

$$d([g], [g']) := \inf \left\{ d(hg, h'g') \mid h, h' \in H \right\} = d(g, Hg')$$

(the last equality follows by left-invariance), which induces the quotient topology. It could well happen that $H = \{e\}$ even in the case that x is not equiregular.

3.5. First order necessary conditions

We now revisit the statement of Theorem 2.50, as well as that of Lemma 2.54, for an extremal $\gamma \in H^1([0,1], \mathbb{G})$. In the case of Carnot groups, the equations take a simpler form involving the structure constants of the Lie algebra $\mathfrak{g}$. This allows to compute explicit formulas for the normal extremals in many concrete Carnot groups. We will also use these reformulations in order to give an example of a strictly abnormal length minimizer in a Carnot group, in Section 4.6.

We can define a frame of left-invariant one-forms $\omega_1, \dots, \omega_n$ on $\mathbb{G}$ such that $\omega_i(X_j^L) = \delta_{ij}$ holds everywhere: it suffices to take the dual basis $\overline{\omega}_1, \dots, \overline{\omega}_n$ of $\mathfrak{g}^*$ (with respect to the fixed basis $X_1, \dots, X_n$ of $\mathfrak{g}$) and to define

$$\omega_i(x) := (L_{x^{-1}})^* \overline{\omega}_i = d\,(L_{x^{-1}})_x^* \,\overline{\omega}_i.$$

The relation $\omega_i(X_j^L) = \delta_{ij}$ now holds since

$$\omega_i(X_j^L) = \left\langle \overline{\omega}_i, dL_{x^{-1}}[X_j^L(x)] \right\rangle = \langle \overline{\omega}_i, X_j \rangle = \delta_{ij}.$$

For any covector $\nu \in T_x^*\mathbb{G}$, we denote by $\nu_1, \dots, \nu_n$ its components with respect to this frame, so that $\nu = \sum_{i=1}^n \nu_i \omega_i(x)$ and $\nu_i = \left\langle \nu, X_i^L(x) \right\rangle$. These coefficients ν_i should not be confused with the components of ν in exponential coordinates.

THEOREM 3.36. Given a biextremal (γ, λ) and $\overline{\nu} \in \mathbb{R}$ such that

$$(\lambda(0), \overline{\nu}) \in \left(\operatorname{im} d\,\widehat{\mathrm{extEnd}}_{\overline{u}}\right)^{\perp} \setminus \{0\},$$

we have

$$\lambda_i(t) + \overline{\nu} u_i(t) = 0 \tag{3.4}$$

for a.e. t and all $i = 1, \dots, r$. Moreover, the dual curve λ solves the system

$$\dot{\lambda}_i(t) = -\sum_{j=1}^r \sum_{k=1}^n c_{ij}^k u_j(t) \lambda_k(t) \tag{3.5}$$

for a.e. t and all $i = 1, \dots, n$. Conversely, any couple of curves $(\gamma, \lambda) \in H^1([0,1], T^*\mathbb{G})$ solving these two systems is a biextremal if $(\lambda(0), \overline{\nu}) \neq 0$; here we write that the couple takes values in $T^*\mathbb{G}$ to mean that γ is the projection on $\mathbb{G}$ of $\lambda \in H^1([0,1], T^*\mathbb{G})$.

Proof. In this proof we will write X_i instead of X_i^L for simplicity. Substitution of $\lambda(t) = \sum_{j=1}^n \lambda_j(t) \omega_j(\gamma(t))$ in (2.2) gives

$$0 = \langle \lambda(t), X_i(\gamma(t)) \rangle + \overline{\nu} u_i(t) = \sum_{j=1}^n \delta_{ij} \lambda_j(t) + \overline{\nu} u_i(t) = \lambda_i(t) + \overline{\nu} u_i(t),$$

for any $i = 1, \dots, r$. So (γ, λ) solves the system (2.2) (for some $\overline{\nu} \in \mathbb{R}$) iff it is a biextremal (see also Remark 2.53), provided λ is the dual curve associated to $\lambda(0)$ and the nontriviality condition $(\lambda(0), \overline{\nu}) \neq 0$ holds.

Now assume that λ lifts γ. Let us recall the identity

$$[X_i, X_j] = \sum_{k=1}^n c_{ij}^k X_k,$$

which holds at e by definition and thus on all of $\mathbb{G}$, as well, by left-invariance. Thinking λ as a covector in $(\mathbb{R}^n)^*$ in exponential coordinates, (2.3) is equivalent to asking

$$\left\langle \dot{\lambda}(t), X_i(\gamma(t)) \right\rangle = -\sum_{j=1}^r u_j(t) \lambda(t) dX_j[X_i](\gamma(t))$$

for all $i = 1, \dots, n$ and a.e. t. But

$$\left\langle \dot{\lambda}(t), X_i(\gamma(t)) \right\rangle = \frac{d}{dt} \langle \lambda(t), X_i(\gamma(t)) \rangle - \left\langle \lambda(t), \frac{d}{dt} X_i(\gamma(t)) \right\rangle = \dot{\lambda}_i(t) - \lambda(t) dX_i(\gamma(t))[\dot{\gamma}(t)]$$

(λ_i should not be confused with the i-th component of λ in coordinates), so, recalling that $\dot\gamma(t) = \sum_{j=1}^r u_j(t)X_j(\gamma(t))$, the above system is equivalent to

$$\begin{aligned}\dot\lambda_i(t) &= -\sum_{j=1}^r u_j(t)\lambda(t)dX_j[X_i](\gamma(t)) + \sum_{j=1}^r u_j(t)\lambda(t)dX_i[X_j](\gamma(t))\\ &= -\sum_{j=1}^r u_j(t)\left\langle \lambda(t), [X_i, X_j](\gamma(t))\right\rangle\\ &= -\sum_{j=1}^r\sum_{k=1}^n c_{ij}^k u_j(t)\lambda_k(t).\end{aligned}$$

□

3.6. A concrete example: the Heisenberg group

We now show how, using the equations derived in the previous section, one can compute explicitly the length minimizers in the case of the Heisenberg group $\mathbb{H}$, which is perhaps the simplest example of a sub-Riemannian manifold which is not Riemannian.

Definition 3.37. The *Heisenberg group* is the Lie group $\mathbb{H} := \left\{ \begin{pmatrix} 1 & * & * \\ 0 & 1 & * \\ 0 & 0 & 1 \end{pmatrix} \right\} < GL(3, \mathbb{R})$.

Its Lie algebra $\mathfrak{g}$ can be identified with $\left\{ \begin{pmatrix} 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & 0 \end{pmatrix} \right\}$, which is a Lie subalgebra of $\mathfrak{gl}(3, \mathbb{R})$. Notice that, letting

$$X_1 := e_{12},\ X_2 := e_{23},\ X_3 := e_{13},$$

we have $X_3 = [X_1, X_2]$. $\mathbb{H}$ is a stratified group, with the stratification

$$\mathfrak{g} = V_1 \oplus V_2, \quad V_1 := \langle X_1, X_2 \rangle,\ V_2 := \langle X_3 \rangle.$$

$\mathbb{H}$ becomes a Carnot group, with $r = s = 2$, by choosing the inner product on V_1 such that X_1, X_2 is an orthonormal basis. Identifying $\mathbb{H}$ with $\mathbb{R}^3$ using the exponential coordinates, the left-invariant vector fields take the well-known expressions

$$X_1^L(x) = \partial_1 - \frac{x_2}{2}\partial_3,\ X_2^L(x) = \partial_2 + \frac{x_1}{2}\partial_3,\ X_3^L(x) = \partial_3,$$

which can be obtained by the same method used to prove Proposition 3.18.

From the proof of Corollary 4.13, which can be carried out without localizing, or simply from Theorem 4.14 (which shows that abnormal biextremals do not exist in $\mathbb{H}$) we know that all length minimizers are normal extremals. Given a unit-speed normal biextremal (γ, λ) defined on $[0, T]$, with $\gamma(0) = e$, equations (3.4) and (3.5) take this simple form (notice that $c_{12}^3 = -c_{21}^3 = 1$, while all the other structure constants vanish):

$$\begin{cases}\lambda_1 = -u_1\\ \lambda_2 = -u_2,\end{cases} \qquad \begin{cases}\dot\lambda_1 = -u_2\lambda_3\\ \dot\lambda_2 = u_1\lambda_3\\ \dot\lambda_3 = 0.\end{cases}$$

Thus, eliminating λ_1 and λ_2 and calling υ the constant value of λ_3, we obtain

$$\begin{cases}\dot u_1 = \upsilon u_2\\ \dot u_2 = -\upsilon u_1.\end{cases}$$

If $\upsilon = 0$ we have $\gamma(t) = \exp(u_1 X_1 + u_2 X_2)$, which is a length minimizer for any $T > 0$ by Proposition 3.21. Assume now that $\upsilon \neq 0$. Taking into account that $u_1^2 + u_2^2 = 1$, we arrive at

$$\begin{cases} u_1(t) = \alpha \sin(\upsilon t) + \beta \cos(\upsilon t) \\ u_2(t) = -\beta \sin(\upsilon t) + \alpha \cos(\upsilon t) \end{cases}$$

for some α, β such that $\alpha^2 + \beta^2 = 1$. Finally, using

$$\dot{\gamma}(t) = u_1(t) X_1^L(\gamma(t)) + u_2(t) X_2^L(\gamma(t)) = u_1(t)\partial_1 + u_2(t)\partial_2 + \frac{u_2(t)\gamma_1(t) - u_1(t)\gamma_2(t)}{2}\partial_3$$

(in exponential coordinates, $\gamma_i(t)$ denoting the i-th coordinate of $\gamma(t)$), we can obtain first γ_1 and γ_2, by taking the antiderivatives of u_1 and u_2 vanishing at 0, and then we can compute γ_3. We arrive at

$$\gamma(t) = \begin{pmatrix} \frac{\alpha(1-\cos(\upsilon t)) + \beta \sin(\upsilon t)}{\upsilon} \\ \frac{-\beta(1-\cos(\upsilon t)) + \alpha \sin(\upsilon t)}{\upsilon} \\ -\frac{\upsilon t - \sin(\upsilon t)}{2\upsilon^2} \end{pmatrix}. \tag{3.6}$$

Conversely, if $\gamma : [0, T] \to \mathbb{H}$ has this form, then γ is a normal extremal. Notice that γ touches the x_3-axis only when $t = \frac{2k\pi}{|\upsilon|}$, for some integer $k \geq 0$. Fix any $\eta \neq 0$: the unit-speed length minimizers $\gamma : [0, T] \to \mathbb{H}$ connecting 0 to $(0, 0, \eta)$ can be found by looking for the minimal $T = \frac{2k\pi}{|\upsilon|}$ (which equals $L(\gamma)$) such that $\gamma(T) = (0, 0, \eta)$, for some γ as above. From

$$T = \frac{2k\pi}{|\upsilon|}, \quad \eta = \gamma(T) = -\frac{k\pi}{\upsilon^2} \operatorname{sgn}(\upsilon)$$

we deduce that υ has sign opposite to η and $|\upsilon| = \sqrt{\frac{k\pi}{|\eta|}}$, which gives $T = 2\sqrt{k\pi\,|\eta|}$. This attains its minimum value when $k = 1$ (we cannot have $k = 0$ since $\eta \neq 0$). So the unit-speed geodesics from 0 to $(0, 0, \eta)$ are the curves $\gamma : \left[0, 2\sqrt{\pi\,|\eta|}\right] \to \mathbb{H}$ given by the above formula, with $\upsilon := -\operatorname{sgn}(\eta)\sqrt{\frac{\pi}{|\eta|}}$. Notice that $(\alpha, \beta) \in S^1$ is arbitrary, so we have a one-parameter family of geodesics.

Finally, we show that, for any $x = (x_1, x_2, x_3) \in \mathbb{H}$ with $(x_1, x_2) \neq (0, 0)$, there exists a unique unit-speed length minimizer from 0 to x. This is clear if $x_3 = 0$ since, in this case, any normal extremal from 0 to x must have $\upsilon = 0$. Otherwise, if $x_3 \neq 0$, it suffices to show that $x = \gamma(t)$ for a unique $\gamma : \left[0, \frac{2\pi}{|\upsilon|}\right] \to \mathbb{H}$ given by (3.6) and a unique $t \in (0, 2\sqrt{\pi\,|\eta|})$. In order to find γ, it is convenient to work with complex numbers. Notice that $\gamma_1(t) = x_1$ and $\gamma_2(t) = x_2$ iff

$$x_1 + i x_2 = \frac{(\alpha - i\beta)(1 - e^{-\upsilon t i})}{\upsilon},$$

since the right-hand side equals $\gamma_1(t) + i\gamma_2(t)$. This equation gives

$$\upsilon^{-2} = \frac{|x_1 + i x_2|^2}{|1 - e^{-\upsilon t i}|^2} = \frac{x_1^2 + x_2^2}{2(1 - \cos(\upsilon t))}.$$

Setting $s := |\upsilon|\, t$ (which now varies in $(0, 2\pi)$) and substituting υ^{-2} into

$$x_3 = -\frac{\upsilon t - \sin(\upsilon t)}{2\upsilon^2} = -\operatorname{sgn}(\upsilon)\frac{s - \sin s}{2\upsilon^2},$$

we arrive at the equation

$$x_3 = -\operatorname{sgn}(\upsilon)\frac{x_1^2 + x_2^2}{4} \cdot \frac{s - \sin s}{1 - \cos s}.$$

Now $s \mapsto \frac{s-\sin s}{1-\cos s}$ is a diffeomorphism from $(0, 2\pi)$ to $(0, +\infty)$: indeed, its derivative equals $1-\frac{(s-\sin s)\sin s}{(1-\cos s)^2}$, which is positive iff $2-2\cos s > s\sin s$, i.e. iff $4\sin^2\left(\frac{s}{2}\right) > 2s\sin\left(\frac{s}{2}\right)\cos\left(\frac{s}{2}\right)$, which is trivial if $s \in [\pi, 2\pi)$ and follows from $\tan\left(\frac{s}{2}\right) > \frac{s}{2}$ if $s \in (0, \pi)$. Thus, the preceding equation uniquely determines s and $\operatorname{sgn}(v)$. Finally, $|v| = \sqrt{\frac{2(1-\cos s)}{x_1^2+x_2^2}}$, $t = \frac{s}{|v|}$ and $\alpha - i\beta = \frac{v(x_1+ix_2)}{1-e^{-vti}}$ are uniquely determined, as well (so that the same is true for γ).

3.7. Extremal polynomials

Recently, in [LLMV13], Le Donne, Leonardi, Monti and Vittone obtained a way to integrate (3.5) and showed that, for a fixed biextremal (γ, λ), the components $\lambda_i(t) := \left\langle \lambda(t), X_i^L(\gamma(t)) \right\rangle$ of $\lambda(t)$ depend only on $\gamma(t)$, by means of suitable *extremal polynomials*, whose definition depends only on $\lambda(0)$. As we will see below, their degree properties imply the smoothness of length minimizers in Carnot groups when $s \leq 3$.

In this section we will identify $\mathbb{G}$ with $\mathbb{R}^n$ by means of the *exponential coordinates of the second kind*, which means that we will implicitly use the map

$$F : \mathbb{R}^n \to \mathbb{G}, \quad (x_1, \dots, x_n) \mapsto \exp(x_n X_n) \cdots \exp(x_1 X_1),$$

as well as the following fact.

PROPOSITION 3.38. $F : \mathbb{R}^n \to \mathbb{G}$ is a diffeomorphism.

Proof. F is injective: assume by contradiction that $F(x) = F(y)$ for some $x \neq y$ and let i be the smallest index such that $x_i \neq y_i$. From $F(x) = F(y)$ we deduce

$$\exp(x_n X_n) \cdots \exp(x_i X_i) = \exp(y_n X_n) \cdots \exp(y_i X_i).$$

Letting $j := d(i)$ and applying the homomorphism $\pi_j : \mathbb{G}_j \to V_j$ (see Lemma 3.24), we get

$$x_i X_i + \cdots + x_{r_j} X_{r_j} = y_i X_i + \cdots + y_{r_j} X_{r_j},$$

which implies $x_i = y_i$ by linear independence, contradiction. F is surjective, as well: fix $x \in \mathbb{G}$ and set $x_{(1)} := x$. Let $\pi_1(x_{(1)}) = \alpha_1 X_1 + \cdots + \alpha_r X_r$ and set

$$x_{(2)} := x_{(1)} \exp(-\alpha_1 X_1) \cdots \exp(-\alpha_r X_r) \in \mathbb{G}_2$$

(the fact that $x_{(2)} \in \mathbb{G}_2$ follows from $\pi_1(x_{(2)}) = 0$). Let $\pi_2(x_{(2)}) = \alpha_{r_1+1} X_{r_1+1} + \cdots + \alpha_{r_2} X_{r_2}$ and

$$x_{(3)} := x_{(2)} \exp(-\alpha_{r_1+1} X_{r_1+1}) \cdots \exp(-\alpha_{r_2} X_{r_2}) \in \mathbb{G}_3$$

(again, the fact that $x_{(3)} \in \mathbb{G}_3$ follows from $x_{(3)} \in \mathbb{G}_2$, together with $\pi_2(x_{(3)}) = 0$). Iterating, after s steps we arrive at $x_{(s+1)} = e$. Thus,

$$\begin{aligned} x = x_{(1)} &= x_{(2)} \exp(\alpha_r X_r) \cdots \exp(\alpha_1 X_1) = x_{(3)} \exp(\alpha_{r_2} X_{r_2}) \cdots \exp(\alpha_1 X_1) \\ &= \cdots = \exp(\alpha_n X_n) \cdots \exp(\alpha_1 X_1). \end{aligned}$$

This argument, which shows the surjectivity of F, in fact exhibits a smooth right inverse $H : \mathbb{G} \to \mathbb{R}^n$, i.e. a map H such that $F \circ H = \mathrm{id}_{\mathbb{G}}$. Since F is bijective, we have $H = F^{-1}$, proving that F is a diffeomorphism. □

DEFINITION 3.39. For any $v \in \mathbb{R}^n$ and any $i = 1, \dots, n$, the *extremal polynomial* $P_i^v \in \mathbb{R}[x_1, \dots, x_n]$ is given by the finite sum

$$P_i^v(x) := \sum_{\alpha \in \mathbb{N}^n} \sum_{k=1}^n \frac{1}{\alpha!} c_{\alpha i}^k v_k x^\alpha$$

(see the remark below), where $\alpha! := \alpha_1! \cdots \alpha_n!$, $x^\alpha := x_1^{\alpha_1} \cdots x_n^{\alpha_n}$ and the *generalized structure constants* $c^k_{\alpha i}$ are defined by

$$\mathrm{ad}(X_n)^{\alpha_n} \cdots \mathrm{ad}(X_1)^{\alpha_1} X_i =: \sum_{k=1}^n c^k_{\alpha i} X_k.$$

These polynomials P_i^v can be viewed as functions on $\mathbb{G}$ by the above identification (i.e. P_i^v corresponds to the smooth function $P_i^v \circ F^{-1} : \mathbb{G} \to \mathbb{R}$).

REMARK 3.40. Since $c^k_{\alpha i} = 0$ when $\sum_{j=1}^n \alpha_j d(j) + d(i) > s$, the sum defining P_i^v is indeed finite and each nonzero monomial in that sum has weighted degree at most $s - d(i)$.

These special polynomials satisfy a remarkable identity, which is the core of the proof of Theorem 3.42.

THEOREM 3.41. For any $v \in \mathbb{R}^n$ and any $i, j = 1, \ldots, n$ we have $P_i^v(0) = v_i$ and

$$X_j^L P_i^v = \sum_{k=1}^n c^k_{ji} P_k^v \tag{3.7}$$

on all of $\mathbb{G}$.

We omit the proof, which is quite long and involves several clever manipulations, and refer the reader to the original paper [LLMV13].

THEOREM 3.42. Given a biextremal (γ, λ) defined on $[0, T]$ with $\gamma(0) = 0$, let $v_j := \langle \lambda(0), X_i \rangle$ and $v := (v_1, \ldots, v_n)$. For any t we have

$$\lambda_i(t) = P_i^v(\gamma(t)).$$

Proof. The thesis is clear when $i = n$: by (3.5) we have $\dot\lambda_n \equiv 0$, so that λ_n is constant; similarly, by (3.7), $X_j^L P_n^v \equiv 0$ on $\mathbb{G}$ (for any j), so that P_n^v is constant on $\mathbb{G}$ and we obtain

$$P_n^v(\gamma(t)) = P_n^v(0) = v_n = \lambda_n(0) = \lambda_n(t)$$

for any t. It now suffices to show the thesis for a fixed i, assuming that it holds for any $i' > i$. For a.e. t we have, using again (3.7) and (3.5),

$$\begin{aligned}
\frac{d}{dt}(P_i^v \circ \gamma)(t) &= \sum_{j=1}^r u_j(t)(X_j^L P_i^v)(\gamma(t)) \\
&= \sum_{j=1}^r \sum_{k=1}^n u_j(t) c^k_{ji} P_k^v(\gamma(t)) \\
&= -\sum_{j=1}^r \sum_{k=1}^n u_j(t) c^k_{ij} \lambda_k(t) \\
&= \dot\lambda_i(t);
\end{aligned}$$

in the second-to-last equality we used the fact that $c^k_{ji} P_k^v(\gamma(t)) = -c^k_{ij}\lambda_k(t)$, which is trivial when $k \leq i$ (since then $c^k_{ij} = 0$) and follows from the inductive hypothesis when $k > i$. Since

$$P_i^v \circ \gamma(0) = P_i^v(0) = v_i = \lambda_i(0)$$

as well, we deduce $P_i^v \circ \gamma(t) = \lambda_i(t)$ for any t. $\square$

Using a result which will be proved later in the thesis (namely, the *Goh conditions*), we can now deduce the smoothness of constant-speed geodesics when $s \leq 3$, which was also proved by Tan and Yang in [TY13] with different techniques.

THEOREM 3.43. Assume that $s \leq 3$ and let $\gamma : [0, T] \to \mathbb{G}$ be a constant-speed length minimizer. Then γ is smooth, i.e. $\gamma \in C^\infty([0, T], \mathbb{G})$.

Proof. Step 1. Let $\gamma : [0, T] \to \mathbb{G}$ be a constant-speed length minimizer and assume by contradiction that γ is not smooth. By left translation, we can assume $\gamma(0) = e$ and, rescaling the speed, that $T = 1$. Since γ is not smooth, it has to be a strictly abnormal extremal (see Theorem 2.55). Such γ cannot exist if $s = 1$ or $s = 2$ (see Corollary 2.51 and the proof of Corollary 4.13), so we must have $s = 3$. We can assume that r is the minimal rank such that such a counterexample exists.

By Theorem 4.12 there exists a dual curve $\lambda : [0, 1] \to T^*\mathbb{G}$ satisfying

$$\left\langle \lambda(t), [X_i^L, X_j^L](\gamma(t)) \right\rangle = 0$$

for any t and any $i, j = 1, \ldots, r$. Since (γ, λ) is an abnormal biextremal, we also have $\left\langle \lambda(t), X_i^L(\gamma(t)) \right\rangle = 0$. Recalling that $V_2 = [V_1, V_1]$ and that $[X_i^L, X_j^L]$ is the left-invariant vector field associated to $[X_i, X_j]$, we deduce

$$\lambda_i(t) = 0, \quad \forall i = 1, \ldots, r_2, \ \forall t \in [0, 1].$$

Step 2. Let $v_i := \lambda_i(0)$ and $v := (v_1, \ldots, v_n) \neq 0$. We now prove that there exists some j with $d(j) = 2$ and such that $P_j^v \not\equiv 0$. If this is not the case, given any m with $d(m) = 3$ we can write (using $V_3 = [V_1, V_2]$)

$$X_m = \sum_{\substack{i:d(i)=1 \\ j:d(j)=2}} \alpha_{ij} [X_i, X_j]$$

for suitable $\alpha_{ij} \in \mathbb{R}$, which implies $\sum \alpha_{ij} c_{ij}^k = \delta_{km}$ (the sum being over i, j as in the displayed equation). Thus, using (3.7),

$$0 \equiv \sum_{\substack{i:d(i)=1 \\ j:d(j)=2}} \alpha_{ij} X_i^L P_j^v = \sum_{\substack{i:d(i)=1 \\ j:d(j)=2}} \sum_{k=1}^n \alpha_{ij} c_{ij}^k P_k^v = \sum_{k=1}^n \delta_{km} P_k^v = P_m^v,$$

which gives in particular $v_m = P_m^v(0) = 0$. Combining this with $v_i = \lambda_i(0) = 0$ for $i \leq r_2$, we obtain $v = 0$, which is absurd.

Step 3. Thus, there exists some j with $d(j) = 2$ and $P_j^v \not\equiv 0$. Since the weighted degree of any monomial in P_j^v is at most 1 (see Remark 3.40) and $P_j^v(0) = v_j = 0$, P_j^v has the form

$$P_j^v(x) = \sum_{i=1}^r \beta_i x_i.$$

Let us define $\omega \in V_1^*$ by $\langle \omega, \sum_{i=1}^r x_i X_i \rangle := \sum_{i=1}^r \beta_i x_i$. Notice that, by Theorem 3.42, $P_j^v(\gamma(t)) = 0$ for any t. But on $\mathbb{G}$ we have $P_j^v = \omega \circ \pi$ (since $\pi \circ F(x) = \sum_{i=1}^r x_i X_i$, as π is a homomorphism by Lemma 3.22). To reach a contradiction, it suffices to show that γ is still a length minimizer in a Carnot subgroup $\mathbb{G}' \subsetneq \mathbb{G}$ with smaller rank ($\mathbb{G}'$ could be even smaller than the subgroup $\pi^{-1}(\{X : \langle \omega, X \rangle = 0\})$, which is not necessarily a stratified subgroup).

Step 4. Let $Y_1, \dots, Y_r$ be an orthonormal basis of V_1 satisfying $\langle \omega, Y_i \rangle = 0$ for $i \le r-1$ and $\langle \omega, Y_r \rangle \neq 0$. We define inductively

$$V_1' := \langle Y_1, \dots, Y_{r-1} \rangle, \quad V_{i+1}' := [V_1', V_i']$$

(for $i = 1, \dots, s-1$) and we set $\mathfrak{g}' := V_1' \oplus \cdots \oplus V_s'$. We obtain $[V_i', V_j'] \subseteq V_{i+j}'$ (with the convention that $V_k' := \{0\}$ if $k > s$) by using Jacobi's identity repeatedly, exactly as in the proof of Proposition 3.3. Thus, $\mathfrak{g}'$ is the Lie subalgebra generated by $Y_1, \dots, Y_{r-1}$ (and is a graded Lie subalgebra of $\mathfrak{g}$).

Define $\mathbb{G}' := \exp(\mathfrak{g}')$, which is a simply connected Lie subgroup whose Lie algebra is $\mathfrak{g}'$. By the preceding discussion, $\mathbb{G}'$ is a stratified group; it becomes a Carnot group with the metric induced by $\overline{g}|_{V_1'}$ (see Definition 3.4). Moreover, $\mathbb{G}'$ has rank $r-1$.

Step 5. We are left to prove that $\gamma([0,1]) \subseteq \mathbb{G}'$: once this is done, γ will be a length minimizer in $\mathbb{G}'$, as well (since, for any horizontal curve δ in $\mathbb{G}'$, its length in $\mathbb{G}'$ equals its length in $\mathbb{G}$), contradicting the minimality of r. Calling $v = (v_1, \dots, v_r)$ the control associated to γ with respect to $Y_1^L, \dots, Y_r^L$, from $\langle \omega, \pi(\gamma(t)) \rangle \equiv 0$ we deduce

$$0 = \frac{d}{dt} \langle \omega, \pi(\gamma(t)) \rangle = \left\langle \omega, d\pi_{\gamma(t)}[\dot{\gamma}(t)] \right\rangle = \sum_{i=1}^{r} v_i(t) \langle \omega, Y_i \rangle = v_r(t) \langle \omega, Y_r \rangle$$

for a.e. t, thanks to the fact that $d\pi_{\gamma(t)}[Y_i^L(\gamma(t))] = Y_i$ (by Lemma 3.25). Thus $v_r \equiv 0$, so the horizontal curve γ' in $\mathbb{G}'$ associated to the control $v' := (v_1, \dots, v_{r-1})$ (with respect to the frame of left-invariant vector fields $Y_1^L, \dots, Y_{r-1}^L$ in $\mathbb{G}'$) satisfies

$$\dot{\gamma}'(t) = \sum_{i=1}^{r} v_i(t) X_i^L(\gamma'(t))$$

in $\mathbb{G}$, as well, for a.e. t. Hence, $\gamma' \equiv \gamma$, proving that $\gamma([0,1]) \subseteq \mathbb{G}'$. $\square$

CHAPTER 4

Second order theory

In this chapter we develop a second order theory, with the aim of providing necessary and sufficient conditions for the minimality of a given strictly abnormal extremal. The necessary conditions can be used to obtain, in the special case of sub-Riemannian manifolds with step 2, the absence of strictly abnormal minimizers and thus the smoothness of all minimizers. On the contrary, the sufficient conditions will enable us to give explicit examples of strictly abnormal minimizers, even in Carnot groups.

In the first part we look for second order necessary conditions for a horizontal curve to be a minimizer. Before delving into the precise definitions and proofs, we sketch the structure of the argument. Let us assume that $\gamma : [0,1] \to M$ minimizes the energy among all horizontal curves (in $H^1([0,1], M)$) joining $\gamma(0)$ and $\gamma(1)$ and that γ is a strictly abnormal extremal. Let $u \in \mathcal{U}$ be the control associated to γ, so that $\gamma(1) = \mathrm{End}(\gamma(0), u)$ (we will omit the dependence on $\gamma(0)$ in the sequel). We have already observed (see Section 2.7) that the extended endpoint map $\mathrm{extEnd} : \mathcal{U} \to M \times \mathbb{R}$ cannot be locally open at u in such case.

As we will see soon, this implies that $\widehat{\mathrm{extEnd}}$ has negative index at u, the index being a generalization of Morse index to maps whose codomain is a manifold, and as a corollary we will obtain that $\widehat{\mathrm{End}}$ has negative index at u as well. Then we will compute the Hessian of $\widehat{\mathrm{End}}$ and, by a blow-up technique, we will deduce that γ admits a dual curve satisfying the so-called *Goh conditions*.

4.1. Hessian and index of maps between manifolds

In order to motivate the definition of the Hessian, consider two finite-dimensional differentiable manifolds N, M and a smooth (or just C^2) map $F : N \to M$. Recall that, given $v \in T_pN$, the differential $dF_p[v] \in T_{F(p)}M$ can be defined as $\frac{d}{dt}(F \circ \sigma)\big|_{t=0}$, where $\sigma : (-\epsilon, \epsilon) \to N$ is any smooth curve such that $\sigma(0) = p$ and $\dot\sigma(0) = v$. We want to define the Hessian of F at p in an intrinsic way, as a bilinear map from (a suitable subspace of) T_pN to (a suitable quotient of) $T_{F(p)}M$.

Let us identify a neighbourhood V of $F(p)$ with $\mathbb{R}^n$ by using some local coordinates. One could then naively define $\mathrm{Hess}\, F_p[v,v] := \frac{d^2}{dt^2}(F \circ \sigma)\Big|_{t=0}$. Let us now see how this formula behaves when we perform a change on coordinates on M, i.e. when we look at $\Phi \circ F$ instead of F, where $\Phi : V \to V'$ is a diffeomorphism. Starting from

$$\frac{d}{dt}(\Phi \circ F \circ \sigma) = d(\Phi \circ F)[\dot\sigma] = d\Phi\,[dF[\dot\sigma]]$$

we get

$$\frac{d^2}{dt^2}(\Phi \circ F \circ \sigma) = d\Phi\left[d^2F[\dot\sigma, \dot\sigma] + dF[\ddot\sigma]\right] + d^2\Phi\,[dF[\dot\sigma], dF[\dot\sigma]]\,,$$

while (using this formula with $\Phi := \mathrm{id}$)

$$\frac{d^2}{dt^2}(F \circ \sigma) = d^2F[\dot\sigma, \dot\sigma] + dF[\ddot\sigma].$$

To guarantee that the formula $\frac{d^2}{dt^2}(F \circ \sigma)\Big|_{t=0}$ gives a well-defined vector in (a quotient of) $T_{F(p)}M$ we thus require that $v \in \ker dF_p$ (so that $d^2\Phi\,[dF[\dot\sigma], dF[\dot\sigma]] = 0$) and we project $\frac{d^2}{dt^2}(F \circ \sigma)\Big|_{t=0}$ into the quotient $T_{F(p)}M/\mathrm{im}\, dF_p = \mathrm{coker}\, dF_p$ (so that we can forget the term $dF[\ddot\sigma]$).

Definition 4.1. The *Hessian* of F at p is the symmetric bilinear map $\mathrm{Hess}\, F_p : \ker dF_p \times \ker dF_p \to \mathrm{coker}\, dF_p$ associated to the quadratic form

$$v \mapsto \frac{d^2}{dt^2}(F \circ \sigma)\Big|_{t=0} \mod \mathrm{im}\, dF_p.$$

Remark 4.2. This definition still makes sense when N is an open subset of a Banach space, or more generally a Banach manifold, and $F \in C^2(N, M)$. This is the case for both the endpoint map $\mathrm{End} : \mathcal{U} \to M$ and the extended endpoint map $\mathrm{extEnd} : \mathcal{U} \to M \times \mathbb{R}$.

Remark 4.3. In the case of $\mathrm{End} : \mathcal{U} \to M$ or $\mathrm{extEnd} : \mathcal{U} \to M \times \mathbb{R}$ (and $p := u$), $\mathcal{U}$ is an open subset of $L^2([0, 1], \mathbb{R}^r)$ and one could use the vector space structure to choose a canonical σ, namely $\sigma(t) := u + tv$. In this case $\ddot\sigma = 0$, so one would not need to quotient by $\mathrm{im}\, dF_p$ in order to define the Hessian. We will not use this remark, though.

In the sequel N will always denote a finite dimensional differentiable manifold or an open subset of a Banach space, while M will always be a finite dimensional differentiable manifold.

Let us now define the index of a bilinear map.

Definition 4.4. Given a symmetric bilinear form $Q : X \times X \to \mathbb{R}$, where X is a real vector space (possibly infinite dimensional), we set $Q(v) := Q(v, v)$. If $Y \subseteq X$ is a subspace, the notation $Q|_Y > 0$ means that Q is positive definite on Y, and similarly for $Q|_Y < 0$. The *index* of Q is

$$\mathrm{ind}\, Q := \sup_{Y : Q|_Y < 0} \dim Y.$$

Definition 4.5. Given a map $F \in C^2(N, M)$, the *index* of F at $p \in N$ is

$$\mathrm{ind}_p F := \inf_{\lambda \in (\mathrm{im}\, dF_p)^\perp \setminus \{0\}} \mathrm{ind}\, \lambda\, \mathrm{Hess}\, F_p - \dim \mathrm{coker}\, dF_p.$$

Here λ varies over the nonzero covectors in $(T_{F(p)})^*$ vanishing on $\mathrm{im}\, dF_p$, so that $\lambda\, \mathrm{Hess}\, F_p$ is well-defined on $\ker dF_p$. Notice that such covectors can be identified with elements of $(\mathrm{coker}\, dF_p)^*$.

Remark 4.6. If $Q : X \times X \to \mathbb{R}$ is a continuous symmetric bilinear form (X being a real vector space), the associated quadratic form (which we still denote by Q) satisfies $dQ_x = 2Q(x, \cdot)$, $d^2Q_x = 2Q(\cdot, \cdot)$ and $\mathrm{ind}_0\, Q = \mathrm{ind}\, Q$.

4.2. A sufficient second order condition for local openness

The main result of this section is that $\operatorname{ind}_p F \geq 0$, i.e. $\operatorname{ind} \lambda \operatorname{Hess} F_p \geq \dim \operatorname{coker} dF_p$ for all nonzero $\lambda \in (\operatorname{im} dF_p)^\perp$, implies local openness at p.

Remark 4.7. Notice that if dF_p is surjective this condition tells us nothing, but in this case local openness is guaranteed by the implicit function theorem.

We start with a special case to illustrate the main ideas.

Theorem 4.8. Let N, M be finite dimensional manifolds and $F \in C^2(N, M)$. Assume $\operatorname{ind}_p F \geq 0$ and $\dim \operatorname{coker} dF_p = 1$. Then F is locally open at p (i.e. $F(p) \in \operatorname{int} F(U)$ for any neighbourhood U of p).

Proof. Working in local coordinates (centered at p and $F(p)$) we can assume $N = \mathbb{R}^n$, $M = \mathbb{R}^m$, $p = 0$ and $F(p) = 0$. The hypotheses tell us that for any $\lambda \in (\operatorname{im} dF_0)^\perp \setminus \{0\}$ we have $\operatorname{ind} \lambda \operatorname{Hess} F_0 \geq 1$. Fix any such λ (which is in fact uniquely determined up to nonzero factors) and put $Q := \lambda \operatorname{Hess} F_0$. The last inequality, applied to λ and $-\lambda$, says that the quadratic form $v \mapsto Q(v, v)$ takes on both positive and negative values. So there exist nonzero $v, w \in \ker dF_0$ such that

$$Q(v, v) = 0, \quad Q(v, w) \neq 0.$$

Indeed, we can find x, y such that $Q(x, x) > 0$ and $Q(y, y) < 0$. Q has to be nondegenerate on the two-dimensional subspace $\langle x, y\rangle$ and there exists $v \in \langle x, y\rangle \setminus \{0\}$ such that $Q(v, v) = 0$. Now for any $w \in \langle x, y\rangle \setminus \langle v\rangle$ we have $Q(v, w) \neq 0$. The fact that $Q(v, v) = 0$ gives $d^2F_0[v, v] \in \operatorname{im} dF_0$, so there exists $z \in T_0N$ such that

$$dF_0[z] = -\frac{1}{2} d^2F_0[v, v].$$

Now let us fix any splitting $\mathbb{R}^n = E \oplus \ker dF_0$ and, for any $\epsilon > 0$, consider the map

$$\Phi_\epsilon : E \times \mathbb{R} \to \mathbb{R}^m, \quad \Phi_\epsilon(x, t) := F\left(\epsilon^2(z + x) + \epsilon(v + tw)\right).$$

Since F is C^2 we can write $F(u) = dF_0[u] + \frac{1}{2}d^2F_0[u, u] + o(|u|^2)$, so

$$\begin{aligned}\Phi_\epsilon(x, t) &= \epsilon^2 dF_0[z + x] + \frac{1}{2}\epsilon^2 d^2F_0[v + tw, v + tw] + o(\epsilon^2)\\ &= \epsilon^2(dF_0[x] + td^2F_0[v, w]) + \frac{1}{2}\epsilon^2 t^2 d^2F_0[w, w] + o(\epsilon^2),\end{aligned}$$

where the error $o(\epsilon^2)$ is uniform as (x, t) vary over an arbitrary compact set. The map

$$(x, t) \mapsto dF_0[x] + td^2F_0[v, w]$$

is linear and bijective, so for t and ϵ small enough, say $|t| \leq \delta$, we can apply Lemma A.1 in Appendix A to $\frac{\Phi_\epsilon}{\epsilon^2}\big|_{\overline{B}_\delta}$, which gives $0 \in \operatorname{int} \Phi_\epsilon(B_1)$ (here $\overline{B}_\delta$ is the Euclidean ball in the product $E \times \mathbb{R}$). Since $\Phi_\epsilon(B_1) \subseteq F(B_{C\epsilon})$ for some positive constant C, the thesis follows. □

We now move to the general case. To attack it we need two technical lemmas.

Lemma 4.9. Let $F \in C^2(\mathbb{R}^n, \mathbb{R}^m)$ with $F(0) = 0$ and assume that the quadratic form $Q := \operatorname{Hess} F_0 : \ker dF_0 \to \operatorname{coker} dF_0$ has a regular zero (i.e. a zero where the differential is surjective). Then F has regular zeros which are arbitrarily close to 0 and F is locally open at 0.

Proof. Fix any splitting $\mathbb{R}^n = E \oplus \ker dF_0$ and pick a regular zero v of Q. Since $Q(v) = 0$ we know that $d^2F_0[v,v] \in \operatorname{im} dF_0$, thus we can write $dF_0[z] = -\frac{1}{2}d^2F_0[v,v]$. Now we basically repeat the final part of the proof of Lemma 4.8: we define

$$\Phi_\epsilon : E \times \ker dF_0 \to \mathbb{R}^m, \quad \Phi_\epsilon(x,y) := F\left(\epsilon^2(z+x) + \epsilon(v+y)\right)$$

and remark that

$$\Phi_\epsilon(x,y) = \epsilon^2\left(dF_0[x] + d^2F_0[v,y]\right) + \frac{1}{2}\epsilon^2 d^2F_0[y,y] + o(\epsilon^2),$$

$$\begin{aligned} d\Phi_\epsilon(x,y)[a,b] =& dF\left(\epsilon^2(z+x) + \epsilon(v+y)\right)[\epsilon^2 a + \epsilon b] \\ =& dF_0[\epsilon^2 a + \epsilon b] + d^2F_0[\epsilon(v+y), \epsilon b] + o(\epsilon^2) \\ =& \epsilon^2\left(dF_0[a] + d^2F_0[v,b]\right) + \epsilon^2 d^2F_0[y,b] + o(\epsilon^2) \end{aligned}$$

(again the error is uniform as x, y, a, b vary in an arbitrary compact set). Therefore, since $(a,b) \mapsto dF_0[a] + d^2F_0[v,b]$ is surjective, the same is true for

$$(a,b) \mapsto dF\left(\epsilon^2(z+x) + \epsilon(v+y)\right)[\epsilon^2 a + \epsilon b],$$

provided y and ϵ are sufficiently small, say $y \in \overline{B}_\delta$, and $x \in \overline{B}_1$. Thus dF is surjective at $\epsilon^2(z+x) + \epsilon(v+y)$, as well. Finally, possibly shrinking δ and then ϵ as in the previous proof, we can apply Lemma A.1 to $\frac{\Phi_\epsilon}{\epsilon^2}\big|_{\overline{B}_\delta}$ to get that $\Phi_\epsilon(B_\delta)$ contains a neighbourhood of 0 (here B_δ denotes the Euclidean ball in the product $E \times \ker dF_0$). In particular $\Phi_\epsilon(x,y) = 0$ for some $(x,y) \in B_\delta \subseteq \overline{B}_1 \times \overline{B}_\delta$, which gives us the required regular zero $\epsilon^2(z+x) + \epsilon(v+y)$ of F. □

Lemma 4.10. Let $Q : \mathbb{R}^n \to \mathbb{R}^m$ be an $\mathbb{R}^m$-valued quadratic form (i.e. Q is componentwise a quadratic form) such that $\operatorname{ind} \lambda Q \geq m$ for any $\lambda \in (\mathbb{R}^m)^* \setminus \{0\}$. Then Q has a regular zero.

Proof. We can assume that the kernel of Q is trivial, i.e. $\{v : Q(v,\cdot) = 0\} = \{0\}$, since otherwise we can replace Q by the induced map

$$\overline{Q} : (\mathbb{R}^n/\ker Q) \times (\mathbb{R}^n/\ker Q) \to \mathbb{R}^m,$$

whose regular zeros correspond to regular zeros of Q. So, recalling that $dQ_v = 2Q(v,\cdot)$, we have $dQ_v \neq 0$ for any $v \neq 0$. Now we argue by strong induction on m, the thesis being clear for $m = 1$ (see the beginning of the proof of Lemma 4.8). Suppose we already know that $Q(v) = 0$ for some $v \neq 0$ and let $k := \dim \operatorname{im} dQ_v$. If $k = m$ we are done; otherwise we have $0 < k < m$ and, for any $\lambda \in (\mathbb{R}^m)^* \setminus \{0\}$,

$$\operatorname{ind} \lambda \operatorname{Hess} Q_v = \operatorname{ind} \lambda Q|_{\ker dQ_v} \geq m - k,$$

since by hypothesis $\operatorname{ind} \lambda Q \geq m$ and $\ker dQ_v$ has codimension k in $\mathbb{R}^n$. Therefore

$$\operatorname{Hess} Q_v : \ker dQ_v \to \operatorname{coker} dQ_v$$

satisfies the hypothesis with m replaced by $m - k$, so by induction it has a regular zero. Thus by Lemma 4.9 Q has a regular zero as well.
We now assume by contradiction that Q vanishes only at 0. First of all, we claim that $Q(\mathbb{R}^n) = \mathbb{R}^m$. In fact $Q(\mathbb{R}^n)$ is a closed cone: indeed, if $Q(x_n) \to y \neq 0$, then $|x_n|^2 \left|Q\left(\frac{x_n}{|x_n|}\right)\right| \to |y|$, so x_n is bounded because $|Q|$ on S^{n-1} is bounded below by a positive constant, hence it converges up to subsequences. Moreover, $Q(\mathbb{R}^n) \setminus \{0\}$ is open: repeating verbatim the preceding argument, for every $v \in \mathbb{R}^n \setminus \{0\}$ and every $\lambda \in (\mathbb{R}^m)^* \setminus \{0\}$ we have $\operatorname{ind} \lambda \operatorname{Hess} Q_v \geq m - \dim \operatorname{im} dQ_v$, which is smaller than m, so $\operatorname{Hess} Q_v$ has a regular zero by the inductive hypothesis, thus by Lemma 4.9 Q is locally open at v (in the trivial case $\dim \operatorname{im} dQ_v = m$ we use instead the implicit function theorem). Thus $Q(\mathbb{R}^n) \setminus \{0\}$ is

both open and closed in $\mathbb{R}^m \setminus \{0\}$, so as $m \geq 2$ it follows that $Q(\mathbb{R}^n) \setminus \{0\} = \mathbb{R}^m \setminus \{0\}$ and $Q(\mathbb{R}^n) = \mathbb{R}^m$.

Consider the map

$$R := \frac{Q}{|Q|} : S^{n-1} \to S^{m-1}$$

and a regular value x of R, which exists by Sard's theorem. Let $v_0 \in R^{-1}(x)$ be such that $|Q(v_0)|$ has the minimum possible value (the fact that $R^{-1}(x) \neq \emptyset$ follows from the surjectivity of Q) and set $a_0 := |Q(v_0)|$. Saying that v_0 minimizes $|Q|$ over $R^{-1}(x)$ and $|Q(v_0)| = a_0$ is equivalent to saying that (a_0, v_0) solves this minimization problem:

$$\text{minimize } a, \quad \text{among all } (a,v) \in (0,+\infty) \times S^{n-1} \text{ such that } Q(v) - ax = 0.$$

We remark that the map $H(a,v) := Q(v) - ax$ (whose domain is $(0,+\infty) \times S^{n-1}$) has surjective differential at (a_0, v_0): in fact, calling $x^\perp := \{y \in \mathbb{R}^m : \langle x, y\rangle = 0\}$ and $\pi_{x^\perp} : \mathbb{R}^m \to x^\perp$ the orthogonal projection, we have $x^\perp = T_x S^{m-1} = \operatorname{im} d\left(\frac{Q}{|Q|}\right)_{v_0} = \pi_{x^\perp}(\operatorname{im} dQ_{v_0})$ (we wrote Q instead of $Q|_{S^{n-1}}$ for simplicity), so $\operatorname{im} dH_{(a_0,v_0)} \supseteq \operatorname{im} dQ_{v_0} + \langle x\rangle = \pi_{x^\perp}(\operatorname{im} dQ_{v_0}) + \langle x \rangle = \mathbb{R}^m$. Therefore the map

$$(0,+\infty) \times \mathbb{R}^n \to \mathbb{R}, \quad (a,v) \mapsto \Big(Q(v) - ax, |v|^2 - 1\Big)$$

has surjective differential at (a_0, v_0), as well. The constraints of the above optimization problem can be encoded as the vanishing of this map, so we can apply Lemma A.3 to obtain that

$$L(a,v) = a + \lambda\left(Q(v) - ax\right) + \mu\left(|v|^2 - 1\right),$$

for some $\lambda \in (\mathbb{R}^m)^*$, $\mu \in \mathbb{R}$, $(\lambda, \mu) \neq 0$, satisfies the first and second order optimality conditions. The first order conditions are

$$\frac{\partial L}{\partial a}(a_0, v_0) = 0, \quad \frac{\partial L}{\partial v}(a_0, v_0) = 0,$$

which can be rewritten as

$$1 - \lambda x = 0, \quad \lambda dQ_{v_0} + 2\mu \langle v_0, \cdot\rangle = 0.$$

Inserting v_0 in the second one gives $\mu = -a_0 \lambda x = -a_0$. The second order conditions now give

$$d^2(\lambda Q)(v_0)[w,w] - 2a_0 |w|^2 \geq 0$$

for any $w \in W := \ker dQ_{v_0} \cap T_{v_0} S^{n-1}$, which surely lies in the kernel of the differential of each constraint (strictly speaking, we are applying the second order conditions to the vector $(0,w) \in \mathbb{R} \times \mathbb{R}^n = T_{(a_0,v_0)}((0,+\infty) \times \mathbb{R}^n)$). Since $a_0 > 0$ and $d^2\lambda Q(v_0) = 2\lambda Q$, we get $\lambda Q(w) \geq 0$, i.e. $\lambda Q|_W \geq 0$. But $\dim \operatorname{im} d\left(Q|_{S^{n-1}}\right)_{v_0} \geq \dim \pi_{x^\perp}\left(\operatorname{im} d\left(Q|_{S^{n-1}}\right)_{v_0}\right) = m-1$, so $\dim \operatorname{im} d\left(Q|_{S^{n-1}}\right)_{v_0} = m-1$ as v_0 is a critical point for $Q|_{S^{n-1}}$ (otherwise we would have $a_0(1-\epsilon)x \in Q(S^{n-1})$ for $\epsilon > 0$ small). Hence, $\dim W = (n-1) - (m-1) = n-m$. To reach a contradiction, notice that

$$\lambda Q|_{W \oplus \langle v_0\rangle} \geq 0$$

as well, since $\lambda Q(v_0, v_0) = \lambda(a_0 x) = a_0 > 0$ and $Q(v_0, w) = 0$ for any $w \in W$. So λQ cannot be negative definite on a subspace of $\mathbb{R}^n$ whose dimension is greater than $n - (n-m+1) = m-1$, contradicting the hypothesis $\operatorname{ind} \lambda Q \geq m$. □

THEOREM 4.11. If $F \in C^2(N,M)$ and $\operatorname{ind}_p F \geq 0$, then F is locally open at p.

Proof. Let $k := \dim \operatorname{coker} dF_p$. The hypothesis says that for every $\lambda \in (\operatorname{im} dF_p)^\perp \setminus \{0\}$ we have $\operatorname{ind} \lambda \operatorname{Hess} F_p \geq k$. Localizing we can assume that N is an open subset of a Banach space X and $M = \mathbb{R}^m$. We can assume $p = 0$ and $F(p) = 0$, as well. As a first step, we reduce to the case that X is finite-dimensional. Call S the unit sphere of $(\operatorname{im} dF_0)^\perp$.

For any $\lambda \in S$ there exists some $E_\lambda \subseteq \ker dF_0$ such that $\dim E_\lambda = k$ and $\lambda \operatorname{Hess} F_0|_{E_\lambda} < 0$. By finite-dimensionality of E_λ, we still have $\lambda' \operatorname{Hess} F_0|_{E_\lambda} < 0$ when λ' lies in a suitable neighbourhood $U_\lambda \subseteq S$ of λ. By compactness of S we can find $\lambda_1, \dots, \lambda_N \in S$ such that $S = \bigcup_{i=1}^N U_{\lambda_i}$. Let $E_i := E_{\lambda_i}$ for simplicity.

Moreover, there exists a finite-dimensional $E_0 \subseteq X$ such that $dF_0(E_0) = \operatorname{im} dF_0$. Now set $E := \left\langle \bigcup_{i=0}^N E_i \right\rangle$: $G := F|_{E \cap N}$ still verifies the hypotheses as $\operatorname{im} dG_0 = \operatorname{im} dF_0$ (because $E_0 \subseteq E$) and, for every i, we have $E_i \subseteq \ker dG_0$ (as $dG_0 = dF_0|_E$). Thus, possibly replacing F by G, X by E and N by $E \cap N$, we can assume that X is finite-dimensional.

Now Lemma 4.10 gives us a regular zero $\overline{v}$ of $\operatorname{Hess} F_0$, so that Lemma 4.9 applies, proving local openness of F at 0. □

4.3. Hessian of the endpoint map and Goh minimality conditions

In this section we are going to apply Theorem 4.11 to prove that, when $\mathcal{D} = \langle X_1, \dots, X_r \rangle$, any constant-speed strictly abnormal length minimizer must satisfy the Goh conditions stated below. An immediate consequence will be the smoothness of constant-speed length minimizers when the step of the sub-Riemannian structure is at most 2.

THEOREM 4.12. Let $\gamma : [0,1] \to M$ be a constant-speed length minimizer and assume that it is a strictly abnormal extremal. Then there exists a dual curve $\lambda : [0,1] \to T^*M$, with $\lambda(t) \neq 0$ for any t, satisfying the *Goh conditions*

$$\langle \lambda(t), [X_i, X_j](\gamma(t)) \rangle = 0,\ \forall t \in [0,1],\ i,j = 1, \dots, r.$$

Proof of Theorem 4.12. Let $x := \gamma(0)$, $y := \gamma(1)$ and recall that γ (being constant-speed) minimizes the energy over $\Omega_{x,y}$, as well. We will omit the dependence of End on the starting point. The proof is divided in several steps.

Step 1. We exploit the minimality of γ as in Section 2.7: calling $u \in \mathcal{U}$ the control associated to γ, the extended endpoint map $\operatorname{extEnd} = (\operatorname{End}, E)$ cannot be locally open at u. Equivalently, $\widehat{\operatorname{extEnd}} = (\widehat{\operatorname{End}}, E)$ is not locally open at u (recall that $\widehat{\operatorname{End}} = \Phi_1^{-1} \circ \operatorname{End}$ and Φ_t is the flow associated to u, for $t \in [0,1]$).

Thus, since $\widehat{\operatorname{extEnd}} \in C^\infty(\mathcal{U}, \mathbb{R}^n \times \mathbb{R})$ (see Corollary B.16), Theorem 4.11 tells us that $\operatorname{ind}_u \widehat{\operatorname{extEnd}} < 0$. We notice that, writing any element of $\left(\operatorname{im} d\,\widehat{\operatorname{extEnd}}_u\right)^\perp$ as $(\lambda, \nu) \in (\mathbb{R}^n)^* \times \mathbb{R}$, we always have $\nu = 0$, since otherwise γ would be a normal extremal. Moreover, $(\lambda, 0) \in \left(\operatorname{im} d\,\widehat{\operatorname{extEnd}}_u\right)^\perp$ iff $\lambda \in \left(\operatorname{im} d\,\widehat{\operatorname{End}}_u\right)^\perp$. So there exists some $\overline{\lambda} \in \left(\operatorname{im} d\,\widehat{\operatorname{End}}_u\right)^\perp \setminus \{0\}$ such that

$$\operatorname{ind} \overline{\lambda} \operatorname{Hess} \widehat{\operatorname{End}}_u < n.$$

In particular, this index is finite. This will be sufficient to deduce the Goh conditions.

Step 2. Let $\Psi_{t,v} := \Phi_t^* \langle v(t), X \rangle$, where $\langle v(t), X \rangle$ is shorthand for $\sum_{i=1}^r v_i(t) X_i$. Let us prove that, when $u, v \in C^\infty([0,1], \mathbb{R}^r)$ and $u \in \mathcal{U}$, we have

$$\left.\frac{\partial^2}{\partial s^2} \widehat{\operatorname{End}}(u + sv)\right|_{s=0} = 2 \int_0^1 \int_0^t d\Psi_{t,v} \left[\Psi_{\tau,v}\right](0)\, d\tau\, dt$$

in local coordinates centered at x. Indeed, from Proposition C.11 we have (for any sufficiently small s)

$$\widehat{\mathrm{End}}_t(u+sv) = s\int_0^t \Psi_{\tau,v}\left(\widehat{\mathrm{End}}_\tau(u+sv)\right)\,d\tau$$

for any $t\in[0,1]$, where $\widehat{\mathrm{End}}_t := \Phi_t^{-1}\circ \mathrm{End}_t$. So

$$\begin{aligned}
\frac{\partial^2}{\partial s^2}\widehat{\mathrm{End}}(u+sv)\Big|_{s=0} &= 2\int_0^1 \frac{\partial}{\partial s}\left(\Psi_{t,v}\left(\widehat{\mathrm{End}}_t(u+sv)\right)\right)\Big|_{s=0} dt\\
&= 2\int_0^1 d\Psi_{t,v}(0)\left[\int_0^t \Psi_{\tau,v}(0)\,d\tau\right]\,dt\\
&= 2\int_0^1\int_0^t d\Psi_{t,v}\left[\Psi_{\tau,v}\right](0)\,d\tau\,dt.
\end{aligned}$$

From the smoothness of $\widehat{\mathrm{End}}$ it follows that the same formula holds for any $u\in\mathcal{U}$ and any $v\in L^2([0,1],\mathbb{R}^r)$, provided we show that the right-hand side is continuous in u, v. But the right-hand side equals

$$2\sum_{i,j}\int_0^1\int_0^t v_i(t)v_j(\tau)d\left(\Phi_t^*X_i\right)\left[\Phi_\tau^*X_j\right](0)\,d\tau\,dt$$

and it suffices to know that, whenever $u_n\to u$ in L^2, we have $\Phi_t^n\to\Phi_t$ uniformly on $[0,1]\times K$, for any sufficiently small compact neighbourhood K of 0, and that the same convergence holds for the first and second spatial derivatives of Φ_t^n (here Φ_t^n is the flow associated to u_n). This follows from Proposition B.12 and the compactness of $[0,1]\times K$.

Moreover, recalling that $v\in\ker d\,\widehat{\mathrm{End}}(u)$ iff $\int_0^1\Psi_{t,v}(0)\,dt=0$ (see Lemma 2.32), we obtain that, for every $v\in\ker d\,\widehat{\mathrm{End}}(u)$,

$$\begin{aligned}
d^2\,\widehat{\mathrm{End}}(u)[v,v] &= 2\int_0^1\int_0^t d\Psi_{t,v}\left[\Psi_{\tau,v}\right](0)\,d\tau\,dt\\
&= \int_0^1\int_0^t d\Psi_{t,v}\left[\Psi_{\tau,v}\right](0)\,d\tau\,dt - \int_0^1\int_t^1 d\Psi_{t,v}\left[\Psi_{\tau,v}\right](0)\,d\tau\,dt\\
&= \int_0^1\int_0^t d\Psi_{t,v}\left[\Psi_{\tau,v}\right](0)\,d\tau\,dt - \int_0^1\int_0^\tau d\Psi_{t,v}\left[\Psi_{\tau,v}\right](0)\,dt\,d\tau\\
&= \int_0^1\int_0^t \left(d\Psi_{t,v}\left[\Psi_{\tau,v}\right](0) - d\Psi_{\tau,v}\left[\Psi_{t,v}\right](0)\right)\,d\tau\,dt\\
&= \int_0^1\int_0^t \left[\Psi_{\tau,v},\Psi_{t,v}\right](0)\,d\tau\,dt,
\end{aligned}$$

where in the second-to-last equality we interchanged the names of t and τ.

Step 3. We remember that $\overline{\lambda}$ gives rise to a dual curve $\lambda(t)$ defined by $\lambda(0):=\overline{\lambda}$ and $\lambda(t):=\left(\Phi_t^{-1}\right)^*(\lambda(0)) = \left(d\left(\Phi_t\right)_x^{-1}\right)^*\overline{\lambda}$. λ is the required dual curve: assume instead that for some t_0 and some i,j we have

$$\langle\lambda(t_0),\left[X_i,X_j\right](\gamma(t_0))\rangle \neq 0.$$

To reach a contradiction, it suffices to show that the bilinear form on $L^2([0,1],\mathbb{R}^r)$

$$B(v,w) := \left\langle\overline{\lambda},\int_0^1\int_0^t \left[\Psi_{\tau,v},\Psi_{t,w}\right](x)\,d\tau\,dt\right\rangle$$

is negative definite on subspaces whose dimension is arbitrarily large, because then (using the fact that $\ker d\widehat{\mathrm{End}}(u)$ has finite codimension) this will also hold with $L^2([0,1],\mathbb{R}^r)$ replaced by $\ker d\widehat{\mathrm{End}}(u)$, where $B(v,v) = \overline{\lambda}\,\mathrm{Hess}\,\widehat{\mathrm{End}}(u)[v,v]$.

Fix any $v, w \in L^2([0,1],\mathbb{R}^r)$ and define $v_\epsilon \in L^2([0,1],\mathbb{R}^r)$ by $v_\epsilon(t) := v\left(\frac{t-t_0}{\epsilon}\right)$ (and $v(t) := 0$ outside the interval $[t_0, t_0+\epsilon]$). w_ϵ is defined similarly. We now compute an asymptotic formula for $\frac{1}{\epsilon^2}B(v_\epsilon, w_\epsilon)$:

$$\begin{aligned}
\frac{1}{\epsilon^2}B(v_\epsilon, w_\epsilon) &= \frac{1}{\epsilon^2}\left\langle \overline{\lambda}, \int_{t_0}^{t_0+\epsilon}\int_{t_0}^{t} \left[\Psi_{\tau,v_\epsilon}, \Psi_{t,w_\epsilon}\right](x)\, d\tau\, dt\right\rangle \\
&= \left\langle \overline{\lambda}, \int_0^1\int_0^s \left[\Psi_{t_0+\epsilon\sigma,v}, \Psi_{t_0+\epsilon s,w}\right](x)\, d\sigma\, ds\right\rangle \\
&\to \left\langle \overline{\lambda}, \int_0^1\int_0^s \left[\Phi_{t_0}^* \langle v(\sigma), X\rangle, \Phi_{t_0}^* \langle w(s), X\rangle\right](x)\, d\sigma\, ds\right\rangle \\
&= \left\langle \overline{\lambda}, \int_0^1\int_0^s \Phi_{t_0}^* \left[\langle v(\sigma), X\rangle, \langle w(s), X\rangle\right](x)\, d\sigma\, ds\right\rangle \\
&= \int_0^1\int_0^s \lambda(t_0)\left[\langle v(\sigma), X\rangle, \langle w(s), X\rangle\right](\gamma(t_0))\; d\sigma\, ds.
\end{aligned}$$

We exploited the continuity (with respect to time) of Φ_t and its spatial derivatives to pass to the limit.

Step 4. Again, it suffices to show that

$$B'(v,w) := \int_0^1\int_0^s \lambda(t_0)\left[\langle v(\sigma), X\rangle, \langle w(s), X\rangle\right](\gamma(t_0))\; d\sigma\, ds$$

is negative definite on subspaces of $L^2([0,1],\mathbb{R}^r)$ with arbitrarily large dimension. Now we choose v to be of the form $v_i(t) := \sum_{k=1}^N a_k \cos(2\pi k t)$, $v_j(t) := \sum_{k=1}^N a_k \sin(2\pi k t)$ and $v_\ell = 0$ for $\ell \neq i,j$. Such controls form an N-dimensional subspace. We obtain

$$\begin{aligned}
B'(v,v) &= \langle \lambda(t_0), [X_i, X_j](\gamma(t_0))\rangle \int_0^1\int_0^s \left(v_i(\sigma)v_j(s) - v_j(\sigma)v_i(s)\right)\, d\sigma\, ds \\
&= \langle \lambda(t_0), [X_i, X_j](\gamma(t_0))\rangle \sum_{k=1}^N\sum_{k'=1}^N \frac{a_k a_{k'}}{2\pi k}\int_0^1 \big(\sin(2\pi k s)\sin(2\pi k' s) \\
&\quad - (1-\cos(2\pi k s))\cos(2\pi k' s)\big)\; ds \\
&= \langle \lambda(t_0), [X_i, X_j](\gamma(t_0))\rangle \sum_{k=1}^N \frac{a_k^2}{2\pi k} < 0
\end{aligned}$$

for any $v \neq 0$, provided we assume $\langle \lambda(t_0), [X_i, X_j](\gamma(t_0))\rangle < 0$ (we always can, possibly inverting the roles of i and j). Since N is arbitrary, this completes the proof. $\square$

Corollary 4.13. If M is a sub-Riemannian manifold with step 2, then all constant-speed length minimizers are smooth.

Proof. Let $\gamma : [0,T] \to M$ be a constant-speed geodesic. Since smoothness is a local issue, we can localize and assume $\mathcal{D} = \langle X_1, \dots, X_r\rangle$. By rescaling the speed of γ, we can also assume that $T = 1$. If γ is a normal extremal, then it is also smooth. In order to conclude, it suffices to show that γ cannot be a strictly abnormal extremal. Assume by contradiction that this is the case: by Theorem 4.12, there exists a dual curve λ satisfying $\langle \lambda(t), [X_i, X_j](\gamma(t))\rangle = 0$, for all $t \in [0,1]$ and all i,j. Since (γ, λ) cannot be a normal biextremal, equation (2.2) holds with $\overline{\nu} = 0$, so $\langle \lambda(t), X_i(\gamma(t)) \rangle = 0$ for all t and all i, as well. Since the vector fields $X_i, [X_i, X_j]$ span pointwise all the tangent space, we deduce $\lambda \equiv 0$, which contradicts the nontriviality of λ. $\square$

When M has rank 2, the Goh conditions (which pointwise become a single constraint on $\lambda(t)$) can be proved in a much easier way, by a direct and elementary computation, as we now see.

THEOREM 4.14. Let M be a sub-Riemannian manifold with rank 2 and $\mathcal{D} = \langle X_1, X_2 \rangle$. If (γ, λ) is an abnormal biextremal, with γ having a constant nonzero speed, then the Goh conditions are automatically satisfied (even without assuming that γ is a minimizer), i.e.

$$\langle \lambda(t), [X_1, X_2](\gamma(t)) \rangle = 0$$

for all t.

Proof. Let us call $u = (u_1, u_2)$ the control associated to γ. We can differentiate the identity $\langle \lambda(t), X_1(\gamma(t)) \rangle = 0$ as follows: for any Lebesgue point t of the control we have, using Proposition C.10 in the appendix,

$$\begin{aligned} 0 = \langle \lambda(t+h), X_1(\gamma(t+h)) \rangle &= \left\langle \left(\Phi_{t,t+h}(\langle u, X \rangle)^{-1} \right)^* \lambda(t), X_1(\gamma(t+h)) \right\rangle \\ &= \left\langle \lambda(t), d\Phi_{t,t+h}(\langle u, X \rangle)^{-1} X_1(\gamma(t+h)) \right\rangle \\ &= \langle \lambda(t), [u_1(t)X_1 + u_2(t)X_2, X_1](\gamma(t)) \rangle h + o(h) \\ &= u_1(t) \langle \lambda(t), [X_2, X_1](\gamma(t)) \rangle h + o(h). \end{aligned}$$

So $u_1(t) \langle \lambda(t), [X_2, X_1](\gamma(t)) \rangle = 0$ and similarly $u_2(t) \langle \lambda(t), [X_1, X_2](\gamma(t)) \rangle = 0$. Since $|u(t)| \neq 0$, we cannot have $u_1(t) = u_2(t) = 0$, thus we deduce

$$\langle \lambda(t), [X_1, X_2](\gamma(t)) \rangle = 0$$

for a.e. $t \in [0, \overline{T}]$. But the left-hand side is continuous in t, so this holds for every t. $\square$

This proof can be generalized to obtain that any constant-speed geodesic is smooth on an open dense subset of times, which we prove below. Nonetheless, what follows is not sufficient to obtain that this subset has full measure (which is still an open problem).

THEOREM 4.15. If M is a sub-Riemannian manifold with rank 2 and $\gamma : [0, T] \to M$ is a constant-speed length minimizer, then γ is smooth (i.e. C^∞-regular) on an open dense subset of $[0, T]$.

Proof. We can assume that γ is unit-speed and, by localizing, that $\mathcal{D} = \langle X_1, X_2 \rangle$. If γ is a normal extremal, there is nothing to prove. Suppose then that γ is abnormal and let λ be a dual curve such that (γ, λ) is an abnormal biextremal. Fix any $\overline{t} \in [0, T]$ and any open neighbourhood U of $\overline{t}$. We define

$$k := \min \left\{ j : \exists t \in U \text{ s.t. } \lambda(t) \notin \mathrm{Lie}^j(\mathcal{D}, \gamma(t))^\perp \right\}$$

(see Definition 2.5). By the bracket-generating condition and the fact that λ is nonzero at all times, the above set is nonempty. Moreover, $k \geq 2$ (as (γ, λ) is an abnormal biextremal). Let $t' \in U$ be such that $\lambda(t') \notin \mathrm{Lie}^k(\mathcal{D}, \gamma(t'))^\perp$: by Remark 2.4, there exists $i_1, \ldots, i_k \in \{1, 2\}$ such that

$$\left\langle \lambda(t), [X_{i_1}, [\cdots, [X_{i_{k-1}}, X_{i_k}] \cdots]](\gamma(t')) \right\rangle \neq 0.$$

Let us call V the open subset of U consisting of the times satisfying the displayed condition and notice that, by the minimality of k, setting

$$Y := [X_{i_2}, [\cdots, [X_{i_{k-1}}, X_{i_k}]\cdots]],$$

we have $\langle\lambda(t), Y(\gamma(t))\rangle \equiv 0$ on U (and in particular on V). Repeating the first part of the preceding proof with X_1 replaced by Y, we obtain

$$u_1(t)\,\langle\lambda(t), [X_1, Y]\,(\gamma(t))\rangle + u_2(t)\,\langle\lambda(t), [X_2, Y]\,(\gamma(t))\rangle = 0$$

for a.e. $t \in V$. Hence, at these times,

$$u(t) = \begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix}, \quad v(t) := \begin{pmatrix} \langle\lambda(t), [Y, X_2]\,(\gamma(t))\rangle \\ \langle\lambda(t), [X_1, Y]\,(\gamma(t))\rangle \end{pmatrix}$$

are linearly dependent as elements of $\mathbb{R}^2$. As v never vanishes on V, we deduce $u(t) = \pm\frac{v(t)}{|v(t)|}$. Let us set $\overline{v}(t) := \frac{v(t)}{|v(t)|}$, so that $u(t) = \alpha(t)\overline{v}(t)$ a.e. for some Borel function $\alpha : V \to \{\pm 1\}$.

Let V' be a connected component of V and fix any $s \in V$. We now prove that γ is smooth on V'. Define $h(t) := \int_s^t \alpha(\tau)\,d\tau$ for $t \in V'$. Plugging $u(t) = \dot{h}(t)\overline{v}(t)$ into $\dot{\gamma}(t) = \langle u(t), X\rangle\,(\gamma(t))$ and into (2.3), observing that $\overline{v}(t)$ depends on t only through $\gamma(t)$ and $\lambda(t)$, we deduce that

$$\frac{d}{dt}(\gamma(t), \lambda(t)) = \dot{h}(t) Z(\gamma(t), \lambda(t))$$

for a suitable smooth autonomous vector field Z in the cotangent bundle T^*M. By Proposition C.13 (whose statement of course also holds with V and s in place of $[0,T]$ and 0), we obtain $\gamma = \delta \circ h$, where δ is the projection on M of the integral curve of Y with initial condition $(\gamma(s), \lambda(s))$ (at time s). Taking into account that γ is a minimizer, we deduce that $h(t)$ is monotone, so that $\alpha(t) = 1$ a.e. or $\alpha(t) = -1$ a.e. (on V'), which give $\gamma(t) = \delta(t - s)$ for any $t \in V'$ or $\gamma(t) = \delta(t - s)$ for any $t \in V'$, respectively. □

Recently Sussmann, in [Sus14], obtained a similar result for any rank, under the additional hypothesis that M, $\mathcal{D}$ and g are real analytic (which is the case for Carnot groups). Namely, he proved that any constant-speed geodesic is real analytic on an open dense subset of its domain.

4.4. Minimality of short nice abnormal extremals

The last results lead naturally to the following question: do strictly abnormal geodesics exist at all? Assuming $\mathcal{D} = \langle X_1, \ldots, X_r\rangle$, the proof of Corollary 4.13 shows that they cannot exist when the step is 2. In [TY13], it is claimed that they do not appear in Carnot groups with step at most 3 (although there is no general consensus on the validity of this paper).

In fact, it was originally believed that length minimizers are all normal, starting from a work by Strichartz ([Str86]) which contained a flawed application of the Pontryagin Maximum Principle. In [Mon94] Montgomery gave the first explicit example of a strictly abnormal minimizer in a sub-Riemannian manifold, while in [GK95] Golé and Karidi showed that this phenomenon occurs also in the context of Carnot groups.

In what follows we will give a sufficient condition, due to Liu and Sussmann, which guarantees that, given a strictly abnormal extremal $\gamma : [0, \overline{T}] \to M$, one can find a small positive $T \leq \overline{T}$ such that $\gamma|_{[0,T]}$ is a length minimizer.

Using this result, we will exhibit an example of a strictly abnormal minimizer in a Lie group, endowed with a left-invariant distribution with step 3.

In the final part of this chapter we will review an example by Golé and Caridi of a strictly abnormal minimizer in a step 4 Carnot group.

DEFINITION 4.16. An abnormal biextremal $(\gamma, \lambda) : [0, \overline{T}] \to \mathbb{R}^n \times T^*(\mathbb{R}^n)$, with γ having unit speed, is said to be a *nice abnormal biextremal* if, for any $t \in [0, \overline{T}]$, we have $\lambda(t) \in \mathrm{Lie}^2(\mathcal{D})^\perp \setminus \mathrm{Lie}^3(\mathcal{D})^\perp$ (see Definition 2.2).

From now on we restrict our attention to the case $r = 2$, r denoting the rank of the distribution $\mathcal{D}$. Since we are interested in showing the minimality of short initial pieces of a given horizontal path, which is a local matter (by Remark 2.22), we also assume $M = \mathbb{R}^n$ and $\mathcal{D} = \langle X_1, X_2 \rangle$. Here X_1 and X_2 are two smooth vector fields which are orthonormal with respect to g, as usual.

Let us begin with some easy observations concerning nice abnormal biextremals.

REMARK 4.17. By Theorem 4.14, the requirement $\lambda(t) \in \mathrm{Lie}^2(\mathcal{D})^\perp$ for all $t \in [0, \overline{T}]$ already follows from the fact that (γ, λ) is a unit-speed abnormal extremal.

REMARK 4.18. Setting $Y := [X_1, X_2]$, one can repeat the part of the proof of Theorem 4.15 coming after the definition of Y, with V and V' replaced by $[0, T]$ (this time, the fact that v is always nonzero comes from the hypothesis $\lambda(t) \notin \mathrm{Lie}^3(\mathcal{D})^\perp$).

In particular, although in this section we are interested in the converse direction, one deduces this fact, which was first observed by Liu and Sussmann (1995): if a nice abnormal extremal is a minimizer, then it is smooth.

We will prove the following converse of the preceding remark.

THEOREM 4.19. If $(\gamma, \lambda) : [0, \overline{T}] \to T^*\mathbb{R}^n$ is a smooth nice abnormal biextremal, then for any sufficiently small $T > 0$ the curve $\gamma|_{[0,T]}$ is a strict length minimizer.

Rather than proving that the curves joining $\gamma(0)$ to $\gamma(T)$ have length at least T (which would force us to use some kind of calibration), we will proceed in the opposite way. More precisely, we will prove a constrained rigidity result (Theorem 4.25): we will show that any curve $\delta : [0, T] \to \mathbb{R}^n$ with speed at most 1 and $\delta(0) = \gamma(0)$ has a final point $\delta(T) \neq \gamma(T)$, unless $\delta \equiv \gamma$, whenever T is small and the control $\overline{u} + u$ of δ is close to the control $\overline{u}$ of γ.

The strategy would be to prove that $\lambda(T) d^2(\mathrm{End}_T)_{\overline{u}}$ is positive definite. It turns out that a weaker estimate is true: see the statement of Lemma 4.23, where $\|\cdot\|_*$ is a sort of H^{-1} norm, whose definition involves the primitive of u. After some reductions and definitions, we will compute the first and second differential of End_T, expressing them in terms of the primitive of u. We will also have to carefully estimate the remainder in the second order expansion of End_T, in order to bound it with $o(\|u\|_*^2)$. However, the main argument is contained in Lemma 4.22, Lemma 4.23 and Theorem 4.25.

In order to prove Theorem 4.19, we can assume that $\gamma(0) = 0$ and $\dot{\gamma}(t) = X_1(\gamma(t))$ for all t (to do this, for instance, we can extend $\dot{\gamma}$ to a vector field Z in a neighbourhood V of $\gamma(0)$; then, shrinking V, we can choose a vector field X_2' on V such that $X_2' \in \mathcal{D}|_V$, $|X_2'| = 1$ and $\langle X_2', Z \rangle = 0$; finally we take again a unit vector field $X_1' \in \mathcal{D}|_V$ satisfying

$\langle X_1', X_2' \rangle = 0$, so that X_1' extends $\dot{\gamma}$ up to replacing X_1' with $-X_1'$, and we are done by shrinking $\overline{T}$, replacing X_1, X_2 with X_1', X_2' and mapping V to $\mathbb{R}^n$).

It will also be useful to have at our disposal two smooth differential forms $\omega_1, \omega_2 : \overline{B}_R \to T^*\mathbb{R}^n$ such that $\omega_i(X_j) = \delta_{ij}$. Their existence is clear for a small positive R, which we fix from now on (it can be proved that such ω_1 and ω_2 exist on the whole of $\mathbb{R}^n$, but we do not need this stronger statement).

We also assume without loss of generality that $\overline{T} \le 1$ and $\gamma([0, \overline{T}]) \subseteq \overline{B}_R$. For any fixed $0 < T \le \overline{T}$, we will call $\overline{u}$ the control associated to γ, i.e. $\overline{u}(t) := (1, 0)$ for all $t \in [0, T]$. We will also call $q := \gamma(T)$. Notice that this is an abuse of notation, since in fact $\overline{u}$ and q depend on T.

The elements of $\mathcal{U}_T$, which is the domain of $\mathrm{End}_T = \mathrm{End}_T(0, \cdot)$, will be conveniently written as $\overline{u} + (\dot{v}, \dot{w})$. This means that, for a generic translated control $u = (u_1, u_2) \in \mathcal{U}_T - \overline{u}$ (corresponding to the real control $\overline{u} + u \in \mathcal{U}_T$), we implicitly define $v(t) := \int_0^t u_1(\tau)\, d\tau$ and $w(t) := \int_0^t u_2(\tau)\, d\tau$ (so that in particular $v, w \in H^1([0, T])$ and $v(0) = w(0) = 0$). We define a weaker norm $\|\cdot\|_*$ on $L^2([0, T], \mathbb{R}^2)$, which is the one induced by the scalar product

$$\langle u, u' \rangle_* := v(T)v'(T) + w(T)w'(T) + \int_0^T \left(vv' + ww'\right)(t)\, dt.$$

Finally, we define

$$\mathcal{V}_T := \{u : \overline{u} + u \in \mathcal{U}_T,\ |\overline{u} + u| \le 1 \text{ a.e.}\}.$$

Notice that, if $u \in \mathcal{V}_T$, then $\dot{v} \le 0$ a.e. This simple fact will be crucial to carry out some estimates.

In the proofs of Lemmas 4.20, 4.21 and 4.24, we will implicitly assume that the involved control u is smooth. The fact that the respective theses hold without this hypothesis follows immediately by a standard density argument and by the smoothness of End_T.

LEMMA 4.20 (first order differential). For any $0 < T \le \overline{T}$ the endpoint map $\mathrm{End}_T : \mathcal{U}_T \to \mathbb{R}^n$ satisfies

$$d(\mathrm{End}_T)_{\overline{u}}[u] = v(T)X_1(q) + w(T)X_2(q) - \int_0^T w(t)\dot{Y}_t^T(q)\, dt, \tag{4.1}$$

where $q = \gamma(T)$, $Y_t^T := d\Phi_{T-t}(X_1)_* X_2$ and $\dot{Y}_t^T := \frac{\partial}{\partial t} Y_t^T$.

Proof. Let us fix any $0 < T \le \overline{T}$. From Proposition C.12 in the appendix and the fact that $\Phi_{t,T}(\langle \overline{u}, X \rangle) = \Phi_{T-t}(X_1)$ (see Definition C.2 for the notation) we have

$$\mathrm{End}_T(\overline{u} + u) = \Phi_{0,T}(0, \langle \overline{u}, X \rangle + \langle u, X \rangle) = \Phi_{0,T}\left(q, \langle u(t), \Phi_{T-t}(X_1)_* X \rangle\right)$$

for any smooth $u = (\dot{v}, \dot{w}) \in L^2\left([0, T], \mathbb{R}^2\right)$ close to 0. Here $\Phi_{T-t}(X_1)_* X$ is shorthand for $(\Phi_{t,T}(X_1)_* X_1, \Phi_{t,T}(X_1)_* X_2)$. Since $\Phi_{T-t}(X_1)_* X_1 = X_1$ (by Proposition C.9),

$$\Phi_{T-t}(X_1)_* X = (X_1, Y_t^T).$$

By Propositions C.11 and C.13,

$$\begin{aligned} \mathrm{End}_T(\overline{u} + u) &= \Phi_{0,T}(\dot{v}X_1) \circ \Phi_{0,T}\left(\dot{w}(t)\Phi_{0,t}(\dot{v}X_1)^* Y_t^T\right)(q) \\ &= \Phi_{v(T)}(X_1) \circ \Phi_{0,T}\left(\dot{w}(t)\Phi_{v(t)}(X_1)^* Y_t^T\right)(q). \end{aligned} \tag{4.2}$$

Now the estimate

$$\Phi_s(X_1)(x) = x + sX_1(x) + o(s) = x + sX_1(q) + o(|x - q|) + o(s),$$

Corollary C.7 and Proposition C.5 (with $k = 1$) allow us to deduce

$$\begin{aligned}
\mathrm{End}_T(\overline{u}+u) &= \Phi_{v(T)}(X_1)\left(\int_0^T \dot{w}(t)Y_t^T(q)\,dt + O\left(\int_0^T |\dot{w}(t)v(t)|\,dt\right)\right)\\
&= \int_0^T \dot{w}(t)Y_t^T(q)\,dt + v(T)X_1(q) + o(\|u\|_2)
\end{aligned}$$

as $u \to 0$. So finally, noticing that $Y_T^T(q) = X_2(q)$ and integrating by parts,

$$\begin{aligned}
d(\mathrm{End}_T)_{\overline{u}}[u] &= v(T)X_1(q) + \int_0^T \dot{w}(t)Y_t^T(q)\,dt\\
&= v(T)X_1(q) + w(T)X_2(q) - \int_0^T w(t)\dot{Y}_t^T(q)\,dt.
\end{aligned} \tag{4.3}$$

□

LEMMA 4.21 (second order differential). For any $0 < T \le \overline{T}$ we have

$$\begin{aligned}
\left\langle \lambda(T), d^2\,\mathrm{End}_{\overline{u}}[u,u]\right\rangle = \Bigg\langle \lambda(T), &\int_0^T w^2(t)\left[\dot{Y}_t^T, Y_t^T\right]dt\\
&+ \int_0^T \left[w(T)X_2 + \int_0^t w(\tau)\dot{Y}_\tau^T\,d\tau, w(t)\dot{Y}_t^T\right]dt\\
&+ \left(v(T)dX_1 + \int_0^T \dot{w}(t)dY_t^T\,dt\right)\left[d(\mathrm{End}_T)_{\overline{u}}[u]\right]\Bigg\rangle.
\end{aligned} \tag{4.4}$$

In this formula it is meant that all the vector fields are evaluated at q, which we omitted for brevity. The last term is an abbreviation of

$$v(T)dX_1\left[d(\mathrm{End}_T)_{\overline{u}}[u]\right] + \int_0^T \dot{w}(t)dY_t^T\left[d(\mathrm{End}_T)_{\overline{u}}[u]\right]dt.$$

Proof. Let $\lambda := \lambda(T)$. Since End_T is smooth, in order to find $d^2(\mathrm{End}_T)_{\overline{u}}[u,u]$ it suffices to compute the second order expansion of $\mathrm{End}_T(\overline{u}+u)$, as we did in the previous proof in order to find the first order differential. Starting from (4.2) and using Corollary C.7 and Proposition C.5 (with $k = 2$), we have

$$\begin{aligned}
\mathrm{End}_T(\overline{u}+u) = \Phi_{v(T)}(X_1)\Bigg(q &+ \int_0^T \dot{w}(t)Y_t^T\,dt + \int_0^T v(t)\dot{w}(t)\left[X_1, Y_t^T\right]dt\\
&+ \int_0^T\int_0^t \dot{w}(t)\dot{w}(\tau)dY_t^T\left[Y_\tau^T\right]d\tau\,dt + o(\|u\|_2^2)\Bigg)
\end{aligned}$$

(again we use the convention that vector fields are evaluated at q when the point is not specified). Since $\Phi_s(X_1)(x) = x + sX_1(x) + \frac{s^2}{2}dX_1[X_1](x) + o(s^2)$ uniformly in a neighbourhood of q, we deduce

$$\begin{aligned}
\mathrm{End}_T(\overline{u}+u) =& q + d(\mathrm{End}_T)_{\overline{u}}[u] + \frac{v^2(T)}{2}dX_1[X_1] + \int_0^T v(t)\dot{w}(t)\left[X_1, Y_t^T\right]dt\\
&+ \int_0^T\int_0^t \dot{w}(t)\dot{w}(\tau)dY_t^T\left[Y_\tau^T\right]d\tau\,dt + v(T)dX_1\left[\int_0^T \dot{w}(t)Y_t^T\,dt\right] + o(\|u\|_2^2).
\end{aligned}$$

Thus we have the following formula for the second order differential:

$$\begin{aligned}\frac{1}{2}d^2(\mathrm{End}_T)_{\overline{u}}[u,u] =& \frac{v^2(T)}{2}dX_1[X_1] + \int_0^T v(t)\dot{w}(t)\left[X_1, Y_t^T\right]\,dt \\ &+ \int_0^T\int_0^t \dot{w}(t)\dot{w}(\tau)dY_t^T\left[Y_\tau^T\right]\,d\tau\,dt + v(T)\int_0^T \dot{w}(t)dX_1\left[Y_t^T\right]\,dt.\end{aligned}$$

Equation (4.1) shows that $\lambda \in (\operatorname{im} d(\mathrm{End}_T)_{\overline{u}})^\perp$ implies $\left\langle \lambda, \left[X_1, Y_t^T\right]\right\rangle = 0$ for any t: choosing $v = 0$ and letting w vary in $C^1([0,T])$, with $w(0) = w(T) = 0$, we obtain $\left\langle \lambda, \dot{Y}_t^T\right\rangle = 0$. This identity, together with Proposition C.9 and the fact that $Y_t^T = \Phi_{t-T}(X_1)^* X_2$, gives the claim. So the term $\int_0^T v(t)\dot{w}(t)\left[X_1, Y_t^T\right]\,dt$ vanishes when it is paired with λ. Now, since $d(\mathrm{End}_T)_{\overline{u}}[u] = v(T)X_1 + \int_0^T \dot{w}(t)Y_t^T\,dt$ (see (4.3)),

$$\begin{aligned}&\left\langle \lambda, \frac{1}{2}d^2\,\mathrm{End}_{\overline{u}}[u,u] - \frac{1}{2}\left(v(T)dX_1 + \int_0^T \dot{w}(t)dY_t^T\,dt\right)\left[d(\mathrm{End}_T)_{\overline{u}}[u]\right]\right\rangle \\ &= \left\langle \lambda, \int_0^T\int_0^t \dot{w}(t)\dot{w}(\tau)dY_t^T\left[Y_\tau^T\right]\,d\tau\,dt + \frac{1}{2}v(T)\int_0^T \dot{w}(t)dX_1\left[Y_t^T\right]\,dt\right. \\ &\quad \left. - \frac{1}{2}v(T)\int_0^T \dot{w}(t)dY_t^T[X_1] - \frac{1}{2}\int_0^T\int_0^T \dot{w}(t)\dot{w}(\tau)dY_t^T\left[Y_\tau^T\right]\,d\tau\,dt\right\rangle.\end{aligned}$$

We remark that

$$\frac{1}{2}v(T)\int_0^T \dot{w}(t)dX_1\left[Y_t^T\right]\,dt - \frac{1}{2}v(T)\int_0^T \dot{w}(t)dY_t^T[X_1] = -\frac{1}{2}v(T)\int_0^T \dot{w}(t)\left[X_1, Y_t^T\right]\,dt,$$

which again gives no contribution due to the presence of $\left[X_1, Y_t^T\right]$, while

$$\begin{aligned}&\int_0^T\int_0^t \dot{w}(t)\dot{w}(\tau)dY_t^T\left[Y_\tau^T\right]\,d\tau\,dt - \frac{1}{2}\int_0^T\int_0^T \dot{w}(t)\dot{w}(\tau)dY_t^T\left[Y_\tau^T\right]\,d\tau\,dt \\ &= \frac{1}{2}\int_0^T\int_0^t \dot{w}(t)\dot{w}(\tau)dY_t^T\left[Y_\tau^T\right]\,d\tau\,dt - \frac{1}{2}\int_0^T\int_t^T \dot{w}(t)\dot{w}(\tau)dY_t^T\left[Y_\tau^T\right]\,d\tau\,dt \\ &= \frac{1}{2}\int_0^T\int_0^t \dot{w}(t)\dot{w}(\tau)dY_t^T\left[Y_\tau^T\right]\,d\tau\,dt - \frac{1}{2}\int_0^T\int_0^\tau \dot{w}(t)\dot{w}(\tau)dY_t^T\left[Y_\tau^T\right]\,dt\,d\tau \\ &= \frac{1}{2}\int_0^T\int_0^t \dot{w}(t)\dot{w}(\tau)\left(dY_t^T\left[Y_\tau^T\right] - dY_\tau^T\left[Y_t^T\right]\right)\,d\tau\,dt = \frac{1}{2}\int_0^T\int_0^t \dot{w}(t)\dot{w}(\tau)\left[Y_\tau^T, Y_t^T\right]\,d\tau\,dt,\end{aligned}$$

where in the second-to-last equality we interchanged the names of t and τ. Finally, we integrate by parts twice:

$$\begin{aligned}&\int_0^T\int_0^t \dot{w}(t)\dot{w}(\tau)\left[Y_\tau^T, Y_t^T\right]\,d\tau\,dt = \int_0^T \dot{w}(t)\left(w(t)\left[Y_t^T, Y_t^T\right] - \int_0^t w(\tau)\left[\dot{Y}_\tau^T, Y_t^T\right]\,d\tau\right)\,dt \\ &= \int_0^T\int_\tau^T \dot{w}(t)w(\tau)\left[Y_t^T, \dot{Y}_\tau^T\right]\,dt\,d\tau \\ &= \int_0^T\left(w(T)w(\tau)\left[X_2, \dot{Y}_\tau^T\right] - w^2(\tau)\left[Y_\tau^T, \dot{Y}_\tau^T\right] - \int_\tau^T w(t)w(\tau)\left[\dot{Y}_t^T, \dot{Y}_\tau^T\right]\,dt\right)\,d\tau \\ &= \int_0^T w^2(t)\left[\dot{Y}_t^T, Y_t^T\right]\,dt + \int_0^T\left[w(T)X_2 + \int_0^t w(\tau)\dot{Y}_\tau^T\,d\tau, w(t)\dot{Y}_t^T\right]\,dt.\end{aligned}$$

Combining this with the two previous computations we arrive at the thesis. $\square$

As we will see below, when $u \in \ker d(\mathrm{End}_T)_{\overline{u}}$, $\left\langle \lambda(T), \int_0^T w^2(t) \left[\dot{Y}_t^T, Y_t^T\right] dt \right\rangle$ is the leading term in the above formula as $T \to 0$. We now show that it always takes nonnegative values.

LEMMA 4.22. Possibly replacing λ by $-\lambda$, there exists some $\eta > 0$ such that

$$\left\langle \lambda(T), \left[\dot{Y}_t^T, Y_t^T\right](q) \right\rangle \geq \eta,$$

for any $0 < T \leq \overline{T}$ and any $t \in [0, T]$.

Proof. As already noticed in Remark 4.18, we have $\langle \lambda(t), [[X_1, X_2], X_1](\gamma(t)) \rangle = 0$ for any $t \in [0, \overline{T}]$. Since $\lambda(t) \notin \mathrm{Lie}^3(\mathcal{D})^\perp$ we deduce that

$$\langle \lambda(t), [[X_1, X_2], X_2](\gamma(t)) \rangle \neq 0$$

for all t. By continuity the left-hand side has constant sign. Possibly replacing λ by $-\lambda$, we can assume that it is always positive and we set

$$\eta := \min_{t \in [0, \overline{T}]} \langle \lambda(t), [[X_1, X_2], X_2](\gamma(t)) \rangle .$$

Let us now fix any $0 < T \leq \overline{T}$. Notice that $\Phi_{T-t}(X_1)_*[[X_1, X_2], X_2] = \left[\left[X_1, Y_t^T\right], Y_t^T\right]$, since Φ_{T-t} is a local diffeomorphism carrying X_2 to Y_t^T and X_1 to itself (by Proposition C.9). Recalling that $\lambda(t) = \Phi_{T-t}(X_1)^* \lambda(T)$ and $\gamma(t) = \Phi_{T-t}(X_1)^{-1}(q)$, we deduce

$$\begin{aligned}
\langle \lambda(t), [X_2, [X_1, X_2]](\gamma(t)) \rangle &= \langle \lambda(T), \left(\Phi_{T-t}(X_1)_*[[X_1, X_2], X_2]\right)(q) \rangle \\
&= \left\langle \lambda(T), \left[\left[X_1, Y_t^T\right], Y_t^T\right](q) \right\rangle .
\end{aligned}$$

Finally, we notice that $\left[X_1, Y_t^T\right] = \dot{Y}_t^T$, by Proposition C.9. $\square$

Since (γ, λ) is a nice abnormal biextremal iff so is $(\gamma, -\lambda)$, we can (and will) assume that

$$\left\langle \lambda(T), \left[\dot{Y}_t^T, Y_t^T\right] \right\rangle \geq \eta.$$

Now we prove that $\lambda(T) d^2 \mathrm{End}_{\overline{u}}$ is positive definite on $\ker d \,\mathrm{End}_{\overline{u}}$. For technical reasons we need positive definiteness on a set which is slightly larger than $\ker d \,\mathrm{End}_{\overline{u}}$.

LEMMA 4.23 (positive definiteness). Provided T is sufficiently small, there exist some $\eta', \epsilon > 0$ such that

$$\left\langle \lambda(T), d^2(\mathrm{End}_T)_{\overline{u}}[u, u] \right\rangle \geq \eta' \left\|u\right\|_*^2$$

whenever $u = (\dot{v}, \dot{w}) \in \mathcal{V}_T$ satisfies $|d(\mathrm{End}_T)_{\overline{u}}[u]| \leq \epsilon \left\|u\right\|_*$. As a consequence, the same holds if $|d(\mathrm{End}_T)_{\overline{u}}[u]| \leq \epsilon' \left\|u\right\|_2$, for some $\epsilon' > 0$.

Proof. Applying ω_1 to (4.1) we get

$$\begin{aligned}
|v(T)| &= \left| \langle \omega_1, v(T) X_1 \rangle (q) \right| \\
&= \left| \left\langle \omega_1, \int_0^T w(t) \dot{Y}_t^T \, dt \right\rangle (q) + \langle \omega_1(q), d(\mathrm{End}_T)_{\overline{u}}[u] \rangle \right| \\
&\leq C \left(\int_0^T |w|(t) \, dt + \epsilon \left\|u\right\|_* \right),
\end{aligned}$$

where C is a constant independent of T (since $\dot{Y}_t^T(q)$ is bounded uniformly in t and T). Similarly we get

$$|w(T)| \leq C \left(\int_0^T |w|(t) \, dt + \epsilon \left\|u\right\|_* \right).$$

In addition, since $\dot v \le 0$ and $v(0)=0$, we have $\max_{t\in[0,T]} |v(t)| = |v(T)|$, so that

$$\|v\|_2 = \left(\int_0^T |v|^2(t)\,dt\right)^{1/2} \le T^{1/2}|v(T)|.$$

Finally, $\int_0^T |w|(t)\,dt \le T^{1/2}\|w\|_2$ and $T\le 1$, so we obtain

$$\begin{aligned}|v(T)| + |w(T)| + \|v\|_2 &\le 2|v(T)| + |w(T)| \le 3C\|w\|_2 + 3C\epsilon\|u\|_* \\ &\le 3C\|w\|_2 + 3C\epsilon\left(|v(T)| + |w(T)| + \|v\|_2 + \|w\|_2\right),\end{aligned}$$

which gives, if ϵ is such that $3C\epsilon \le \frac{1}{2}$,

$$\frac{1}{2}\left(|v(T)| + |w(T)| + \|v\|_2\right) \le \left(3C + \frac{1}{2}\right)\|w\|_2 .$$

This shows that $\|u\|_* \le C'\|w\|_2$ for some constant C' independent of T. In particular, $|v(T)|, |w(T)| \le C'\|w\|_2$ (by the definition of $\|\cdot\|_*$). Recalling (4.4) and writing

$$\begin{aligned}&\left(v(T)dX_1 + \int_0^T \dot w(t) dY_t^T\,dt\right)[d(\mathrm{End}_T)_{\overline{u}}[u]] \\ &= \left(v(T)dX_1 + w(T)dX_2 - \int_0^T w(t)d\dot Y_t^T\,dt\right)[d(\mathrm{End}_T)_{\overline{u}}[u]],\end{aligned}$$

we arrive at

$$\begin{aligned}\left\langle \lambda(T), d^2(\mathrm{End}_T)_{\overline{u}}[u,u]\right\rangle \ge& \eta\|w\|_2^2 - |\lambda(T)|\left|\int_0^T \left[w(T)X_2 + \int_0^t w(\tau)\dot Y_\tau^T\,d\tau, w(t)\dot Y_t^T\right](q)\,dt\right| \\ &- |\lambda(T)|\left|\left(v(T)dX_1 + w(T)dX_2 - \int_0^T w(t)d\dot Y_t^T\,dt\right)[d(\mathrm{End}_T)_{\overline{u}}[u]]\right|.\end{aligned}$$

The second term can be estimated by

$$\begin{aligned}&C''\left(|w(T)|\int_0^T |w(t)|\,dt + \int_0^T\int_0^t |w(t)|\,|w(\tau)|\,d\tau\,dt\right) \\ &\le C''\left(C'\|w\|_2\cdot T^{1/2}\|w\|_2 + T\|w\|_2^2\right) = C''(C'T^{1/2}+T)\|w\|_2^2,\end{aligned}$$

while the third term is bounded by

$$C''\|u\|_*\,|d(\mathrm{End}_T)_{\overline{u}}[u]| \le C''\epsilon\|u\|_*^2,$$

for some suitable constant C'' independent of T. So finally

$$\begin{aligned}\left\langle \lambda(T), d^2(\mathrm{End}_T)_{\overline{u}}[u,u]\right\rangle \ge& \eta\|w\|_2^2 - C''(C'T^{1/2} - T)\|w\|_2^2 - C''\epsilon\|u\|_*^2 \\ \ge& \eta(C')^{-2}\|u\|_*^2 - (C''C'T^{1/2} + C''T + C''\epsilon)\|u\|_*^2 .\end{aligned}$$

The first part of the thesis follows once T is chosen sufficiently small and ϵ satisfies also $C''\epsilon \le \frac{1}{2}\eta(C')^{-2}$. The second part follows from the fact that $\|u\|_* \le \widehat{C}\|u\|_2$ (for some $\widehat{C} > 0$ depending on T): we can choose $\epsilon' := \left(\widehat{C}\right)^{-1}\epsilon$. □

LEMMA 4.24 (remainder estimate). For any $0 < T \le \overline{T}$ we have

$$\left\langle \lambda(T), \mathrm{End}_T(\overline{u}+u) - q - \frac{1}{2}d^2(\mathrm{End}_T)_{\overline{u}}[u,u]\right\rangle = o(\|u\|_*^2)$$

as $u \to 0$ in $\mathcal{V}_T$ (with respect to the usual L^2 topology on controls).

Proof. Let $\lambda := \lambda(T)$. We need a different expansion of End_T where, contrary to the one previously used, $\dot{v}$ appears instead of $\dot{w}$. This will allow us to exploit the fact that $\dot{v}$ has constant sign when estimating a third order term. This alternative expansion is obtained as follows:

(4.5)

$$\begin{aligned}
\mathrm{End}_T(\overline{u}+u) =&\Phi_{0,T}\left(0,(1+\dot{v}(t))X_1+\dot{w}(t)X_2\right)\\
=&\Phi_{w(T)}(X_2)\circ\Phi_{0,T}\left((1+\dot{v}(t))\Phi_{w(t)}(X_2)^*X_1\right)(0)\\
=&\Phi_{w(T)}(X_2)\circ\Phi_{0,T}\left((1+\dot{v}(t))\left(\Phi_{w(t)}(X_2)^*X_1-X_1\right)+(1+\dot{v}(t))X_1\right)(0)\\
=&\Phi_{w(T)}(X_2)\circ\Phi_{0,T}\left((1+\dot{v}(t))\left(\Phi_{(T-t)+(v(T)-v(t))}(X_1)_*\Phi_{w(t)}(X_2)^*X_1-X_1\right)\right)\\
&\circ\Phi_{T+v(T)}(X_1)(0)\\
=&\Phi_{w(T)}(X_2)\circ\Phi_{0,T}\left((1+\dot{v}(t))\left(\Phi_{v(T)-v(t)}(X_1)_*\Phi_{w(t)}(Y_t^T)^*X_1-X_1\right)\right)\\
&\circ\Phi_{v(T)}(X_1)(q).
\end{aligned}$$

We used Propositions C.11, C.12 and C.13. Moreover, in the last equality we used the fact that $\Phi_{T-t}(X_1)$ is a local diffeomorphism carrying X_2 to Y_t^T and X_1 to itself (in evaluating the expression $\Phi_{w(t)}(Y_t^T)$, Y_t^T should be considered as an autonomous vector field).

Now, using Corollary C.8,

$$\begin{aligned}
&\Phi_{0,T}\left((1+\dot{v}(t))\left(\Phi_{v(T)-v(t)}(X_1)_*\Phi_{w(t)}(Y_t^T)^*X_1-X_1\right)\right)(x)\\
&=\Phi_{0,T}\left((1+\dot{v}(t))\left(w(t)\left[Y_t^T,X_1\right]+\frac{w^2(t)}{2}\left[Y_t^T,\left[Y_t^T,X_1\right]\right]\right.\right.\\
&\qquad\left.\left.+(v(t)-v(T))w(t)\left[X_1,\left[Y_t^T,X_1\right]\right]\right)\right)(x)+o(\|u\|_*^2).
\end{aligned}$$

So we can replace the intermediate flow in (4.5) by the above approximation.

In addition, if we expand all the three flows using Proposition C.5 with $k=2$ (by the same method as in the proof of Lemma 4.21), the error is still $o(\|u\|_*^2)$: indeed, we can bound $|\dot{v}|\leq 2$ and $\|v\|_\infty$, $\|w\|_\infty$ are arbitrarily small as $u\to 0$, so that the error in the expansion of the intermediate flow is $O(\|v\|_2^3)=o(\|u\|_*^2)$, while the claim is clear for the other two flows.

We are left with an expression containing the second order expansion of $\mathrm{End}_T(\overline{u}+u)$ and some higher order terms, each consisting of a (possibly multiple) integral containing a vector field (evaluated at q) and a monomial of degree at least three in the variables $\dot{v}(t)$, $v(t)$, $w(t)$, $v(T)$ and $w(T)$. In order to conclude it suffices to estimate these terms when they are paired with λ.

The terms where the combined multiplicity of the v's and w's (i.e. the factors different from $\dot{v}(t)$) is at least three are $o(\|u\|_*^2)$, since again we can bound $|\dot{v}|\leq 2$ and all but two of the v's and w's by an arbitrarily small constant; then, if necessary, we can decouple the two remaining factors by the Cauchy-Schwarz inequality. For example

$$\int_0^T\int_0^t|vw|(t)\,|\dot{v}w|(\tau)\,d\tau\,dt\leq o(1)\left(\int_0^T|w|\,dt\right)\left(\int_0^T|w|\,d\tau\right)\leq o(1)T\,\|w\|_2^2=o(\|u\|_*^2).$$

The same expression can also be treated in this way:

$$\int_0^T\int_0^t|vw|(t)\,|\dot{v}w|(\tau)\,d\tau\,dt\leq o(1)\int_0^T|vw|\,dt\leq o(1)\,\|v\|_2\,\|w\|_2=o(\|u\|_*^2).$$

Terms containing $v(T)$ or $w(T)$ are also easy to estimate: if $\dot v(t)$ appears, then also $w(t)$ does; now by Cauchy-Schwarz we can gain the desired $o(1)$ also from $\|\dot v\|_2$. For example

$$|v(T)| \int_0^T |\dot v w| \; dt \le |v(T)| \, \|\dot v\|_2 \, \|w\|_2 = o(\|u\|_*^2).$$

The following five terms are left to be bound:

- $\int_0^T \dot v(t) \frac{w^2(t)}{2} \left[Y_t^T, \left[Y_t^T, X_1\right]\right] \; dt$;
- $\int_0^T \dot v(t) v(t) w(t) \left[X_1, \left[Y_t^T, X_1\right]\right] \; dt$;
- $\int_0^T \int_0^t \dot v(t) w(t) w(\tau) d\left[Y_t^T, X_1\right]\left[\left[Y_\tau^T, X_1\right]\right] \; d\tau \, dt$;
- $\int_0^T \int_0^t w(t) \dot v(\tau) w(\tau) d\left[Y_t^T, X_1\right]\left[\left[Y_\tau^T, X_1\right]\right] \; d\tau \, dt$;
- $\int_0^T \int_0^t \dot v(t) w(t) \dot v(\tau) w(\tau) d\left[Y_t^T, X_1\right]\left[\left[Y_\tau^T, X_1\right]\right] \; d\tau \, dt$.

The last three terms are estimated by the Cauchy-Schwarz inequality: for example

$$\begin{aligned} &\left| \int_0^T \int_0^t \dot v(t) w(t) \dot v(\tau) w(\tau) d\left[Y_t^T, X_1\right]\left[\left[Y_\tau^T, X_1\right]\right] \; d\tau \, dt \right| \\ &\le O(1) \left(\int_0^T |\dot v w| \; dt \right) \left(\int_0^T |\dot v w| \; d\tau \right) \le O(1) \, \|\dot v\|_2^2 \, \|u\|_*^2 \end{aligned}$$

and $\|\dot v\|_2$ is arbitrarily small. The second term gives no contribution, since

$$\left\langle \lambda, \left[X_1, \left[Y_t^T, X_1\right]\right] \right\rangle = 0,$$

which is obtained by differentiating the identity $\left\langle \lambda, [X_1, Y_t^T] \right\rangle = 0$ (as in the proof of Theorem 4.14). Finally, to treat the first term we exploit the fact that $\dot v(t) \le 0$:

$$\left| \int_0^T \dot v(t) \frac{w^2(t)}{2} \left[Y_t^T, \left[Y_t^T, X_1\right]\right] \; dt \right| \le O(1) \int_0^T (-\dot v(t)) w^2(t) \, dt.$$

Integrating by parts we obtain

$$\int_0^T \dot v(t) w^2(t) \, dt = v(T) w^2(T) - 2 \int_0^T v(t) w(t) \dot w(t) \, dt$$

and $v(T) w^2(T) = o(\|u\|_*^2)$, while

$$\left| \int_0^T v(t) w(t) \dot w(t) \, dt \right| \le |v(T)| \int_0^T |w \dot w| \; dt \le |v(T)| \, \|w\|_2 \, \|\dot w\|_2 = o(\|u\|_*^2),$$

thanks to the fact that $|v(t)|$ is monotone increasing. □

Theorem 4.25 (constrained rigidity). Provided T is sufficiently small, there exists some neighbourhood V of 0 in $\mathcal{V}_T$ such that $\mathrm{End}_T(\overline{u} + u) \ne q$ whenever $u \in V \setminus \{0\}$.

We recall that $\mathcal{V}_T$ is the subset of $\mathcal{U}_T - \overline{u}$ consisting of the translated controls u such that $|\overline{u} + u| \le 1$ a.e.

Proof. Fix any T such that Lemma 4.23 applies, providing us with the two constants η' and ϵ'. We can write

$$L^2([0,T], \mathbb{R}^2) = \ker d(\mathrm{End}_T)_{\overline{u}} \oplus N$$

for some finite-dimensional subspace N. We decompose any $u \in \mathcal{V}_T$ as $u = \xi + \zeta$, where $\xi \in \ker d(\mathrm{End}_T)_{\overline{u}}$ and $\zeta \in N$. Notice that $d(\mathrm{End}_T)_{\overline{u}}$ restricts to an isomorphism between N and $\mathrm{im}\, d(\mathrm{End}_T)_{\overline{u}}$, so for some $C > 0$ the following inequalities hold:

$$C^{-1} \|\zeta\|_2 \le |d(\mathrm{End}_T)_{\overline{u}}[\zeta]| = |d(\mathrm{End}_T)_{\overline{u}}[u]| \le C \, \|\zeta\|_2 \, .$$

By Lemma 4.24 we can find a neighbourhood V of 0 in $\mathcal{V}_T$ such that

$$\left|\left\langle \lambda(T), \mathrm{End}_T(\overline{u}+u) - q - \frac{1}{2}d^2(\mathrm{End}_T)_{\overline{u}}[u,u]\right\rangle\right| < \eta' \|u\|_*^2 . \tag{4.6}$$

for any $u \in V \setminus \{0\}$. Shrinking V if necessary, we can assume that on $V \setminus \{0\}$ we also have

$$|\mathrm{End}_T(\overline{u}+u) - q - d(\mathrm{End}_T)_{\overline{u}}[u]| < C^{-2}\epsilon' \|u\|_2 .$$

Now let us fix any $u \in V \setminus \{0\}$. We distinguish two cases: if $\|\zeta\|_2 > C^{-1}\epsilon' \|u\|_2$, then

$$\begin{aligned}|\mathrm{End}_T(\overline{u}+u) - q| &\geq |d(\mathrm{End}_T)_{\overline{u}}[u]| - |\mathrm{End}_T(\overline{u}+u) - q - d(\mathrm{End}_T)_{\overline{u}}[u]| \\ &> C^{-1}\|\zeta\|_2 - C^{-2}\epsilon' \|u\|_2 > 0,\end{aligned}$$

so the thesis is true in this case. On the other hand, if $\|\zeta\|_2 \leq C^{-1}\epsilon' \|u\|_2$ we deduce

$$|d(\mathrm{End}_T)_{\overline{u}}[u]| \leq C \|\zeta\|_2 \leq \epsilon' \|u\|_2 .$$

Thus Lemma 4.23 and (4.6) give

$$\langle \lambda(T), \mathrm{End}_T(\overline{u}+u) - q\rangle \geq \left\langle \lambda(T), d^2(\mathrm{End}_T)_{\overline{u}}[u,u]\right\rangle - \eta' \|u\|_*^2 > 0,$$

so again $\mathrm{End}_T(\overline{u}+u) \neq q$. □

Theorem 4.19 can be deduced as a corollary of Theorem 4.25.

Proof of Theorem 4.19. By Theorem 4.25 there exist some $0 < T' \leq \overline{T}$ and a neighbourhood V of 0 in $\mathcal{V}_{T'}$ such that $\mathrm{End}_{T'}(\overline{u}+u) \neq \gamma(T')$ whenever $u \in V \setminus \{0\}$. We now fix any $0 < T \leq T'$ such that, whenever $u \in \mathcal{V}_{T'}$ satisfies $\|u\|_2 \leq 2T^{1/2}$, we have $u \in V$.

Assume by contradiction that there exists a constant-speed horizontal path $\delta : [0,T] \to \mathbb{R}^n$, $\delta \neq \gamma|_{[0,T]}$ connecting 0 to $\gamma(T)$ with $L(\delta) \leq L\left(\gamma|_{[0,T]}\right)$. Let $h \in L^2([0,T], \mathbb{R}^2)$ be the control associated to δ. Since γ has unit speed, we have $|h| = \left|\dot{\delta}\right| \leq 1$ a.e. on $[0,T]$. Let us extend h to a new control $\overline{h} \in L^2([0,T'], \mathbb{R}^2)$ in this way:

$$\overline{h}(t) := \begin{cases} h(t) & \text{if } t \in [0,T] \\ \overline{u}(t) & \text{if } t \in [T,T'] \end{cases}$$

It is clear that $u := \overline{h} - \overline{u} \in \mathcal{V}_{T'}$, since the trajectory obtained by travelling along δ and then along $\gamma|_{[T,T']}$ is horizontal, it has speed at most 1 a.e. and its associated control is $\overline{h}$. This argument also shows that $\mathrm{End}_T(\overline{u}+u) = \gamma(T')$. Finally, $u \neq 0$ and

$$\|u\|_2 = \left(\int_0^T |h - \overline{u}|^2 \, dt\right)^{1/2} \leq 2T^{1/2},$$

which implies $u \in V \setminus \{0\}$. So $\mathrm{End}_{T'}(\overline{u}+u) \neq \gamma(T')$ and we have the desired contradiction. □

4.5. A strictly abnormal minimizer in a Lie group

This example is due to Liu and Sussmann. Let $\mathbb{G} := SO(3) \times \mathbb{R}$, which is a Lie group with Lie algebra $\mathfrak{so}(3) \times \mathbb{R}$. The factor $\mathfrak{so}(3)$ can be identified with the Lie algebra consisting of all skew-symmetric 3×3 matrices

$$\left\{A \in \mathbb{R}^{3\times 3} : A^t = -A\right\},$$

endowed with the usual Lie bracket $[A,B] := AB - BA$. This vector space has a basis K_1, K_2, K_3 satisfying

$$[K_1, K_2] = K_3, \; [K_2, K_3] = K_1, \; [K_3, K_1] = K_2.$$

In fact, one can choose

$$K_1 := \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \; K_2 := \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \; K_3 := \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Now, identifying also $\mathfrak{so}(3) \times \mathbb{R}$ with the left-invariant vector fields over $\mathbb{G}$, we define

$$X_1 := (K_1, 1), \quad X_2 := (K_1 + K_2, 2).$$

Declaring that X_1, X_2 form an orthonormal basis of the smooth distribution $\mathcal{D} := \langle X_1, X_2 \rangle$ makes $\mathbb{G}$ a sub-Riemannian manifold.

Notice that

$$[X_1, X_2] = (K_3, 0), \; [X_1, [X_1, X_2]] = -(K_2, 0), \; [X_2, [X_1, X_2]] = (K_1 - K_2, 0) = 2X_1 - X_2.$$

We deduce that $X_1, X_2, [X_1, X_2], [X_1, [X_1, X_2]]$ are everywhere linearly independent (by their left invariance, it suffices to check that they are linearly independent as elements of $\mathfrak{so}(3) \times \mathbb{R}$). We also notice that $[X_2, [X_1, X_2]] \in \mathcal{D} \backslash \langle X_1 \rangle$ (pointwise): this is the fundamental property of this distribution which will enable us to prove the strict abnormality of the curve defined below.

Let γ be any integral curve for X_2 and, more precisely, let $\gamma : [0,1] \to \mathbb{G}$ satisfy $\gamma(0) = (I, 0)$ and $\dot{\gamma} = X_2$ on $[0,1]$. γ cannot be a normal extremal: its associated control is $u \equiv (0,1)$, so if (γ, λ) were a normal biextremal we would have $\langle \lambda, X_1 \rangle = -u_1 \equiv 0$ on $[0,1]$. Arguing as in the proof of Theorem 4.14, we would also have

$$\langle \lambda, [X_2, X_1] \rangle \equiv 0, \quad \langle \lambda, [X_2, [X_2, X_1]] \rangle \equiv 0.$$

But $[X_2, [X_2, X_1]] = X_2 - 2X_1$, so we should have $\langle \lambda, X_2 \rangle \equiv 0$, as well. This would contradict the fact that $\langle \lambda, X_2 \rangle = -u_2 \equiv -1$.

We now prove that (γ, λ) is a nice abnormal biextremal, for some suitable dual curve λ: choose any $\overline{\lambda} \in T_{\gamma(1)}\mathbb{G} \setminus \{0\}$ satisfying

$$\left\langle \overline{\lambda}, X_1(\gamma(1)) \right\rangle = \left\langle \overline{\lambda}, X_2(\gamma(1)) \right\rangle = \left\langle \overline{\lambda}, [X_1, X_2](\gamma(1)) \right\rangle = 0$$

and define $\lambda(t) := \Phi_{1-t}(X_2)^* \overline{\lambda}$. Let us check that $\langle \lambda, X_1 \rangle \equiv \langle \lambda, X_2 \rangle \equiv 0$. As before, we have

$$\frac{d}{dt} \langle \lambda, X_2 \rangle = \langle \lambda, [X_2, X_2] \rangle \equiv 0.$$

Moreover,

$$\frac{d}{dt} \langle \lambda, X_1 \rangle = \langle \lambda, [X_2, X_1] \rangle, \; \frac{d}{dt} \langle \lambda, [X_2, X_1] \rangle = [X_2, [X_2, X_1]] = \langle \lambda, X_2 - 2X_1 \rangle = -2 \langle \lambda, X_1 \rangle.$$

As all the functions that we are differentiating vanish at $t = 1$, we deduce that they vanish on $[0,1]$, as well. So (γ, λ) is an abnormal extremal. Since $X_1, X_2, [X_1, X_2], [X_1, [X_1, X_2]]$ form a basis of the tangent space at every point, $\lambda(t) \neq 0$ implies

$$\langle \lambda(t), [X_1, [X_1, X_2]](\gamma(t)) \rangle \neq 0$$

for any $t \in [0,1]$, proving that (γ, λ) is a nice abnormal biextremal. Finally, Theorem 4.19 tells us that any sufficiently short initial piece of γ is a strictly abnormal length minimizer.

4.6. Golé-Karidi's example

We now give an example of a strictly abnormal minimizer in a Carnot group $\mathbb{G}$ with $n = 6$, $r = 2$, $s = 4$. Let $\mathfrak{g}$ be the stratified Lie algebra

$$\mathfrak{g} := \langle X_1, X_2 \rangle \oplus \langle X_3 \rangle \oplus \langle X_4, X_5 \rangle \oplus \langle X_6 \rangle ,$$

where the Lie bracket is given by

$$[X_1, X_2] := X_3,\ [X_1, X_3] := X_4,\ [X_2, X_3] := X_5,\ [X_1, X_4] := X_6$$

(and $[X_1, X_5], [X_2, X_4], [X_2, X_5] := 0$). The fact that Jacobi's identity holds can be checked directly on triples $\{X_i, X_j, X_k\}$ of distinct elements of the basis: we can assume that the sum of their degrees is at most 4 (as otherwise Jacobi's identity is trivially satisfied), so the only triple that has to be checked is $\{X_1, X_2, X_3\}$. But

$$[X_1, [X_2, X_3]] + [X_2, [X_3, X_1]] + [X_3, [X_1, X_2]] = [X_1, X_5] - [X_2, X_4] + [X_3, X_3] = 0.$$

So $\mathfrak{g}$ is indeed a stratified Lie algebra. Let $\mathbb{G}$ be the Carnot group associated to $\mathfrak{g}$.

A unit-speed abnormal biextremal (γ, λ) with control $u = (u_1, u_2)$ has to satisfy the following equations (see (2.3) and (3.4)):

$$\begin{cases} \dot\lambda_1 = -u_2\lambda_3 \\ \dot\lambda_2 = u_1\lambda_3 \\ \dot\lambda_3 = u_1\lambda_4 + u_2\lambda_5 \\ \dot\lambda_4 = u_1\lambda_6 \\ \dot\lambda_5 = 0 \\ \dot\lambda_6 = 0, \end{cases} \tag{4.7}$$

together with $\lambda_1 = \lambda_2 = 0$ and $u_1^2 + u_2^2 = 1$. Notice that the first two equations imply that $\lambda_3 = 0$ as well (we already knew this by Theorem 4.14). Let $f : \mathbb{R} \to \mathbb{R}$ solve

$$\dot f(t) = \frac{1}{\sqrt{1 + f^2(t)}},$$

with $f(0) = 0$ (a global solution exists since the right-hand side is bounded). If we set

$$u_1 := \frac{1}{\sqrt{1+f^2}}, \quad u_2 := -\frac{f}{\sqrt{1+f^2}}, \quad \lambda(t) := (0, 0, 0, f, 1, 1)$$

for $t \in [0, 1]$ (for instance), it is immediate to check that all equations are satisfied. One can find this particular solution by imposing $\lambda_5 = \lambda_6 = 1$: then one gets $u_1 = \dot\lambda_4$, $u_2 = -u_1\lambda_4$ and the unit-speed constraint implies $\dot\lambda_4^2(1 + \lambda_4^2) = 1$.

Let $\gamma : [0, 1] \to \mathbb{G}$ be the curve associated to the control u, with $\gamma(0) = 0$. Since $\lambda_5 = 1 \neq 0$, γ is a smooth nice abnormal biextremal, Theorem 4.19 guarantees that any sufficiently short initial piece $\gamma|_{[0,T]}$ is a length minimizer.

We are left to check that $\gamma|_{[0,T]}$ is a strictly abnormal extremal. Assume by contradiction that it is also normal: then there exists a dual curve $\mu : [0, T] \to \mathbb{R}^6$ which satisfies the same set of equations (4.7), which we rewrite for the reader's convenience:

$$\begin{cases} \dot\mu_1 = -u_2\mu_3 \\ \dot\mu_2 = u_1\mu_3 \\ \dot\mu_3 = u_1\mu_4 + u_2\mu_5 \\ \dot\mu_4 = u_1\mu_6 \\ \dot\mu_5 = 0 \\ \dot\mu_6 = 0, \end{cases}$$

together with $\mu_1 = -u_1$ and $\mu_2 = -u_2$.

In order to derive a contradiction, we are going to express all the variables in terms of u_1 and λ_4. We will obtain that u_1 has to be constant, which is clearly a contradiction. First of all, we notice that $u_2 = -\lambda_4 u_1$. Moreover, from the second equation we deduce

$$\mu_3 = \frac{\dot{\mu}_2}{u_1} = -\frac{\dot{u}_2}{u_1} = \frac{1}{u_1}\left(\frac{1}{1+\lambda_4^2} - \frac{\lambda_4^2}{(1+\lambda_4^2)^2}\right) = u_1^3.$$

So, inserting this into the third equation,

$$3u_1^2\dot{u}_1 = u_1(\mu_4 - \lambda_4\mu_5).$$

Thus (as $u_1 > 0$) $3u_1\dot{u}_1 = \mu_4 - \lambda_4\mu_5$, which gives, together with the fourth equation,

$$\mu_6 u_1 = \dot{\mu}_4 = 3\left(\dot{u}_1\right)^2 + 3u_1\ddot{u}_1 + \mu_5 u_1. \tag{4.8}$$

Now let us compute $\dot{u}_1$ and $\ddot{u}_1$:

$$\dot{u}_1 = -\lambda_4 u_1^4$$

(since $\dot{\lambda}_4 = u_1$), while

$$\ddot{u}_1 = -u_1^5 - 4\lambda_4 u_1^3\dot{u}_1 = -u_1^5 + 4\lambda_4^2 u_1^7.$$

Substituting into (4.8), we arrive at

$$(\mu_6 - \mu_5)u_1 = 15\lambda_4^2 u_1^8 - 3u_1^6$$

and finally, taking into account that $\lambda_4^2 = u_1^{-2} - 1$,

$$(\mu_6 - \mu_5)u_1 = 15u_1^6 - 15u_1^8 - 3u_1^6.$$

This equation forces u_1 to assume finitely many values, because μ_5 and μ_6 are constant. By continuity u_1 has to be constant, which contradicts the fact that λ_4 is strictly increasing on $[0, T]$.

CHAPTER 5

New regularity results

In 2015, Hakavuori and Le Donne obtained the following theorem, which constitutes the first general regularity result for sub-Riemannian geodesics.

THEOREM 5.1 (Hakavuori-Le Donne). Given any sub-Riemannian manifold M and any constant-speed length minimizer $\gamma : [0,1] \to M$, γ cannot have *corner-like singularities*, i.e. we must have $\dot{\gamma}_-(t_0) = \dot{\gamma}_+(t_0)$ for any $t_0 \in (0,1)$ such that the left and right derivatives $\dot{\gamma}_-(t_0), \dot{\gamma}_+(t_0)$ exist.

Its proof, contained in [HL16], is based on a blow-up argument and a clever cut-and-adjust technique (which was first devised in a similar form in the work [LM08], by Leonardi and Monti).

The aim of this chapter is to provide an overview of the proof of this theorem and to improve it in the context of Carnot groups with rank 2. By following a quantitative approach, we will obtain two interesting refinements of Theorem 5.1, whose statements are given after the following definition.

DEFINITION 5.2. Given a Carnot group $\mathbb{G}$ with rank $r = 2$, for any horizontal curve $\gamma : [a,b] \to \mathbb{G}$ we define $\underline{\gamma} : [a,b] \to V_1$ by $\underline{\gamma} := \pi \circ \gamma$, where π is given by Lemma 3.22. Moreover, we define the *excess* of γ on any subinterval $I \subseteq [a,b]$ to be the quantity

$$\operatorname{Exc}(\gamma, I) := \left(\fint_I \left| \dot{\underline{\gamma}}(t) - \left(\fint_I \dot{\underline{\gamma}}(s)\, ds \right) \right|^2 dt \right)^{1/2}.$$

In probabilistic terms, $\operatorname{Exc}(\gamma, I)$ is the square root of the variance of $\dot{\underline{\gamma}}$ on I. This quantity (or, to be precise, its multidimensional analogue) is ubiquitous in the regularity theory for minimal surfaces and for elliptic PDEs.

THEOREM 5.3 (small excess on arbitrarily small scales, one-sided version). Given a constant-speed length minimizer $\gamma : [0,T] \to \mathbb{G}$, there exists a sequence of scales $\eta_i \downarrow 0$ (depending on γ) such that

$$\operatorname{Exc}(\gamma, [0, \eta_i]) \to 0.$$

THEOREM 5.4 (small excess on arbitrarily small scales, two-sided version). Given a constant-speed length minimizer $\gamma : [-T,T] \to \mathbb{G}$, there exists a sequence of scales $\eta_i \downarrow 0$ (depending on γ) such that

$$\operatorname{Exc}(\gamma, [-\eta_i, \eta_i]) \to 0.$$

REMARK 5.5. We notice that, in the special case where $M = \mathbb{G}$ is a Carnot group with rank 2, Theorem 5.4 improves Theorem 5.1.

Proof of the remark. Indeed, for any $t_0 \in (0,1)$, Theorem 5.4 can be applied to the curve

$$\alpha : [-1,1] \to \mathbb{G}, \quad \alpha(t) := \gamma(t_0)^{-1}\gamma(t_0 + \epsilon t),$$

with $\epsilon := t_0 \wedge (1 - t_0)$, which is still a length minimizer. Assuming (up to subsequences) that $\fint_{[-\eta_i,\eta_i]} \underline{\dot\alpha} \to v \in \mathbb{R}^2$, we obtain that $\underline{\dot\alpha_i} \to v$ in $L^2([-1,1])$, where

$$\alpha_i : [-1,1] \to \mathbb{G}, \quad \alpha_i(t) := \delta_{1/\eta_i}(\delta(\eta_i t))$$

and $\delta_\lambda : \mathbb{G} \to \mathbb{G}$ denotes the intrinsic dilation by the factor $\lambda > 0$, since $\underline{\dot\alpha_i}(t) = \underline{\dot\alpha}(\eta_i t)$. So, recalling Lemma 3.26, the control associated to α_i tends to the constant $\overline{u}$, given by

$$v = \sum_{i=1}^{r} \overline{u}_i X_i,$$

and, since $\alpha_i(0) = e$, this implies that α_i tends uniformly on $[-1,1]$ to the line $t \mapsto \exp(tv)$.

In particular, α cannot have different left and right derivatives at 0 (provided that they exist): indeed, assume by contradiction that $\dot\alpha_-(0) \neq \dot\alpha_+(0)$. Since α has constant speed, from Lemma 5.21 below we get

$$\pi_k \circ \alpha(t) = O(|t|^k), \ \forall k = 1, \dots, s$$

(see Section 3.3 for the notation). Thus, since $\overline{\pi}_k = d(\pi_k)_e$, we get $\overline{\pi}_k(\dot\alpha_\pm(0)) = 0$ for $k > 1$, i.e. $\dot\alpha_\pm(0) \in V_1$. But then

$$\underline{\alpha}(t) = \mathbf{1}_{\{t \geq 0\}}\dot\alpha_+(0)t + \mathbf{1}_{\{t<0\}}\dot\alpha_-(0)t + o(t),$$

which gives

$$\underline{\alpha_i} \to \mathbf{1}_{\{t \geq 0\}}\dot\alpha_+(0)t + \mathbf{1}_{\{t<0\}}\dot\alpha_-(0)t,$$

uniformly on $[-1,1]$, thanks to the identity $\pi \circ \delta_{1/\eta_i} = \frac{1}{\eta_i}\pi$. This clearly contradicts the fact that $\underline{\alpha_i}(t) \to tv$ uniformly on $[-1,1]$. $\square$

5.1. Cut, correction devices and preliminary remarks

In the next section, we will give a different presentation of the proof of Theorem 5.1, with respect to the one contained in [HL16]. Our presentation is based on the subsequent application of suitable *correction devices* (which appear in a less explicit way in the original paper), which we are going to define in this section. This will require a certain amount of notation, but is better suited for the proof of our refinements, namely Theorems 5.3 and 5.4.

In this section, we will work in a generic Carnot group (whose rank is not necessarily 2). We will use the notation and the facts contained in Section 3.3, as well as the notation introduced in Definition 2.28.

To begin with, let us choose, for any $Y \in \mathfrak{g}$, a unit-speed geodesic $\delta_Y : [0, \ell_Y] \to \mathbb{G}$ from e to $\exp(Y)$ (so that $\ell_Y = d(e, \exp(Y))$): its existence is guaranteed by Proposition 3.13. Let us call u_Y the associated control in $H^1([0,\ell_Y], \mathbb{R}^r)$.

Definition 5.6. Given two curves $\alpha \in H^1([a, a+a'], \mathbb{G})$ and $\beta \in H^1([b, b+b'], \mathbb{G})$, let us define their *join* $\alpha * \beta \in H^1([a, a+(a'+b')], \mathbb{G})$ by the formula

$$\alpha * \beta : [a, a+(a'+b')] \to \mathbb{G}, \quad \alpha * \beta(t) := \begin{cases} \alpha(t) & t \leq a + a' \\ \alpha(a+a')\beta(b)^{-1}\beta(t + b - (a+a')) & t \geq a + a'. \end{cases}$$

It will be useful to allow the controls to be defined on arbitrary compact intervals $[a, a + a'] \subseteq \mathbb{R}$. We can define the join of two controls in a similar way (generalizing Definition 2.28).

DEFINITION 5.7. Given two controls $u \in L^2([a, a+a'], \mathbb{R}^r)$ and $u' \in L^2([b, b+b'], \mathbb{R}^r)$, we define their *join*

$$u * u' \in L^2([a, a + (a' + b')], \mathbb{R}^r), \quad u * u'(t) := \begin{cases} u(t) & t \leq a + a' \\ u(t + b - (a + a')) & t > a + a'. \end{cases}$$

REMARK 5.8. Notice that $\alpha * \beta$ is continuous. It is horizontal iff so are α and β. Moreover, calling u and u' the controls associated to α and β, $u * u'$ is precisely the control associated to $\alpha * \beta$.

DEFINITION 5.9. Let $\gamma : [a, b] \to \mathbb{G}$ be a unit-speed horizontal curve, with control $u \in L^2([a, b], \mathbb{R}^r)$, and let $[s, s'] \subseteq [a, b]$ be any subinterval. We choose any $w \in V_1$, $|w| = 1$ such that $\langle w, \underline{\gamma}(s') - \underline{\gamma}(s) \rangle = |\underline{\gamma}(s') - \underline{\gamma}(s)|$ and we let $v : [0, |\underline{\gamma}(s') - \underline{\gamma}(s)|] \to \mathbb{R}^2$ be the constant control $v(t) := (w_1, w_2)$ (where $w = w_1 X_1 + w_2 X_2$). We define the *cutted curve* $\mathrm{Cut}(\gamma, [s, s'])$ to be the curve associated to the control

$$u|_{[a,s]} * v * u|_{[s',b]}$$

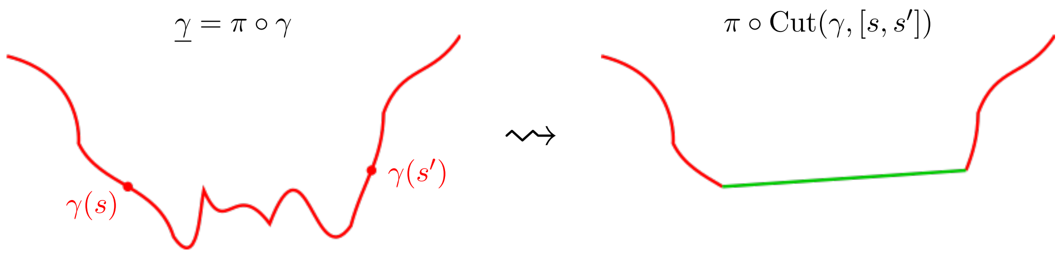
and the same starting point as γ. Equivalently,

$$\mathrm{Cut}(\gamma, [s, s']) := \gamma|_{[a,s]} * \exp(\cdot w)|_{[0, |\underline{\gamma}(s') - \underline{\gamma}(s)|]} * \gamma|_{[s',b]}.$$

Notice that w is uniquely determined if $\underline{\gamma}(s') \neq \underline{\gamma}(s)$, while if $\underline{\gamma}(s') = \underline{\gamma}(s)$ the cutted curve is still well-defined and is simply given by

$$\mathrm{Cut}(\gamma, [s, s']) = \gamma|_{[a,s]} * \gamma|_{[s',b]}.$$

The following picture shows the effect of the cut operation (looking only at the projection on the first layer V_1).



REMARK 5.10. We observe that $\mathrm{Cut}(\gamma, [s, s'])$ is still unit-speed and horizontal. Moreover,

$$L\left(\mathrm{Cut}(\gamma, [s, s'])\right) = L(\gamma) - (s' - s) + |\underline{\gamma}(s') - \underline{\gamma}(s)|$$

and the domain of $\mathrm{Cut}(\gamma, [s, s'])$ is $[a, b - (s' - s) + |\underline{\gamma}(s') - \underline{\gamma}(s)|]$. In particular, as $|\underline{\gamma}(s') - \underline{\gamma}(s)| \leq \int_s^{s'} |\dot{\underline{\gamma}}| \, d\mathcal{L}^1 = s' - s$, we always have $L(\mathrm{Cut}(\gamma, [s, s'])) \leq L(\gamma)$. Assuming $s < s'$, this inequality is strict unless $\dot{\underline{\gamma}}$ is constant on $[s, s']$ (in which case $\gamma(s + t) = \gamma(s) \exp(tv)$ for $t \in [0, |\underline{\gamma}(s') - \underline{\gamma}(s)|]$). Lemma 5.26 provides a quantitative strengthening of this assertion.

Remark 5.11. The final point of the cutted curve has the same projection on V_1 as the final point of γ, i.e.

$$\pi\left(\mathrm{Cut}(\gamma,[s,s'])\left(\tilde{b}\right)\right)=\pi(\gamma(b)),$$

where $\tilde{b}:=b-(s'-s)+\left|\underline{\gamma}(s')-\underline{\gamma}(s)\right|$: indeed, using the formula

$$\mathrm{Cut}(\gamma,[s,s'])\left(\tilde{b}\right)=\gamma(s)\exp\left(\left|\underline{\gamma}(s')-\underline{\gamma}(s)\right|w\right)\gamma(s')^{-1}\gamma(b)$$

and Lemma 3.22, we get

$$\begin{aligned}\pi\left(\mathrm{Cut}(\gamma,[s,s'])\left(\tilde{b}\right)\right)&=\underline{\gamma}(s)+\left|\underline{\gamma}(s')-\underline{\gamma}(s)\right|w+\left(\underline{\gamma}(b)-\underline{\gamma}(s')\right)\\&=\underline{\gamma}(s)+\left(\underline{\gamma}(s')-\underline{\gamma}(s)\right)+\left(\underline{\gamma}(b)-\underline{\gamma}(s')\right)\\&=\underline{\gamma}(b).\end{aligned}$$

Definition 5.12. Given a unit-speed horizontal curve $\gamma:[a,b]\to\mathbb{G}$, with control $u\in L^2([a,b],\mathbb{R}^r)$, any subinterval $[s,s']\subseteq[a,b]$ and any $Y\in\mathfrak{g}$, we define the *corrected curve* $\mathrm{Dev}(\gamma,[s,s'],Y):[a,b+2\ell_Y]\to\mathbb{G}$ to be the curve associated to the control

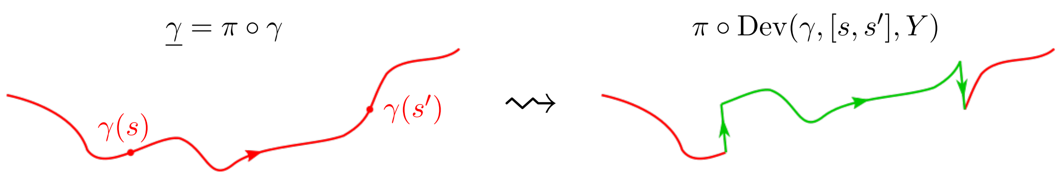
$$u|_{[a,s]}*u_Y*u|_{[s,s']}*\check{u}_Y*u|_{[s',b]},$$

with the same starting point as γ (i.e. $\gamma(a)$). Equivalently,

$$\mathrm{Dev}(\gamma,[s,s'],Y):=\gamma|_{[a,s]}*\delta_Y*\gamma|_{[s,s']}*\delta_Y(\ell_Y-\cdot)*\gamma|_{[s',b]}.$$

Here $\delta_Y(\ell_Y-\cdot)$ denotes δ_Y traveled backwards. We will refer to the process of transforming γ into $\mathrm{Dev}(\gamma,[s,s'],Y)$ as the application of the *correction device* associated to $[s,s']$ and Y.

The following picture shows the appearance of $\pi\circ\mathrm{Dev}(\gamma,[s,s'],Y)$ when $Y\in V_1$ (in which case $\underline{\delta_Y}$ is a straight segment).



Let us compute the displacement of the final point when we apply a correction device.

Definition 5.13. We will use the compact notation $\gamma|_a^b:=\gamma(a)^{-1}\gamma(b)$.

Lemma 5.14. Setting $b':=b+2\ell_Y$, the displacement $\gamma(b)^{-1}\mathrm{Dev}(\gamma,[s,s'],Y)(b')$ is given by the formula

$$\gamma(b)^{-1}\mathrm{Dev}(\gamma,[s,s'],Y)(b')=C_{\gamma|_b^s}\left(\left[\exp(Y),\gamma|_s^{s'}\right]\right),\tag{5.1}$$

where $C_g(h):=ghg^{-1}$ denotes the conjugation by g and $[g,h]:=ghg^{-1}h^{-1}$ is the commutator in $\mathbb{G}$, which should not be confused with the Lie bracket in $\mathfrak{g}$.

Proof. Indeed, it holds

$$\begin{aligned}\mathrm{Dev}(\gamma,[s,s'],Y)(b') &= \gamma(s)\exp(Y)\,\gamma|_s^{s'}\exp(-Y)\,\gamma|_{s'}^{b}\\ &= \gamma(s)\left[\exp(Y),\gamma|_s^{s'}\right]\gamma|_s^{s'}\,\gamma|_{s'}^{b}\\ &= \gamma(s)\left[\exp(Y),\gamma|_s^{s'}\right]\gamma|_s^{b},\end{aligned}$$

from which we deduce, using the identities $\gamma(b)^{-1}\gamma(s)=\gamma|_b^s$ and $\gamma|_s^b=(\gamma|_b^s)^{-1}$,

$$\begin{aligned}\gamma(b)^{-1}\,\mathrm{Dev}(\gamma,[s,s'],Y)(b') &= \gamma|_b^s\left[\exp(Y),\gamma|_s^{s'}\right](\gamma|_b^s)^{-1}\\ &= C_{\gamma|_b^s}\left(\left[\exp(Y),\gamma|_s^{s'}\right]\right).\end{aligned}$$

□

We now state and prove two useful lemmas, which tell us how the homomorphisms π_j behave when dealing with conjugations and commutators.

Lemma 5.15. $\mathbb{G}_j$ is a normal subgroup: more precisely, if $g\in\mathbb{G}$ and $h\in\mathbb{G}_j$, then

$$ghg^{-1}\in\mathbb{G}_j,\quad \pi_j(ghg^{-1})=\pi_j(g).$$

Proof. Writing $g=\exp(X)$ and $h=\exp(Y)$, recall the general formula for Lie groups

$$C_g(h)=C_g(\exp(Y))=\exp(\mathrm{Ad}(g)Y),\quad \mathrm{Ad}(g)=\mathrm{Ad}(\exp(X))=e^{\mathrm{ad}(X)},$$

where $C_g(x):=gxg^{-1}$ denotes the conjugation by g and is an automorphism of $\mathbb{G}$, while $\mathrm{Ad}:=d(C_g)_e$ is the corresponding automorphism of $\mathfrak{g}$. Combining them, we have

$$\exp^{-1}(ghg^{-1})=\mathrm{Ad}(g)Y=e^{\mathrm{ad}\,X}Y=\sum_{k=0}^{\infty}\frac{(\mathrm{ad}\,X)^k}{k!}Y=Y+R,$$

where $R\in W_{j+1}$, since all the terms with $k\geq 1$ belong to W_{j+1} (as $Y\in W_j$). Hence, $ghg^{-1}\in\mathbb{G}_j$ and

$$\pi_j(ghg^{-1})=\overline{\pi}_j\circ\exp^{-1}(ghg^{-1})=\overline{\pi}_j(Y+R)=\overline{\pi}_j(Y)=\pi_j(h).$$

□

Lemma 5.16. Fix some $1\leq j<s$. If $g\in\mathbb{G}$ and $h\in\mathbb{G}_j$, then

$$[g,h]:=ghg^{-1}h^{-1}\in\mathbb{G}_{j+1},\quad \pi_{j+1}([g,h])=[\pi(g),\pi_j(h)].$$

Similarly, if $g\in\mathbb{G}_j$ and $h\in\mathbb{G}$, we have

$$[g,h]:=ghg^{-1}h^{-1}\in\mathbb{G}_{j+1},\quad \pi_{j+1}([g,h])=[\pi_j(g),\pi(h)].$$

Proof. Combining Lemma 5.15 with Lemma 3.24, we obtain $[g,h]=(ghg^{-1})h^{-1}\in\mathbb{G}_j$ and

$$\pi_j([g,h])=\pi_j(ghg^{-1})+\pi_j(h^{-1})=\pi_j(h)-\pi_j(h)=0,$$

so that $[g,h]\in\mathbb{G}_{j+1}$. Now, writing $g=\exp(X)$, $h=\exp(Y)$ and using the formula $\exp^{-1}(ghg^{-1})=e^{\mathrm{ad}\,X}Y$ as in the previous proof, we obtain

$$\exp^{-1}(ghg^{-1})=\sum_{k=0}^{\infty}\frac{(\mathrm{ad}\,X)^k}{k!}Y=Y+[X,Y]+R',$$

where the remainder R' is the sum of all terms with $k \geq 2$ and thus belongs to W_{j+2}. As $h^{-1} = \exp(-Y)$, the Baker-Campbell-Hausdorff formula gives

$$\exp^{-1}([g,h]) = P(Y + [X,Y] + R', -Y) = [X,Y] + R'',$$

where R'' is given by the double sum in (3.2). Now, thinking each term of the double sum as a $(k_1 + \ell_1 + \cdots + k_p + \ell_p + 1)$-multilinear function (and expanding each factor containing $Y + [X,Y] + R'$ accordingly), we obtain that R'' is a linear combination of elements of the form

$$(\operatorname{ad} Z_1) \cdots (\operatorname{ad} Z_k) Z_{k+1},$$

where $k \geq 1$ and $Z_i \in \{Y, [X,Y], R'\}$. Those elements where only Y appears vanish, while the other terms belong to W_{j+2} (since $[X,Y], R' \in W_{j+1}$ and $k \geq 1$). We deduce that $R'' \in W_{j+2}$. Finally,

$$\pi_{j+1}([g,h]) = \overline{\pi}_{j+1} \circ \exp^{-1}([g,h]) = \overline{\pi}_{j+1}([X,Y] + R'') = \overline{\pi}_{j+1}([X,Y]) = [\overline{\pi}(X), \overline{\pi}_j(Y)],$$

since $X = \overline{\pi}(X) + R_X$ and $Y = \overline{\pi}_j(Y) + R_Y$, with $R_X \in W_2$ and $R_Y \in W_{j+1}$.

The second part of the thesis follows from the first one, using the identity $[g,h] = [h,g]^{-1}$. □

COROLLARY 5.17. If $Y \in V_j$, then $\gamma(b)^{-1} \operatorname{Dev}(\gamma, [s,s'], Y)(b') \in \mathbb{G}_{j+1}$ and

$$\pi_{j+1}\left(\gamma(b)^{-1} \operatorname{Dev}(\gamma, [s,s'], Y)(b')\right) = [Y, \underline{\gamma}(s') - \underline{\gamma}(s)].$$

Proof. By Lemma 3.22 we have

$$\pi\left(\gamma|_s^{s'}\right) = \pi(\gamma(s)^{-1}\gamma(s')) = \pi(\gamma(s')) - \pi(\gamma(s)) = \underline{\gamma}(s') - \underline{\gamma}(s).$$

Moreover, $\pi_j(\exp(Y)) = Y$. Hence, using Lemma 5.16, we obtain

$$\left[\exp(Y), \gamma|_s^{s'}\right] \in G_{j+1}, \quad \pi_{j+1}\left(\left[\exp(Y), \gamma|_s^{s'}\right]\right) = [Y, \underline{\gamma}(s') - \underline{\gamma}(s)].$$

The thesis now follows from equation (5.1) and Lemma 5.15. □

We conclude this section by introducing some other useful notation.

DEFINITION 5.18. When dealing with curves γ defined on symmetric intervals, it is convenient to use modified versions of Cut and Dev, namely $\operatorname{Cut}'(\gamma, [s,s'])$ and $\operatorname{Dev}'(\gamma, [s,s'], Y)$: these new curves are simply obtained from $\operatorname{Cut}(\gamma, [s,s'])$ and $\operatorname{Dev}(\gamma, [s,s'], Y)$ by precomposing them with a time translation, in such a way that their new domain is still a symmetric interval. Thus, for instance, if $\gamma : [-1,1] \to \mathbb{G}$ satisfies $\underline{\gamma}(0) = \underline{\gamma}\left(\frac{2}{3}\right)$, we have

$$\operatorname{Cut}'\left(\gamma, \left[0, \frac{2}{3}\right]\right) : \left[-\frac{2}{3}, \frac{2}{3}\right] \to \mathbb{G}, \quad \operatorname{Cut}'\left(\gamma, \left[0, \frac{2}{3}\right]\right) = \operatorname{Cut}\left(\gamma, \left[0, \frac{2}{3}\right]\right)\left(\cdot - \frac{1}{3}\right).$$

DEFINITION 5.19. Given a unit-speed $\gamma : I \to \mathbb{G}$, two subintervals $[s,s'], [t,t'] \subseteq I$ with $s' \leq t$ and two elements Y, Y', we use the compact notation

$$\operatorname{Dev}(\gamma, [s,s'], Y, [t,t'], Y') := \operatorname{Dev}\left(\operatorname{Dev}(\gamma, [s,s'], Y), [t + 2\ell_Y, t' + 2\ell_Y], Y'\right)$$

for the subsequent application of two devices, as well as its symmetric counterpart

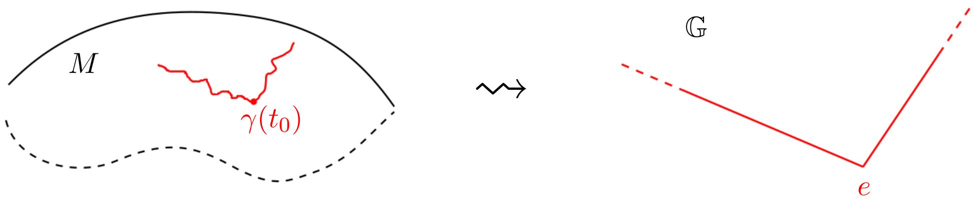
$$\begin{aligned}\operatorname{Dev}'(\gamma, [s,s'], Y, [t,t'], Y') &:= \operatorname{Dev}'\left(\operatorname{Dev}(\gamma, [s,s'], Y), [t + 2\ell_Y, t' + 2\ell_Y], Y'\right) \\ &= \operatorname{Dev}'\left(\operatorname{Dev}'(\gamma, [s,s'], Y), [t + \ell_Y, t' + \ell_Y], Y'\right).\end{aligned}$$

5.2. Overview of Hakavuori-Le Donne's argument

In this section we prove Theorem 5.1. Let us assume by contradiction that, for some $t_0 \in (0,1)$, $\dot\gamma_-(t_0), \dot\gamma_+(t_0)$ exist and $\dot\gamma_-(t_0) \neq \dot\gamma_+(t_0)$. By a desingularization technique (which is presented, for instance, in [Jea14, Section 2.4]) we can assume that $\mathcal{D}$ is equiregular at $\gamma(t_0)$. Now, as was already pointed out in [LM08], by a blow-up argument (which essentially uses Theorem 3.34 at $\gamma(t_0)$), we obtain a Carnot group $\mathbb{G}$ and a curve $\alpha : \mathbb{R} \to \mathbb{G}$ which is still a length minimizer between any couple of its points and has the form

$$\alpha(t) = \begin{cases} \exp(tY_1) & t \le 0 \\ \exp(tY_2) & t \ge 0, \end{cases}$$

with $Y_1, Y_2 \in V_1$, $|Y_1| = |Y_2| > 0$ and $Y_1 \neq Y_2$.



Thus, to derive a contradiction, it suffices to prove that (for any $\mathbb{G}$ and any such curve α) $\alpha|_{[-1,1]}$ is not a length minimizer. By rescaling time, we reduce to the case that $|Y_1| = |Y_2| = 1$. We can clearly assume that Y_1, Y_2 are linearly independent, since otherwise $Y_1 = -Y_2$ and our claim is trivial. Moreover, we can also assume that $\mathbb{G}$ has rank 2: indeed, as $\alpha(\mathbb{R}) \subseteq \exp(\langle Y_1, Y_2 \rangle)$, we can replace $\mathbb{G}$ with $\mathbb{G}' := \exp(\mathfrak{g}')$, where $\mathfrak{g}'$ is the (graded) Lie subalgebra generated by Y_1, Y_2; $\mathbb{G}'$ is a Carnot subgroup (see the argument used in Step 4 in the proof of Theorem 3.43).

The main contribution of [HL16] consists exactly in the proof of the non-minimality of $\alpha|_{[-1,1]}$ under these hypotheses. This is achieved by cutting the curve $\alpha|_{[-1,1]}$ (in such a way that the projection on V_1 becomes a straight segment), obtaining a shorter curve β_1 with the same starting point, but with a final point which is only guaranteed to have the same projection on V_1 as $\alpha(1)$. Then one replaces β_1 with suitable corrected curves β_k, whose final point y_k satisfies $y_k \in \alpha(1)\mathbb{G}_{k+1}$. A crucial fact is that the error $\pi_{k+1}(\alpha(1)^{-1}y_k)$ on the layer V_{k+1} and the extra length needed to correct it scale with different powers of r under the dilation δ_r: this is exploited in order to guarantee that all the corrected curves are shorter than $\alpha|_{[-1,1]}$.

We now give the full proof of the fact that $\alpha|_{[-1,1]}$ is not a length minimizer, following the above sketch. Set $\beta_1 := \mathrm{Cut}'\left(\alpha|_{[-1,1]}, [-1,1]\right)$ and call $[-T_1, T_1]$ its domain: recalling Definition 5.9, Remark 5.11 and Remark 5.10, we see that β_1 satisfies the hypotheses of the next theorem, with $k = 1$. The non-minimality of $\alpha|_{[-1,1]}$ then follows by applying it repeatedly: we obtain a finite sequence $\beta_1, \dots, \beta_s$; the final curve β_s connects $\alpha(-1)$ to $\alpha(1)$ and has $L(\beta_s) < L\left(\alpha|_{[-1,1]}\right)$.

Theorem 5.20. Fix an integer $1 \le k \le s-1$. If there exists a unit-speed horizontal $\beta_k : [-T_k, T_k] \to \mathbb{G}$ such that

$$\beta_k(-T_k) = \alpha(-1), \quad \alpha(1)^{-1}\beta_k(T_k) \in \mathbb{G}_{k+1}, \quad L(\beta_k) < L\left(\alpha|_{[-1,1]}\right)$$

(the last condition being equivalent to $T_k < 1$, as α is unit-speed), then there exists a unit-speed horizontal $\beta_{k+1} : [-T_{k+1}, T_{k+1}] \to \mathbb{G}$ satisfying

$$\beta_k(-T_{k+1}) = \alpha(-1), \quad \alpha(1)^{-1}\beta_{k+1}(T_{k+1}) \in \mathbb{G}_{k+2}, \quad L(\beta_{k+1}) < L\left(\alpha|_{[-1,1]}\right).$$

We use the convention that $\mathbb{G}_j := \{e\}$ for any $j > s$.

Proof. Set $\epsilon := L\left(\alpha|_{[-1,1]}\right) - L(\beta_1)$ and write $\alpha(1)^{-1}\beta_k(T_k) = \exp(E)$, for some $E \in W_{k+1} = V_{k+1} \oplus \cdots \oplus V_s$. Recalling that $[V_k, V_1] = V_{k+1}$ and that Y_1, Y_2 form a basis of V_1, we can find $Z_1, Z_2 \in V_k$ such that

$$\overline{\pi}_{k+1}(E) = [Z_1, X_1] + [Z_2, X_2].$$

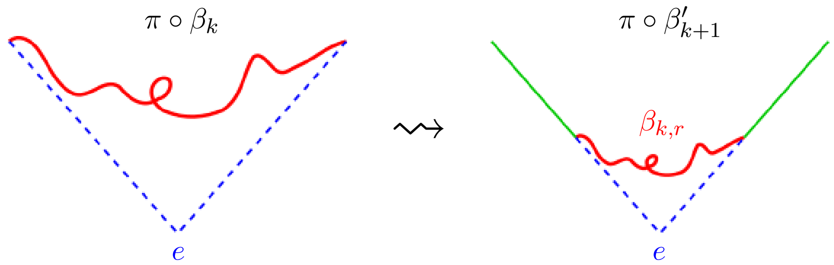
Let $\beta_{k,r} : [-rT_k, rT_k] \to \mathbb{G}$ be the unit-speed horizontal curve given by

$$\beta_{k,r}(t) := \delta_r \circ \beta_k\left(\frac{t}{r}\right)$$

and call β'_{k+1} the reparametrization of

$$\alpha|_{[-1,-r]} * \beta_{k,r} * \alpha|_{[r,1]}$$

obtained by translating the time, in such a way that the domain of β'_{k+1} is a symmetric interval $[-T'_{k+1}, T'_{k+1}]$.



Notice that $\beta'_{k+1}(-T'_{k+1}) = \alpha(-1)$, as well as $\beta'_{k+1}|_{[-rT_k, rT_k]} \equiv \beta_{k,r}$ (since $\alpha(-r) = \beta_{k,r}(-rT_k)$). Moreover,

$$L\left(\alpha|_{[-1,1]}\right) - L(\beta'_{k+1}) = L\left(\alpha|_{[-r,r]}\right) - L(\beta_{k,r}) = \epsilon r.$$

Thus, we obtain

$$\begin{aligned}\alpha(1)^{-1}\beta'_{k+1}(T'_{k+1}) &= \left(\alpha(r)\,\alpha|_r^1\right)^{-1}\left(\beta_{k,r}(rT_k)\,\alpha|_r^1\right)\\ &= C_{\alpha|_1^r}\left(\alpha(r)^{-1}\beta_{k,r}(rT_k)\right)\\ &= C_{\alpha|_1^r}\left(\delta_r \circ \exp(E)\right),\end{aligned}$$

which gives (by Lemma 5.15)

$$\pi_{k+1}\left(\alpha(1)^{-1}\beta'_{k+1}(T'_{k+1})\right) = r^{k+1}\overline{\pi}_{k+1}(E). \tag{5.2}$$

We define

$$\beta_{k+1} := \mathrm{Dev}'\left(\beta'_{k+1}, \left[-T'_{k+1}, -T'_{k+1} + \frac{1}{2}\right], -2r^{k+1}Z_1, \left[T'_{k+1} - \frac{1}{2}, T_{k+1}\right], -2r^{k+1}Z_2\right)$$

and call $[-T_{k+1}, T_{k+1}]$ its symmetric domain. If r is small enough, β_{k+1} is the desired curve: indeed, $\beta_{k+1}(-T_{k+1}) = \alpha(-1)$ and

$$\begin{aligned} L(\beta_{k+1}) &= L(\beta'_{k+1}) + 2\ell_{-2r^{k+1}Z_1} + 2\ell_{-2r^{k+1}Z_2} \\ &= L\left(\alpha|_{[-1,1]}\right) - \epsilon r + O\left(r^{(k+1)/k}\right) \\ &< L\left(\alpha|_{[-1,1]}\right) \end{aligned}$$

when r is small (the estimates $\ell_{-2r^{k+1}Z_1}, \ell_{-2r^{k+1}Z_2} = O\left(r^{(k+1)/k}\right)$ follow from Proposition 3.19). Finally, noticing that

$$\begin{aligned} &\left[-2r^{k+1}Z_1, \underline{\beta'_{k+1}}\left(-T'_{k+1} + \frac{1}{2}\right) - \underline{\beta'_{k+1}}(-T'_{k+1})\right] + \left[-2r^{k+1}Z_2, \underline{\beta'_{k+1}}(T'_{k+1}) - \underline{\beta'_{k+1}}\left(T'_{k+1} - \frac{1}{2}\right)\right] \\ &= \left[-2r^{k+1}Z_1, \underline{\alpha}\left(-\frac{1}{2}\right) - \underline{\alpha}(-1)\right] + \left[-2r^{k+1}Z_2, \underline{\alpha}(1) - \underline{\alpha}\left(\frac{1}{2}\right)\right] \\ &= -r^{k+1}\overline{\pi}_{k+1}(E) \end{aligned}$$

and recalling Corollary 5.17, Lemma 3.24 and (5.2), we obtain

$$\begin{aligned} \pi_{k+1}\left(\alpha(1)^{-1}\beta_{k+1}(T_{k+1})\right) &= \pi_{k+1}\left(\alpha(1)^{-1}\beta'_{k+1}(T'_{k+1})\right) + \pi_{k+1}\left(\beta'_{k+1}(T'_{k+1})^{-1}\beta_{k+1}(T_{k+1})\right) \\ &= r^{k+1}\overline{\pi}_{k+1}(E) - r^{k+1}\overline{\pi}_{k+1}(E) \\ &= 0, \end{aligned}$$

i.e. $\alpha(1)^{-1}\beta_{k+1}(T_{k+1}) \in \mathbb{G}_{k+2}$. □

5.3. A quantitative refinement

It turns out that, in order to correct the error (on the final point) produced by shortening a curve γ taking values in a Carnot group $\mathbb{G}$ of rank 2, all we need is to find, on any subinterval I of the domain of γ, two increments $\underline{\gamma}(b_1) - \underline{\gamma}(a_1)$ and $\underline{\gamma}(b_2) - \underline{\gamma}(a_2)$ which are linearly independent (in a quantitative way) and such that $\left|\underline{\gamma}(b_1) - \underline{\gamma}(a_1)\right|$ and $\left|\underline{\gamma}(b_2) - \underline{\gamma}(a_2)\right|$ are both comparable with $\mathcal{L}^1(I)$. On these two intervals one can then apply the appropriate correction devices.

As a side note, we remark that, since V_2 is 1-dimensional, we just need a single device to correct the error on the layer V_2. To this aim, we only need to know that some $[a, b] \subseteq I$ exists such that $\left|\underline{\gamma}(b) - \underline{\gamma}(a)\right|$ is comparable with $\mathcal{L}^1(I)$ (see Lemma 5.23 below).

Before proving Theorems 5.3 and 5.4, which are the goal of this section, we state and prove some useful lemmas.

Lemma 5.21. If $\gamma \in H^1([0,T], \mathbb{G})$ is a horizontal path, then $\left|\pi_k(\gamma(0)^{-1}\gamma(T))\right| \leq C_k L(\gamma)^k$ (with respect to the inner product on $\mathfrak{g}$ which makes $X_1, \ldots, X_n$ an orthonormal basis), where $C_k > 0$ depends only on k and $\mathbb{G}$.

Proof. We can assume that γ has unit speed and, possibly replacing γ with $L_{\gamma(0)^{-1}} \circ \gamma$, that $\gamma(0) = e$. Recall that, by Proposition 3.18, in exponential coordinates X_1^L and X_2^L have the form:

$$X_i^L(x) = \partial_i + \sum_{j:d(j)>d(i)} f_{ij}(x)\partial_j,$$

where $f_{ij}(x)$ are homogeneous polynomials with weighted degree $d(j)-1$. Let us denote by $\gamma_k(t)$ the k-th component of $\gamma(t)$ in exponential coordinates. We now prove that there exist $C'_1,\dots,C'_n>0$ such that

$$|\gamma_j(t)| \le C'_j t^{d(j)}$$

for any t and any $j=1,\dots,s$, by induction on j: if $j=1,2$ we have

$$|\gamma_j(t)| \le \int_0^t |u_j(\tau)|\, d\tau \le \int_0^t |u(\tau)|\, d\tau = t,$$

where $u=(u_1,u_2)$ is the control of γ. For $j>2$ we have

$$\begin{aligned}
|\gamma_j(t)| &\le \int_0^t \left|u_1(\tau)\left\langle dx^j, X_1\right\rangle(\gamma(\tau)) + u_2(\tau)\left\langle dx^j, X_2\right\rangle(\gamma(\tau))\right|\, d\tau\\
&\le \int_0^t |f_{1j}(\gamma(\tau))|\, d\tau + \int_0^t |f_{2j}(\gamma(\tau))|\, d\tau\\
&\le C''_j \int_0^t \tau^{d(j)-1}\, d\tau + C''_j \int_0^t \tau^{d(j)-1}\, d\tau\\
&= \frac{2C''_j}{d(j)} t^{d(j)}.
\end{aligned}$$

We estimated $|u_1(t)|, |u_2(t)| \le 1$. Moreover, we used the fact that $f_{1j}(\gamma(\tau)), f_{2j}(\gamma(\tau))$ involve only the components $\gamma_i(\tau)$ with $d(i)<d(j)$ (which are already estimated by the inductive hypothesis) and f_{1j}, f_{2j} are homogeneous with weighted degree $d(j)-1$. The thesis follows from the obvious inequality

$$|\pi_k(\gamma(T))| \le \sum_{j:d(j)=k} |\gamma_k(T)|.$$

□

LEMMA 5.22 (compactness of minimizers). Let $\gamma_k:[0,1]\to\mathbb{G}$ be a sequence of unit-speed length minimizers with $\gamma_k(0)=e$. Then there exist a subsequence γ_{k_p} and a unit-speed length minimizer $\gamma_\infty:[0,1]\to\mathbb{G}$ (with $\gamma_\infty(0)=e$) such that $\gamma_{k_p}\to\gamma_\infty$ uniformly and $\underline{\dot\gamma_{k_p}}\to\underline{\dot\gamma_\infty}$ in $L^2([0,1],V_1)$.

Proof. Let us call $u_k=(u_{k,1},u_{k,2})$ the control associated to γ_k. Since, for any k,

$$\gamma_k([0,1]) \subseteq \overline{\mathbb{B}}_1(e),$$

which is compact, and since all the curves γ_k are 1-Lipschitz with respect to the distance d, we can find a subsequence γ_{k_p} converging uniformly to some curve γ_∞.

Since $\|u_{n_k}\|_2=1$, up to further subsequences we can assume $u_{n_k}\rightharpoonup\overline{u}$ in $L^2([0,1],\mathbb{R}^2)$. Identifying $\mathbb{G}$ with $\mathbb{R}^n$ (using the exponential coordinates), this implies that

$$u_{k_p,1}(t)X_1^L(\gamma_k(t)) + u_{k_p,2}(t)X_2^L(\gamma_k(t)) \rightharpoonup \overline{u}_1(t)X_1^L(\gamma_\infty(t)) + \overline{u}_2(t)X_2^L(\gamma_\infty(t))$$

in $L^2([0,1],\mathbb{R}^n)$. Thus, passing to the limit (as $p\to\infty$) in

$$\gamma_{k_p}(t) = \int_0^t \left(u_{k_p,1}(\tau)X_1^L(\gamma_k(\tau)) + u_{k_p,2}(\tau)X_2^L(\gamma_k(\tau))\right)\, d\tau,$$

we obtain

$$\gamma_\infty(t) = \int_0^t \left(\overline{u}_1(\tau)X_1^L(\gamma_\infty(\tau)) + \overline{u}_2(\tau)X_2^L(\gamma_\infty(\tau))\right)\, d\tau.$$

This proves that $\dot{\gamma_\infty}(t) = \overline{u}_1(t)X_1^L(\gamma_\infty(t)) + \overline{u}_2(t)X_2^L(\gamma_\infty(t))$ for a.e. t, so γ_∞ is horizontal with associated control $\overline{u}=:u_\infty$.

Moreover,

$$\|u_\infty\|_2 \geq L(\gamma_\infty) \geq d(\gamma_\infty(0),\gamma_\infty(1)) = \lim_{p\to\infty} d(\gamma_{k_p}(0),\gamma_{k_p}(1)) = 1. \tag{5.3}$$

But we already know that $\|u_\infty\|_2 \leq 1$ (as $u_{k_p} \rightharpoonup u_\infty$ and $\left\|u_{k_p}\right\|_2 \leq 1$), so $\left\|u_{k_p}\right\|_2 \to \|u_\infty\|_2$ and, since $L^2([0,1],\mathbb{R}^2)$ is a Hilbert space, this gives $u_{k_p} \to u_\infty$. Hence, $|u_\infty| = 1$ a.e. and $\underline{\dot\gamma_{k_p}} \to \underline{\dot\gamma_\infty}$, as well (since we have $\underline{\dot\gamma_{k_p}}(s) = u_{k_p,1}(s)X_1 + u_{k_p,2}(s)X_2$ a.e. and the analogous identity for γ_∞: see Lemma 3.26). As all inequalities in (5.3) must be equalities, we obtain $L(\gamma_\infty) = d(\gamma_\infty(0),\gamma_\infty(1))$, thus γ_∞ is a length minimizer. $\square$

Lemma 5.23. If $\gamma : I \to \mathbb{G}$ is a unit-speed length minimizer defined on a compact interval I, there exists some subinterval $[a,b] \subseteq I$ such that $\left|\underline{\gamma}(b) - \underline{\gamma}(a)\right| \geq c\mathcal{L}^1(I)$, with $c > 0$ depending only on $\mathbb{G}$.

Proof. We can assume $I = [0,T]$, $\gamma(0) = e$ and, by rescaling (i.e. by replacing γ with $\delta_{1/T} \circ \gamma(T\cdot)$), we can assume $T = 1$. Assume by contradiction that such universal constant $c > 0$ does not exist. Then we can find $\gamma_k : [0,1] \to \mathbb{G}$ unit-speed minimizers, with $\gamma_k(0) = e$, such that for any $0 \leq a < b \leq 1$ we have $\left|\underline{\gamma_k}(b) - \underline{\gamma_k}(a)\right| \leq 2^{-k}$. By Lemma 5.22, some subsequence γ_{k_p} converges uniformly to a unit-speed length minimizer $\gamma_\infty : [0,1] \to \mathbb{G}$. Such γ_∞ satisfies

$$\left|\underline{\gamma_\infty}(b) - \underline{\gamma_\infty}(a)\right| = \lim_{p\to\infty} \left|\underline{\gamma_{k_p}}(b) - \underline{\gamma_{k_p}}(a)\right| \leq \lim_{p\to\infty} 2^{-k_p} = 0,$$

for any $0 \leq a < b \leq 1$. So $\underline{\gamma_\infty}$ is constant and γ_∞ has to be constant as well, contradiction. $\square$

Lemma 5.24. Fix any $\epsilon > 0$. If $\gamma : I \to \mathbb{G}$ is a unit-speed length minimizer with $\operatorname{Exc}(\gamma, I) \geq \epsilon > 0$, there exist some $c > 0$, depending only on ϵ and $\mathbb{G}$, and some subintervals $[a_1,b_1],[a_2,b_2] \subseteq I$ (with $b_1 \leq a_2$) such that

$$\left|\det\left(\underline{\gamma}(b_1) - \underline{\gamma}(a_1), \underline{\gamma}(b_2) - \underline{\gamma}(a_2)\right)\right| \geq c(\mathcal{L}^1(I))^2.$$

Here the determinant is defined by means of the identification $V_1 \cong \mathbb{R}^2$, $aX_1 + bX_2 \leftrightarrow \begin{pmatrix} a \\ b \end{pmatrix}$.

Proof. Again, we can assume $I = [0,1]$ and $\gamma(0) = e$ (since the excess does not change by rescaling, i.e. $\operatorname{Exc}\left(\delta_\lambda \circ \gamma\left(\lambda^{-1}\cdot\right), \lambda I\right) = \operatorname{Exc}(\gamma, I)$ for any $\lambda > 0$). By contradiction, there exist unit-speed length minimizers $\gamma_k : [0,1] \to \mathbb{G}$ (with $\gamma_k(0) = e$) such that, for any $0 \leq a_1 < b_1 \leq a_2 < b_2 \leq 1$, we have

$$\left|\det\left(\underline{\gamma_k}(b_1) - \underline{\gamma_k}(a_1), \underline{\gamma_k}(b_2) - \underline{\gamma_k}(a_2)\right)\right| \leq 2^{-k}.$$

By Lemma 5.22, there exists a subsequence (γ_{k_p}) such that $\gamma_{k_p} \to \gamma_\infty$ uniformly and $\underline{\dot\gamma_{k_p}} \to \underline{\dot\gamma_\infty}$ in $L^2([0,1],V_1)$, for some unit-speed length minimizer γ_∞. For any $0 \leq a_1 < b_1 \leq a_2 < b_2 \leq 1$ we have

$$\begin{aligned}
&\left|\det\left(\underline{\gamma_\infty}(b_1) - \underline{\gamma_\infty}(a_1), \underline{\gamma_\infty}(b_2) - \underline{\gamma_\infty}(a_2)\right)\right| \\
&= \lim_{p\to\infty} \left|\det\left(\underline{\gamma_\infty}(b_1) - \underline{\gamma_\infty}(a_1), \underline{\gamma_\infty}(b_2) - \underline{\gamma_\infty}(a_2)\right)\right| \\
&\leq \lim_{p\to\infty} 2^{-k_p} \\
&= 0.
\end{aligned}$$

We deduce that

$$\det\left(\underline{\gamma_\infty}(b_1) - \underline{\gamma_\infty}(a_1), \underline{\gamma_\infty}(b_2) - \underline{\gamma_\infty}(a_2)\right) = 0. \tag{5.4}$$

Now we choose two differentiability points $0 < s < t < 1$ for $\underline{\gamma_\infty}$, which exist as $\underline{\gamma_\infty}$ is Lipschitz. Setting

$$a_1 := s,\ b_1 := s + \delta,\ a_2 := t,\ b_2 := t + \delta$$

and letting $\delta \downarrow 0$ in (5.4), we deduce $\det(\underline{\dot\gamma_\infty}(s), \underline{\dot\gamma_\infty}(t)) = 0$. So all the vectors $\underline{\dot\gamma_\infty}(s)$ (as s varies among the differentiability points) are multiples of some fixed unit vector $v \in V_1$, i.e. $\underline{\dot\gamma_\infty}(s) = \alpha(s)v$ for some Borel function $\alpha : [0,1] \to \{\pm 1\}$, on a Borel subset of $[0,1]$ having full measure. Setting $\beta(s) := \int_0^s \alpha(s')\, ds'$ and writing $v = v_1 X_1 + v_2 V_2$, the curve $s \mapsto \exp(\beta(s)v)$ has the same control $(v_1\alpha, v_2\alpha)$ as γ_∞ and the same starting point. So

$$\gamma_\infty(s) = \exp(\beta(s)v).$$

As γ_∞ is a minimizer, β must be monotone increasing or decreasing, i.e. $\alpha(s) = 1$ a.e. or $\alpha(s) = -1$. This gives $\underline{\gamma_\infty}(s) = s$ for all s or $\underline{\gamma_\infty}(s) = -s$ for all s. In both cases we obtain $\mathrm{Exc}(\gamma_\infty, [0,1]) = 0$. But

$$\mathrm{Exc}(\gamma_\infty, [0,1]) = \lim_{p\to\infty} \mathrm{Exc}(\gamma_{k_p}, [0,1]) \geq \epsilon,$$

since $\underline{\dot\gamma_{k_p}} \to \underline{\dot\gamma_\infty}$ in $L^2([0,1], V_1)$, which is a contradiction. □

We will need the following elementary estimates.

LEMMA 5.25. Let $w_1, w_2 \in \mathbb{R}^2$ be two linearly independent vectors; let us write

$$e_1 = c_{11}w_1 + c_{12}w_2, \quad e_2 = c_{21}w_1 + c_{22}w_2.$$

Then, for any $i, j = 1, 2$, we have the estimate

$$|c_{ij}| \leq \frac{\max\{|w_1|, |w_2|\}}{|\det(w_1, w_2)|}.$$

Proof. In fact, denoting by w_{ij} the j-th component of w_i, it is immediate to check that

$$e_1 = \frac{w_{22}w_1 - w_{12}w_2}{\det(w_1, w_2)}, \quad e_2 = \frac{-w_{21}w_1 + w_{11}w_2}{\det(w_1, w_2)}.$$

□

LEMMA 5.26. Let $\gamma : I \to \mathbb{G}$ be a unit-speed horizontal curve and $J \subseteq I$ a compact subinterval (with $\mathcal{L}^1(J) > 0$). Then

$$L(\gamma) - L(\mathrm{Cut}(\gamma, J)) \geq \frac{\mathcal{L}^1(J)}{2} \mathrm{Exc}(\gamma, J)^2.$$

Proof. Let us write $J = [s, s']$ for some $s < s'$. Let $w \in V_1$ be any unit vector such that $\langle w, \underline{\gamma}(s') - \underline{\gamma}(s)\rangle = |\underline{\gamma}(s') - \underline{\gamma}(s)|$, as in Definition 5.9. Notice that this also gives

$$\left\langle w, \fint_s^{s'} \underline{\dot\gamma}\, d\mathcal{L}^1 \right\rangle = \left| \fint_s^{s'} \underline{\dot\gamma}\, d\mathcal{L}^1 \right|.$$

Since $|\underline{\dot\gamma}| = 1$ a.e., as well, we deduce

$$|\underline{\dot\gamma} - w|^2 = 2\left(1 - \langle w, \underline{\dot\gamma}\rangle\right)$$

and finally

$$
\begin{aligned}
\operatorname{Exc}(\gamma,[s,s'])^2 &= \fint_s^{s'} \left| \underline{\dot\gamma} - \left(\fint_s^{s'} \underline{\dot\gamma}\, d\mathcal{L}^1 \right) \right|^2 d\mathcal{L}^1 \\
&\le \fint_s^{s'} \left| \underline{\dot\gamma} - w \right|^2 d\mathcal{L}^1 \\
&= 2\left(1 - \left\langle w, \fint_s^{s'} \underline{\dot\gamma}\, d\mathcal{L}^1 \right\rangle \right) \\
&= 2\left(1 - \left| \fint_s^{s'} \underline{\dot\gamma}\, d\mathcal{L}^1 \right| \right).
\end{aligned}
$$

Multiplying by $\mathcal{L}^1(J) = s' - s$, we arrive at

$$\mathcal{L}^1(J)\operatorname{Exc}(\gamma,J)^2 = 2\left((s'-s) - \left|\underline{\gamma}(s') - \underline{\gamma}(s)\right|\right) = 2\left(L(\gamma) - L(\operatorname{Cut}(\gamma,J))\right)$$

(see Remark 5.10). □

We are now ready to prove the first of the two announced theorems, which we state again here for the reader's convenience.

THEOREM (one-sided version). Given a constant-speed length minimizer $\gamma : [0,T] \to \mathbb{G}$, there exists a sequence of scales $\eta_i \downarrow 0$ such that

$$\operatorname{Exc}(\gamma,[0,\eta_i]) \to 0.$$

Proof. Step 1. Possibly reparametrizing γ, we can assume that $\gamma : [0,T] \to \mathbb{G}$ is unit-speed. Assume by contradiction that $\operatorname{Exc}(\gamma,[0,t]) \ge \epsilon$ for any sufficiently small t. We will inductively build new unit-speed horizontal curves $\gamma^{(k)} : [0,T_k] \to \mathbb{G}$, for $k = 1,\dots,s$, in such a way that

(i) $\gamma^{(k)}(0) = \gamma(0)$,

(ii) $\gamma(T)^{-1}\gamma^{(k)}(T_k) \in \mathbb{G}_{k+1}$,

(iii) $L\left(\gamma^{(k)}\right) < L(\gamma)$.

In particular, $\gamma^{(k)}$ will be a horizontal curve with the same endpoints as γ, but with smaller length: this clearly contradicts the minimality of γ.

We define $\gamma^{(1)} := \operatorname{Cut}(\gamma,[0,\eta])$, where the parameter $\eta > 0$ will be chosen later. In fact, any sufficiently small η will work; in this proof, the notations $O(\cdot)$ and $o(\cdot)$ will be used for asymptotic estimates which hold as $\eta \to 0$. Notice that, by Remarks 5.11 and 5.10, $\gamma^{(1)}$ satisfies (i), (ii) and (iii) (with $k = 1$).

Step 2. Let us fix parameters $\beta > 0$ and $\rho_1 := 1 > \rho_2 > \cdots > \rho_s > 0$ such that

$$\frac{(k+1)\rho_k - \rho_{k+1}}{k} > 1 + 2\beta$$

for all $k = 1,\dots,s-1$. This is possible if β is chosen to be very small: the last inequality can be rewritten as

$$\rho_k > \frac{\rho_{k+1} + k}{k+1} + \frac{2k}{k+1}\beta$$

and one can proceed by choosing $\rho_s \in (0,1)$ at will, then $\rho_{s-1} < 1$ so as to verify the (strict) inequality when $\beta = 0$ and $k = s-1$, then ρ_{s-2} similarly and so on. By continuity, the inequalities will still hold for a small enough $\beta > 0$.

For any $k = 1, \dots, s-1$, we define $I_k := [0, \eta^{\rho_k}]$. Notice that (as soon as $\eta < 1$)

$$[0, \eta] = I_1 \subseteq I_2 \subseteq \cdots \subseteq I_{s-1}.$$

By Lemma 5.26, the length gain obtained by performing the cut is

$$L(\gamma) - L(\gamma^{(1)}) \geq \eta \frac{\epsilon^2}{2} \geq \eta^{1+2\beta}$$

if η is small enough (since $\operatorname{Exc}(\gamma, [0, \eta]) \geq \epsilon$).

The curves $\gamma^{(k)} : [0, T_k] \to \mathbb{G}$ will be constructed inductively so as to satisfy (i), (ii) and (iii), as well as these additional properties, which already hold for $\gamma^{(1)}$:

(iv) $T_k \geq T_{k-1}$ if $k \geq 2$;

(v) $L\left(\gamma^{(k)}\right) \leq L(\gamma) - (1 + o(1))\eta^{1+2\beta}$ (which is clearly stronger than (iii), when η is small);

(vi) $\underline{\gamma^{(k)}}\big|_{[2\eta^{\rho_k}, T_k]} \equiv \underline{\gamma}\big|_{[2\eta^{\rho_k} + (T - T_k), T]}(\cdot + (T - T_k))$, i.e. on $[2\eta^{\rho_k}, T_k]$ $\gamma^{(k)}$ has the same projection on V_1 as the corresponding final piece of γ;

(vii) $\left\| \underline{\gamma^{(k)}} - \underline{\gamma}\big|_{[0, T_k]} \right\|_\infty = O(\eta)$.

Step 3. Assume that $\gamma^{(k)}$ has been constructed, for some $1 \leq k \leq s-1$, and write

$$\gamma(T)^{-1} \gamma^{(k)}(T_k) = \exp(E_k)$$

for a suitable $E_k \in W_{k+1} = V_{k+1} \oplus \cdots \oplus V_s$. Let us estimate $\overline{\pi}_{k+1}(E_k)$: first of all, by (vi) and the uniqueness part of Lemma 3.26,

$$\gamma^{(k)}\Big|_{2\eta^{\rho_k}}^{T_k} = \gamma\Big|_{\tau_k}^{T},$$

where $\tau_k := 2\eta^{\rho_k} + (T - T_k)$. Hence, defining $g_k := \gamma(\tau_k)^{-1} \gamma^{(k)}(2\eta^{\rho_k})$, we have

$$\begin{aligned}
\gamma^{(k)}(T_k) &= \gamma^{(k)}(2\eta^{\rho_k})\, \gamma^{(k)}\Big|_{2\eta^{\rho_k}}^{T_k} \\
&= \gamma(\tau_k) g_k\, \gamma|_{\tau_k}^{T} \\
&= \gamma(\tau_k)\, \gamma|_{\tau_k}^{T}\, C_{\gamma|_T^{\tau_k}}(g_k) \\
&= \gamma(T) C_{\gamma|_T^{\tau_k}}(g_k).
\end{aligned}$$

As $\gamma(T)^{-1} \gamma^{(k)}(T_k) \in \mathbb{G}_{k+1}$, by Lemma 5.15 we obtain $g_k \in \mathbb{G}_{k+1}$, as well, and

$$\overline{\pi}_{k+1}(E_k) = \pi_{k+1}\left(\gamma(T)^{-1} \gamma^{(k)}(T_k)\right) = \pi_{k+1}(g_k) = O\left(\eta^{(k+1)\rho_k}\right) :$$

the last estimate comes from Lemma 5.21, applied to the curve

$$\gamma|_{[0, \tau_k]}(\tau_k - \cdot) * \gamma^{(k)}\Big|_{[0, 2\eta^{\rho_k}]}$$

connecting $\gamma(\tau_k)$ to $\gamma^{(k)}(2\eta^{\rho_k})$, whose length is $\tau_k + 2\eta^{\rho_k} \leq 5\eta^{\rho_k}$ (as, by (iv), $T - T_k \leq T - T_1 \leq \eta \leq \eta^{\rho_k}$).

Step 4. Let us now build $\gamma^{(k+1)}$. As $V_{k+1} = [V_k, X_1] + [V_k, X_2]$, we can find $Y_1, Y_2 \in V_k$ with

$$[Y_1, X_1] + [Y_2, X_2] = \overline{\pi}_{k+1}(E_k), \quad |Y_1|, |Y_2| = O\left(\eta^{(k+1)\rho_k}\right).$$

Furthermore, we have

$$\operatorname{Exc}(\gamma, I_{k+1}) \geq \epsilon$$

whenever η is small enough. We can apply Lemma 5.24 to I_{k+1}, finding $[a_1, b_1], [a_2, b_2] \subseteq I_{k+1}$ (with $b_1 \le a_2$) such that

$$\left|\det\left(\underline{\gamma}(b_1) - \underline{\gamma}(a_1), \underline{\gamma}(b_2) - \underline{\gamma}(a_2)\right)\right| \ge c\eta^{2\rho_{k+1}}.$$

By (vii) we have

$$\left|\det\left(\underline{\gamma^{(k)}}(b_1) - \underline{\gamma^{(k)}}(a_1), \underline{\gamma^{(k)}}(b_2) - \underline{\gamma^{(k)}}(a_2)\right)\right| \ge c\eta^{2\rho_{k+1}} - O(\eta^{1+\rho_{k+1}}) \ge \frac{c}{2}\eta^{2\rho_{k+1}}$$

for small η. By Lemma 5.25, writing

$$X_i = c_{i1}\left(\underline{\gamma^{(k)}}(b_1) - \underline{\gamma^{(k)}}(a_1)\right) + c_{i2}\left(\underline{\gamma^{(k)}}(b_2) - \underline{\gamma^{(k)}}(a_2)\right)$$

for $i = 1, 2$, we have $|c_{ij}| = O\left(\eta^{-\rho_{k+1}}\right)$. So, defining $Z_1 := c_{11}Y_1 + c_{21}Y_2$ and $Z_2 := c_{12}Y_1 + c_{22}Y_2$, we obtain

$$\overline{\pi}_{k+1}(E_k) = [Z_1, \underline{\gamma^{(k)}}(b_1) - \underline{\gamma^{(k)}}(a_1)] + [Z_2, \underline{\gamma^{(k)}}(b_2) - \underline{\gamma^{(k)}}(a_2)],$$

with $|Z_1|, |Z_2| = O\left(\eta^{(k+1)\rho_k - \rho_{k+1}}\right)$. Finally, we let

$$\gamma^{(k+1)} := \mathrm{Dev}(\gamma^{(k)}, [a_1, b_1], -Z_1, [a_2, b_2], -Z_2).$$

By Proposition 3.19, the extra length needed to create this couple of correction devices is

$$O\left(|Z_1|^{1/k}\right) + O\left(|Z_2|^{1/k}\right) = O\left(\eta^{\frac{(k+1)\rho_k - \rho_{k+1}}{k}}\right) = o\left(\eta^{1+2\beta}\right),$$

by the inequalities imposed on the parameters ρ_k. Thus,

$$L(\gamma^{(k+1)}) \le L(\gamma^{(k)}) + o(\eta^{1+2\beta}).$$

Step 5. Let us check that $\gamma^{(k+1)}$ has the desired properties. We have just verified (iii) and (v), while (i), (iv) and (vii) are trivial. In order to check (vi), we remark that

$$\underline{\gamma^{(k+1)}} = \underline{\gamma^{(k+1)}}\Big|_{[0, \eta^{\rho_{k+1}} + (T_{k+1} - T_k)]} * \underline{\gamma^{(k)}}\Big|_{[\eta^{\rho_{k+1}}, T_k]}$$

and the final point of the first curve in the join coincides with the starting point of the second one. Since $T_{k+1} - T_k = o\left(\eta^{1+2\beta}\right) = o\left(\eta^{\rho_{k+1}}\right)$, if η is small enough we obtain

$$\begin{aligned}\underline{\gamma^{(k+1)}}\Big|_{[2\eta^{\rho_{k+1}}, T_{k+1}]} &\equiv \underline{\gamma^{(k)}}\Big|_{[2\eta^{\rho_{k+1}} - (T_{k+1} - T_k), T_k]}(\cdot - (T_{k+1} - T_k)) \\ &\equiv \underline{\gamma}\Big|_{[2\eta^{\rho_{k+1}} - (T - T_{k+1}), T]}(\cdot + (T - T_{k+1}))\end{aligned}$$

(the last equality holds because $2\eta^{\rho_{k+1}} - (T_{k+1} - T_k) \ge 2\eta^{\rho_k}$ when η is small). Thus, $\gamma^{(k+1)}$ satisfies (vi).

Finally, let us check (ii): applying Lemmas 3.24 and 5.17 (and recalling Definition 5.19), we have

$$\gamma(T)^{-1}\gamma^{(k+1)}(T_{k+1}) = \left(\gamma(T)^{-1}\gamma^{(k)}(T_k)\right)\left(\gamma^{(k)}(T_k)^{-1}\gamma^{(k+1)}(T_{k+1})\right) \in \mathbb{G}_{k+1}$$

and

$$\begin{aligned}\pi_{k+1}\left(\gamma(T)^{-1}\gamma^{(k+1)}(T_{k+1})\right) &= \pi_{k+1}(\exp(E_k)) + \pi_{k+1}\left(\gamma^{(k)}(T_k)^{-1}\gamma^{(k+1)}(T_{k+1})\right) \\ &= \overline{\pi}_{k+1}(E_k) + [-Z_1, \underline{\gamma}(b_1) - \underline{\gamma}(a_1)] + [-Z_2, \underline{\gamma}(b_2) - \underline{\gamma}(a_2)] \\ &= 0.\end{aligned}$$

This finishes the proof. $\square$

Let us now see what changes are needed to prove the two-sided version, which we restate.

THEOREM (two-sided version). Given a constant-speed length minimizer $\gamma : [-T,T] \to \mathbb{G}$, there exists a sequence of scales $\eta_i \downarrow 0$ (depending on γ) such that

$$\mathrm{Exc}(\gamma, [-\eta_i, \eta_i]) \to 0.$$

Proof. The proof is analogous to the preceding one, but now all the constraints imposed on the curves $\gamma^{(k)}$, as well as the cut and correction operations, have to be replaced by their symmetric counterparts. So (again assuming without loss of generality that γ has unit-speed) $\gamma^{(k)} : [-T_k, T_k] \to \mathbb{G}$ is a unit-speed horizontal curve satisfying

(i') $\gamma^{(k)}(-T_k) = \gamma(-T)$;

(ii') $\gamma(T)^{-1}\gamma^{(k)}(T_k) \in \mathbb{G}_{k+1}$;

(iii') $L\left(\gamma^{(k)}\right) < L(\gamma)$;

(iv') $T_k \geq T_{k-1}$ if $k \geq 2$;

(v') $L\left(\gamma^{(k)}\right) \leq L(\gamma) - (1+o(1))\eta^{1+2\beta}$;

(vi') $\underline{\gamma^{(k)}}\big|_{[2\eta^{\rho_k}, T_k]} \equiv \underline{\gamma}\big|_{[2\eta^{\rho_k}+(T-T_k), T]}(\cdot + (T - T_k))$ and $\underline{\gamma^{(k)}}\big|_{[-T_k, -2\eta^{\rho_k}]} \equiv \underline{\gamma}\big|_{[-T, -2\eta^{\rho_k}-(T-T_k)]}(\cdot - (T - T_k))$;

(vii') $\left\|\underline{\gamma^{(k)}} - \underline{\gamma}\big|_{[-T_k, T_k]}\right\|_\infty = O(\eta)$.

and the first curve is $\gamma^{(1)} := \mathrm{Cut}'(\gamma, [-\eta, \eta])$. We now list the necessary modifications in the various steps.

The estimate $\pi_{k+1}(g_k) = O\left(\eta^{(k+1)\rho_k}\right)$ follows by applying Lemma 5.21 to

$$\gamma\big|_{[-\tau_k, \tau_k]}(\tau_k - \cdot) * \gamma^{(k)}\Big|_{[-2\eta^{\rho_k}, 2\eta^{\rho_k}]}$$

and noticing that, by (i'), (vi') and Lemma 3.27, $\gamma(-\tau_k) = \gamma^{(k)}(-2\eta^{\rho_k})$.

Finally, in Step 5, the fact that $\gamma^{(k+1)}$ satisfies (vi) follows from the identity

$$\underline{\gamma^{(k+1)}} = \underline{\gamma^{(k)}}\Big|_{[-T_k, -\eta^{\rho_{k+1}}]} * \underline{\gamma^{(k+1)}}\Big|_{[-\eta^{\rho_{k+1}}-(T_{k+1}-T_k), \eta^{\rho_{k+1}}+(T_{k+1}-T_k)]} * \underline{\gamma^{(k)}}\Big|_{[\eta^{\rho_{k+1}}, T_k]},$$

where the final point of each curve in the join coincides with the starting point of the next one. $\square$

5.4. A toy problem

One could wonder whether the techniques introduced in this chapter can be used to obtain stronger regularity results. The proof of Theorem 5.3 in fact shows also the following statement (where one can take $\tau_1 := \rho_2$ and $\tau_2 := \rho_s$).

THEOREM 5.27. Fix $\epsilon > 0$. There exist some $\beta > 0$, $0 < \tau_2 < \tau_1 < 1$ (independent of ϵ) and some threshold $\overline{\eta} = \overline{\eta}(\epsilon)$ such that, for any $\eta \in (0, \overline{\eta})$, this implication holds:

$$\mathrm{Exc}(\delta, [0,t]) \geq \epsilon \ \forall t \in [\eta^{\tau_1}, \eta^{\tau_2}] \Rightarrow \mathrm{Exc}(\delta, [0,\eta]) \leq \eta^\beta.$$

This suggests that constant-speed length minimizers should $C^{1,\beta}$-regular for some small β, depending solely on $\mathbb{G}$: indeed, if a curve γ satisfies $\mathrm{Exc}(\gamma, J) \leq (\mathcal{L}^1(J))^\beta$ for any sufficiently small subinterval $J \subseteq I$, then (by Campanato's embedding theorem) $\underline{\gamma} \in C^{1,\beta}$, so that $\gamma \in C^{1,\beta}$, as well. Of course, this cannot be obtained by the above statement, due to the presence of the hypothesis of big excess.

An interesting toy problem in Carnot groups $\mathbb{G}$ with rank 2 is to prove the non-minimality for this special family of curves: given $\alpha \in (1,2)$, let $\gamma_\alpha : [0,+\infty) \to \mathbb{G}$ be the horizontal curve such that $\gamma_\alpha(0) = e$ and

$$\underline{\gamma_\alpha}(t) = tX_1 + t^\alpha X_2.$$

Since any constant-speed reparametrization of γ_α is not $C^{1,\beta}$, for any $\beta > \alpha$, the following result can be viewed as an additional piece of evidence for the above heuristic.

THEOREM 5.28. If $\alpha > 1$ is sufficiently close to 1, then, for any positive T, $\gamma_\alpha|_{[0,T]}$ is not a length minimizer.

The proof uses the same techniques contained in the proof of Theorem 5.3, but is much simpler. In particular, the notion of excess is no longer involved.

Contrary to the preceding section, we will not have the necessity to deal exclusively with unit-speed curves. Thus, in order to lighten the notation, for any horizontal curve $\gamma : [0,T] \to \mathbb{G}$ we replace $\mathrm{Cut}(\gamma,[s,s'])$ and $\mathrm{Dev}(\gamma,[s,s'],Y)$ by suitable reparametrizations $\overline{\mathrm{Cut}}(\gamma,[s,s'])$ and $\overline{\mathrm{Dev}}(\gamma,[s,s'],Y)$, which still have $[0,T]$ as their domain and satisfy

$$\pi \circ \gamma \equiv \pi \circ \overline{\mathrm{Cut}}(\gamma,[s,s']) = \pi \circ \overline{\mathrm{Cut}}(\gamma,[s,s'],Y)$$

on $[0,T] \setminus (s,s')$. Explicitly, we call $\overline{\mathrm{Cut}}(\gamma,[s,s'])$ the curve built in Definition 5.9, with $\exp(\cdot w)|_{[0,|\underline{\gamma}(s')-\underline{\gamma}(s)|]}$ replaced by $\exp\left(\frac{|\underline{\gamma}(s')-\underline{\gamma}(s)|}{s'-s}\cdot w\right)\Big|_{[0,s'-s]}$. Similarly, we let

$$\overline{\mathrm{Dev}}(\gamma,[s,s'],Y) := \gamma|_{[0,s]} * \sigma * \gamma|_{[s',T]},$$

where σ is the constant-speed reparametrization of $\delta_Y * \gamma|_{[s,s']} * \delta_Y(\ell_Y - \cdot)$ with domain $[s,s']$.

Proof. Call $\gamma := \gamma_\alpha|_{[0,T]}$. Let $\beta := \alpha - 1$ be so small that we can find parameters $\rho_1 := 1 > \rho_2 > \cdots > \rho_s > 0$ satisfying

$$\frac{(k+1)\rho_k - (1+\beta)\rho_{k+1}}{k} > 1 + 2\beta,$$

for all $k = 1,\ldots,s-1$. Let η be any positive parameter to be chosen later (any sufficiently small η will work) and $\gamma^{(1)} := \overline{\mathrm{Cut}}(\gamma,[0,\eta])$. The length gain produced by this cut is

$$L(\gamma) - L(\gamma^{(1)}) = \int_0^\eta \sqrt{1+\alpha^2 t^{2\beta}}\,dt - \sqrt{\eta^2+\eta^{2\alpha}} = (c+o(1))\eta^{1+2\beta}$$

for some $c = c(\alpha) > 0$. Assume now that $\gamma^{(k)}$ has been constructed in such a way that

$$\gamma^{(k)}(0) = e, \quad \gamma(T)^{-1}\gamma^{(k)}(T) \in \mathbb{G}_{k+1}, \quad L(\gamma^{(k)}) < L(\gamma), \quad \underline{\gamma^{(k)}} \equiv \underline{\gamma} \text{ on } [3\eta^{\rho_k},T] \tag{5.5}$$

for some $1 \le k \le s-1$. Writing

$$\gamma(T)^{-1}\gamma^{(k)}(T) = \exp(E_k),$$

by Lemma 5.21 we have (arguing as in Step 3 of the proof of Theorem 5.3)

$$\overline{\pi}_{k+1}(E_k) = \pi_{k+1}(\gamma(3\eta^{\rho_k})^{-1}\gamma^{(k)}(3\eta^{\rho_k})) = O\left(\eta^{(k+1)\rho_k}\right).$$

So we can find $Y_1, Y_2 \in V_k$ with

$$[Y_1,X_1] + [Y_2,X_2] = \overline{\pi}_{k+1}(E_k), \quad |Y_1|,|Y_2| = O\left(\eta^{(k+1)\rho_k}\right).$$

Since

$$\det\left(\underline{\gamma}(2\eta^{\rho_{k+1}}) - \underline{\gamma}(\eta^{\rho_{k+1}}), \underline{\gamma}(3\eta^{\rho_{k+1}}) - \underline{\gamma}(2\eta^{\rho_{k+1}})\right) = c'\eta^{(2+\beta)\rho_{k+1}},$$

using Lemma 5.25 we can write

$$\overline{\pi}_{k+1}(E_k) = \left[Z_1, \underline{\gamma}(2\eta^{\rho_{k+1}}) - \underline{\gamma}(\eta^{\rho_{k+1}})\right] + \left[Z_2, \underline{\gamma}(3\eta^{\rho_{k+1}}) - \underline{\gamma}(2\eta^{\rho_{k+1}})\right]$$

for suitable $Z_1, Z_2 \in V_k$, with $|Z_1|, |Z_2| = O\left(\eta^{(k+1)\rho_k - (1+\beta)\rho_{k+1}}\right)$. Notice that $\underline{\gamma}$ agrees with $\underline{\gamma^{(k)}}$ at $\eta^{\rho_{k+1}}, 2\eta^{\rho_{k+1}}, 3\eta^{\rho_{k+1}}$ if η is small. Finally, we define

$$\gamma^{(k+1)} := \overline{\mathrm{Dev}}(\overline{\mathrm{Dev}}(\gamma^{(k)}, [\eta^{\rho_{k+1}}, 2\eta^{\rho_{k+1}}], -Z_1), [2\eta^{\rho_{k+1}}, 3\eta^{\rho_{k+1}}], -Z_2).$$

The extra length needed to create the devices is $o(\eta^{1+2\beta})$. Thus, $\gamma^{(k+1)}$ satisfies (5.5) (with $k+1$ in place of k). The final curve $\gamma^{(s)}$ has the same endpoints as γ, but smaller length (if η is chosen sufficiently small). $\square$

APPENDIX A

Some well-known analytic facts

A.1. Local openness of perturbations of invertible linear maps

LEMMA A.1. Let $A : \mathbb{R}^N \to \mathbb{R}^N$ be an invertible linear map and let $F : \overline{B}_r \to \mathbb{R}^N$ (where $\overline{B}_r := \overline{B}_r(0) \subset \mathbb{R}^N$) be a continuous map such that $\|F - A\|_{C^0(\partial B_r, \mathbb{R}^N)} < r \left\|A^{-1}\right\|^{-1}$ (the last norm is any submultiplicative one, such as the operator norm). Then we have $0 \in \operatorname{int} F(B_r)$.

Proof. Let us assume first that $A = I$ and $r = 1$. Call $\epsilon := 1 - \|F - \mathrm{id}\|_{C^0(S^{N-1}, \mathbb{R}^N)} > 0$ and assume by contradiction that some $y \in B_\epsilon$ does not belong to $F(B_1)$. Since for any $x \in S^{N-1}$ we have $|F(x)| \geq 1 - \|F - \mathrm{id}\|_{C^0(S^{N-1}, \mathbb{R}^N)} = \epsilon$, we get that $y \notin F(\overline{B}_1)$ as well. So we can define

$$G(x) := \frac{y - F(x)}{|y - F(x)|},$$

which maps $\overline{B}_1 \to S^{N-1}$ continuously. Any fixed point x of G has to belong to S^{N-1} and here $G(x) = x$ is equivalent to $\langle x, G(x) \rangle = 1$, but

$$\langle x, y - F(x) \rangle = \langle x, y \rangle + \langle x, x - F(x) \rangle - 1 \leq |y| + \|F - \mathrm{id}\|_{C^0(S^{N-1}, \mathbb{R}^N)} - 1 = |y| - \epsilon < 0.$$

So G maps $\overline{B}_1$ into itself and has no fixed point, contradicting Brouwer's fixed point theorem. This concludes the proof for the case $A = I$ and $r = 1$.
We can always reduce to this situation by considering $\overline{F} : \overline{B}_1 \to \mathbb{R}^N$, $\overline{F}(x) := \frac{1}{r} A^{-1} F(rx)$, which satisfies

$$\left|\overline{F}(x) - x\right| = \left|\frac{1}{r} A^{-1} F(rx) - x\right| = \frac{1}{r} \left|A^{-1}\left(F(rx) - A(rx)\right)\right| < 1$$

for any $x \in \partial B_1$. The previous discussion gives $0 \in \operatorname{int} \overline{F}(B_1)$, so $0 \in \operatorname{int} F(B_r)$ as well. $\square$

COROLLARY A.2. Let $\Omega \subseteq \mathbb{R}^N$ be an open set containing 0 and $F : \Omega \to \mathbb{R}^N$ continuous such that $F(x) = Ax + o(|x|)$ (as $x \to 0$), where $A : \mathbb{R}^N \to \mathbb{R}^N$ is an invertible linear map. Then F is locally open at 0, i.e. $0 \in \operatorname{int} F(U)$ for any open $U \subseteq \Omega$ containing 0. In particular $0 \in \operatorname{int} F(\Omega)$.

Proof. It suffices to apply Lemma A.1 to $F|_{\overline{B}_r}$, which satisfies the hypotheses for any sufficiently small r. $\square$

A.2. First and second order optimality conditions for constrained problems

LEMMA A.3. Let $\Omega \subseteq \mathbb{R}^N$ be open and $f, h_1, \ldots, h_m \in C^2(\Omega)$. Let

$$\mathcal{V} := \{x \in \Omega : h_1(x) = \cdots = h_m(x) = 0\}$$

and assume that $\overline{x}$ is a local minimum for $f|_{\mathcal{V}}$ and that $dh_1(x), \ldots, dh_m(x)$ are linearly independent. Then there exist unique real numbers $\lambda_1, \ldots, \lambda_m$ such that

$$df(\overline{x}) + \sum_{i=1}^{m} \lambda_i dh_i(\overline{x}) = 0.$$

Moreover, we have the second order optimality condition

$$d^2 f(\overline{x}) + \sum_{i=1}^{m} \lambda_i d^2 h_i(\overline{x}) \geq 0$$

on the vector subspace $Z := \cap_{i=1}^{m} \ker dh_i(\overline{x})$.

Proof. By the implicit function theorem, possibly replacing Ω by a sufficiently small ball centered at $\overline{x}$, $\mathcal{V}$ is an $(N-m)$-dimensional embedded manifold with $T_{\overline{x}}\mathcal{V} = Z$. So for any $v \in Z$ there exists a C^2 curve $\sigma : (-\epsilon, \epsilon) \to \mathcal{V}$ such that $\sigma(0) = \overline{x}$ and $\dot{\sigma}(0) = v$. Now, since $f \circ \sigma$ has a local minimum at 0, we have

$$df(\overline{x})[v] = \frac{d}{dt}(f \circ \sigma)\Big|_{t=0} = 0,$$

so $df(\overline{x})$ vanishes on Z. Hence, it defines a linear functional on the quotient $\mathbb{R}^N/Z$, whose dual has $dh_1(\overline{x}), \ldots, dh_m(\overline{x})$ as a basis, thus we get the first part of the thesis.
As for the second part, fixing $v \in Z$ and σ as before, differentiating twice the identity $h_i \circ \sigma \equiv 0$ gives

$$d^2 h_i(\overline{x})[v,v] + dh_i(\overline{x})[\ddot{\sigma}(0)] \equiv 0. \tag{A.1}$$

Again from local minimality we have $\frac{d^2}{dt^2}(f \circ \sigma)\Big|_{t=0} \geq 0$, i.e.

$$d^2 f(\overline{x})[v,v] + df(\overline{x})[\ddot{\sigma}(0)] \geq 0.$$

Multiplying (A.1) by λ_i and summing these equations together with the last one we arrive to

$$d^2 f(\overline{x})[v,v] + \sum_i \lambda_i d^2 h_i(\overline{x})[v,v] + df(\overline{x})[\ddot{\sigma}(0)] + \sum_i \lambda_i dh_i(\overline{x})[\ddot{\sigma}(0)] \geq 0$$

and we are done since $df(\overline{x}) + \sum_i \lambda_i dh_i(\overline{x}) = 0$. □

A.3. Absolutely continuous functions

We recall that, given a continuous increasing function $h : [0,T] \to [0,T']$, there exists a unique finite positive measure μ on the Borel σ-algebra $\mathcal{B}([0,T])$ such that $h(t) = \mu([0,t])$ for any $0 \leq t \leq T$. Moreover, such μ satisfies $\mu(\{t\}) = 0$ for any t. Let us begin with a simple useful identity.

Lemma A.4. Given h and μ as above, if we also have $h(0) = 0$ and $h(T) = T'$, then $h_*\mu = \mathcal{L}^1 \llcorner [0,T']$.

Proof. The additional condition $h(0) = 0$, $h(T) = T'$ amounts to saying that h is surjective. So, given any $0 \leq s < t \leq T'$, we can write $h^{-1}(s) = [a,a']$ and $h^{-1}(t) = [b,b']$ (possibly with $a = a'$ or $b = b'$). Now

$$h_*\mu([s,t]) = \mu\left(h^{-1}([s,t])\right) = \mu([a,b']) = \mu((a,b']) = h(b') - h(a) = t - s = \mathcal{L}^1([s,t]).$$

Thus $h_*\mu$ and $\mathcal{L}^1 \llcorner [0,T']$ agree on open subsets of $[0,T']$, so they coincide by outer regularity. □

We now turn to the case of *AC* increasing functions. We recall that, in this case, we have $\mu \ll \mathcal{L}^1$ and $\mu = \dot{h}\mathcal{L}^1$. All the facts which are stated and proved in the remainder of this section are special cases of the *area formula* for *AC* functions, which is discussed, for instance, in [AT04, Section 3.4].

LEMMA A.5. Given an *AC* increasing function $h : [0,T] \to \mathbb{R}$, for any $\mathcal{L}^1$-negligible Borel set $N \subseteq [0,T]$ we have

$$\mathcal{L}^1\left(h\left(N \cup \left\{\dot{h} = 0\right\}\right)\right) = 0.$$

Here $\dot{h}$ denotes the classical derivative of h, which exists a.e. (so $\left\{\dot{h} = 0\right\}$ is the set of all $t \in [0,T]$ such that $\dot{h}(t)$ exists and vanishes).

Proof. We can assume, without loss of generality, that $h(0) = 0$ and $T' := h(T) > 0$. Let $S := \bigcup_t h^{-1}(t)$, where t varies over the elements of $[0,T']$ whose preimage is greater than a singleton. Notice that this union is at most countable, so S is a Borel set where $\dot{h} = 0$ a.e., proving that $\mu(S) = \int_S \dot{h}(t)\,dt = 0$. Thus, for any $B \in \mathcal{B}([0,T])$,

$$\mathcal{L}^1(h(B)) = \mu(h^{-1}(h(B))) = \mu(B \cap S) + \mu\left(h^{-1}(h(B)) \setminus S\right) = \mu(B \cap S) + \mu(B \setminus S) = \mu(B),$$

since $h^{-1}(h(B)) \setminus S = B \setminus S$. Now $\left\{\dot{h} = 0\right\}$ is a Borel set and $\mu\left(\left\{\dot{h} = 0\right\}\right) = \mu(N) = 0$ (as $\mu \ll \mathcal{L}^1$), so we deduce

$$\mathcal{L}^1\left(h\left(N \cup \left\{\dot{h} = 0\right\}\right)\right) = \mu\left(N \cup \left\{\dot{h} = 0\right\}\right) = 0.$$

$\square$

LEMMA A.6. Given an *AC* increasing function $h : [0,T] \to \mathbb{R}$, we have

$$\mathcal{L}^1\left(h^{-1}(N) \setminus \left\{\dot{h} = 0\right\}\right) = 0$$

for any $\mathcal{L}^1$-negligible Borel set N.

Proof. Lemma A.4 tells us that

$$\mu\left(h^{-1}(N)\right) = h_*\mu(N) = \mathcal{L}^1(N) = 0,$$

so, recalling that $\mu = \dot{h}\mathcal{L}^1$, we obtain

$$0 = \int_{h^{-1}(N)} \dot{h}(t)\,dt = \int_{h^{-1}(N)\setminus\left\{\dot{h}=0\right\}} \dot{h}(t)\,dt.$$

But $\dot{h} > 0$ a.e. on $h^{-1}(N) \setminus \left\{\dot{h} = 0\right\}$, so this set has to be $\mathcal{L}^1$-negligible. $\square$

LEMMA A.7 (change of variables). If $\phi \in L^1([0,T'])$ (possibly undefined on a negligible subset) and $h : [0,T] \to [0,T']$ is *AC*, increasing and surjective, then $(\phi \circ h)\dot{h}$ is well-defined a.e. and is measurable, with the convention that the right-hand side vanishes at t whenever $\dot{h}(t) = 0$ (even when ϕ is not defined at $h(t)$). Moreover, it is summable and satisfies the identity

$$\int_0^T \phi \circ h(t)\dot{h}(t)\,dt = \int_0^{T'} \phi(t')\,dt'.$$

Proof. The fact that $(\phi \circ h)\dot{h}$ is defined a.e. and is measurable follows from Lemma A.6. In order to prove the second statement, we can assume that ϕ is defined and nonnegative everywhere and that it is Borel. By Lemma A.4 we have

$$\int_0^T \phi \circ h(t)\dot{h}(t)\,dt = \int_{[0,T]} \phi \circ h\,d\mu = \int_{[0,T']} \phi\,dh_*\mu = \int_0^{T'} \phi(t')\,dt'$$

(notice that this proves the summability of $(\phi \circ h)\dot{h}$ in the general case). $\square$

Corollary A.8 (chain rule). If $u : [0,T'] \to \mathbb{R}$ is AC and $h : [0,T] \to [0,T']$ is AC, increasing and surjective, then $u \circ h$ is AC as well. Moreover, its classical derivative is given a.e. by the formula

$$\frac{d}{dt}(u \circ h) = (\dot{u} \circ h)\dot{h}.$$

Proof. It suffices to apply Lemma A.7 to $h|_{[0,t]}$, for any $0 \le t \le T$:

$$u \circ h(t) - u \circ h(0) = u(h(t)) - u(0) = \int_0^{h(t)} \dot{u}(s')\,ds' = \int_0^t \dot{u} \circ h(s)\dot{h}(s)\,ds.$$

$\square$

APPENDIX B

Existence, uniqueness and regularity of flows

The aim of this section is to briefly revisit the classical Cauchy-Lipschitz theory for ordinary differential equations of the form

$$\dot{\gamma}(t) = \sum_{i=1}^{r} u_i(t) X_i(\gamma(t)),$$

where the X_i's are globally defined smooth vector fields on a smooth manifold M and $u_i \in L^2(I)$, for some fixed interval $I \subseteq \mathbb{R}$. This equation is intended to hold a.e. and we require that the solution γ belongs to H^1_{loc}, whose meaning is made precise below.

Definition B.1. Fix any interval $(a,b) \subseteq \mathbb{R}$, with $-\infty \leq a < b \leq +\infty$. We say that $\gamma : (a,b) \to M$ belongs to H^1_{loc} if γ is continuous and, for any $t \in (a,b)$, there exist a local chart $\phi : U \to \mathbb{R}^n$ and some $\epsilon > 0$ such that $\gamma\left([t-\epsilon, t+\epsilon]\right) \subseteq U$ and $\phi \circ \gamma \in H^1\left([t-\epsilon, t+\epsilon], \mathbb{R}^n\right)$. The meaning of $\gamma \in H^1_{loc}(I, M)$ when I is a non-open interval is defined similarly.

When I is a compact interval, we write $H^1(I, M)$ in place of $H^1_{loc}(I, M)$ to emphasize the fact that, for any $\gamma \in H^1(I, M)$, its energy $\int_I |\dot{\gamma}|^2(t)\, dt$ is finite for any fixed Riemannian metric on M.

Remark B.2. Since transition maps are smooth and curves in H^1 are well-behaved under composition with a smooth map (i.e. for any $\sigma \in H^1([c,d], \mathbb{R}^n)$ and any $\psi \in C^1(\Omega, \mathbb{R}^n)$ with $\sigma([c,d]) \subseteq \Omega$ we still have $\psi \circ \sigma \in H^1([c,d], \mathbb{R}^n)$), this definition immediately implies that, given $\gamma \in H^1_{loc}((a,b), M)$, for *any* local chart $\phi : U \to \mathbb{R}^n$ we have $\phi \circ \gamma \in H^1_{loc}\left(\gamma^{-1}(U), \mathbb{R}^n\right)$. Moreover, we can give a precise meaning to the equation, by requiring that any time t has a neighbourhood where it holds a.e. in a local chart. This is a good definition since for $\mathbb{R}^n$-valued curves belonging to H^1_{loc} the time derivative can be interpreted a.e. as a classical one, and thus gives a well-defined vector in TM.

To begin with, we prove a local existence result.

Proposition B.3. Given $x \in \mathbb{R}^n$, $u = (u_1, \ldots, u_r) \in L^2\left([0,1], \mathbb{R}^r\right)$ and smooth vector fields $X_1, \ldots, X_r$ on $\mathbb{R}^n$, there exists some $0 < T \leq 1$ (depending on x, u and the vector fields) such that the equation $\dot{\gamma}(t) = \sum_i u_i(t) X_i(\gamma(t))$ has a solution $\gamma \in H^1([0,T], \mathbb{R}^n)$ with initial condition $\gamma(0) = x$.

Proof. We fix some positive T, to be chosen later. Consider the map

$$F : C^0([0,T], \mathbb{R}^n) \to H^1([0,T], \mathbb{R}^n) \subseteq C^0([0,T], \mathbb{R}^n)$$

given by the formula

$$F(\gamma)(t) := x + \int_0^t \sum_i u_i(s) X_i(\gamma(s))\, ds.$$

Fix any radius $R > 0$ and call $S_R := \{\gamma \in C^0([0,T],\mathbb{R}^n) : \|\gamma - x\|_\infty \le R\}$. If $\gamma \in S_R$ we have the simple estimate

$$\begin{aligned}\|F(\gamma) - x\|_\infty^2 &\le T \int_0^T \left(\sum_i u_i(s) X_i(\gamma(s))\right)^2 ds \\ &\le T\,\|u\|_2^2 \int_0^T \sum_i |X_i(\gamma(s))|^2\, ds \\ &\le CT^2\,\|u\|_2^2\end{aligned}$$

for some positive constant C (explicitly $C := \max_{\overline{B}_R} \sum_i |X_i|^2$). If T is sufficiently small we have $CT^2 \le R^2$, so F maps the closed set S_R into itself. Moreover, for any $\gamma, \delta \in S_R$

$$\|F(\gamma) - F(\delta)\|_\infty^2 \le T\,\|u\|_2^2 \int_0^T |X_i(\gamma(s)) - X_i(\delta(s))|^2\, ds \le C'T^2\,\|\gamma - \delta\|_\infty^2$$

for some $C' > 0$ (explicitly $C' := r \max_i \max_{\overline{B}_R} |dX_i|^2$). So, possibly shrinking T so that $C'T^2 < 1$ as well, we obtain that $F|_{S_R}$ is a contraction and the Banach contraction principle tells us that there is a (unique) fixed point $\gamma \in S_R$. As such, γ belongs to $H^1([0,T],\mathbb{R}^n)$ and solves the differential equation (for a.e. t) with the required initial condition. $\square$

Remark B.4. Clearly the same proof works even when the vector fields are only defined on an open neighbourhood of x and $[0,1]$ is replaced by any other compact interval. We can also solve the differential equation for small negative times, i.e. on $[-T,0]$ (provided u is defined here, of course), simply by reflecting time, which amounts to replace $u(t)$ by $-u(-t)$. Joining the solutions for positive and negative times gives a solution on a neighbourhood of 0, as well.

We now prove a global uniqueness result.

Proposition B.5. Given $x \in M$, $t_0 \in \mathbb{R}$ and $u \in L^2(I,\mathbb{R}^r)$, the equation $\dot\gamma(t) = \sum_i u_i(t) X_i(\gamma(t))$ has at most one solution $\gamma \in H^1_{loc}(I,M)$ such that $\gamma(t_0) = x$, for any interval I containing t_0.

Proof. Let γ, δ be distinct solutions. We can assume that $\gamma(t') \ne \delta(t')$ for some $t' > t_0$ (in the other case it suffices to reflect time, considering the solutions $t \mapsto \gamma(-t)$, $t \mapsto \delta(-t)$ to the same differential equation with $u(t)$ replaced by $-u(-t)$, and t_0 replaced by $-t_0$). Let

$$\bar t := \max\{t \ge t_0 : \gamma \equiv \delta \text{ on } [t_0, t]\} < t'$$

(continuity of γ and δ allows us to write a maximum instead of a supremum). By composing with a local chart near $x := \gamma(\bar t) = \delta(\bar t)$, we can assume that $M = \mathbb{R}^n$. We now fix any $R > 0$ and by continuity we have $\gamma([\bar t, t_1]), \delta([\bar t, t_1]) \subseteq \overline{B}_r(x)$, for some $t_1 > \bar t$. As in the proof of Proposition B.3, if t_1 is chosen sufficiently close to $\bar t$ the map F defined above is a contraction on the closed set S_R (with the interval $[0,T]$ replaced by $[\bar t, t_1]$), hence it has a unique fixed point. But $\gamma|_{[\bar t, t_1]}$ and $\delta|_{[\bar t, t_1]}$ are distinct fixed points of F and we arrive to a contradiction. $\square$

Remark B.6. Fix a nontrivial interval $I \subseteq \mathbb{R}$ and $t_0 \in I$. Uniqueness allows to define, for any $u \in L^2(I,\mathbb{R}^r)$ and any initial condition $\gamma(t_0) = x$, the maximal interval $I_{max}(x,u) \subseteq I$ containing t_0 where the solution exists. Clearly I_{max} has to be relatively open in I (if for example $I = [0,1]$, $t_0 = 0$ and $I_{max} = [0,T]$, with $0 < T < 1$, Proposition B.3 would provide a solution locally near T, with initial condition $\gamma(T)$ at time T, which we could then join to γ, contradicting the maximality of I_{max}).

Definition B.7. From now on, let us denote by $I_{max}(x,u)$ the maximal interval where the solution associated to (x,u) exists, as in the previous remark, where we choose $t_0 := 0$ and $I := [0,1]$. Moreover, we define

$$\mathcal{V} := \left\{(x,u,t) : x \in M,\ u \in L^2([0,1],\mathbb{R}^r),\ [0,t] \subseteq I_{max}(x,u)\right\}.$$

Definition B.8. Let $(x,u,t) \in \mathcal{V}$. We will use the notation $\mathrm{End}_t(x,u) := \gamma(t)$ for the flow map (here γ is the solution associated to (x,u)). End_t is the *endpoint map* at time t. When t is omitted, it is meant to be 1, so that $\mathrm{End} := \mathrm{End}_1$.

Let us move to the regularity properties of the flow, beginning with continuity. In the sequel we will assume that controls are defined on $[0,1]$, but similar results hold for any compact interval $I \subseteq \mathbb{R}$.

Lemma B.9. If $M = \mathbb{R}^n$ and γ_0 solves the equation on $[a,b] \subseteq [0,1]$ with initial condition $\gamma_0(a) = x_0$ and control u_0, for any $\delta > 0$ there exists some $\epsilon > 0$ such that, for all $x \in B_\epsilon(x_0)$ and all u satisfying $\|u - u_0\|_2 < \epsilon$, a solution γ with initial condition $\gamma(a) = x$ and control u exists on $[a,b]$ and satisfies $\|\gamma - \gamma_0\|_\infty < \delta$ on $[a,b]$.

Proof. Fix a large open ball B containing $\gamma_0([a,b])$ and set $R := \operatorname{dist}(\gamma_0([a,b]), B^c) > 0$. The proof of Proposition B.3 shows that there exists some $T > 0$ such that, for any $x \in \overline{B}$ and any initial condition $\gamma(s) = x$, a solution exists on the interval $[s, (s+T) \wedge b]$. Now let γ be the maximal solution associated to x and u, defined on $[a,\bar{t})$ (for some $\bar{t} < b$) or on the whole of $[a,b]$. In the second case we let $\bar{t} := b$. For any $a \le s < t \le b$ such that $\|\gamma - \gamma_0\|_{C^0([s,t])} \le R$ we have $\gamma([s,t]) \subseteq \overline{B}$, so

$$\begin{aligned}
\|\gamma - \gamma_0\|_{C^0([s,t])} &\le |\gamma(s) - \gamma_0(s)| + \int_s^t \left|\sum_i (u_i X_i \circ \gamma - (u_0)_i X_i \circ \gamma_0)\right| \\
&\le |\gamma(s) - \gamma_0(s)| + C\sqrt{t-s}\,\|u - u_0\|_2 + \int_s^t \left|\sum_i (u_0)_i\,(X_i \circ \gamma - X_i \circ \gamma_0)\right| \\
&\le |\gamma(s) - \gamma_0(s)| + C\sqrt{t-s}\,\|u - u_0\|_2 + C\sqrt{t-s}\,\|\gamma - \gamma_0\|_{C^0([s,t])},
\end{aligned}$$

for some $C > 0$ depending only on the choice of B. Choose now any $\alpha \in (0,1)$. If $t - s \le \alpha^2 C^{-2}$ we deduce

$$(1-\alpha)\,\|\gamma - \gamma_0\|_{C^0([s,t])} \le |\gamma(s) - \gamma_0(s)| + C\sqrt{t-s}\,\|u - u_0\|_2\,. \tag{B.1}$$

Now we subdivide the interval $K := \left[a, \left(\bar{t} - \frac{T}{2}\right) \vee a\right]$ in a finite number of subintervals of length at most $\alpha^2 C^{-2}$. Since the number of the needed subintervals can be bounded by a constant N depending only on C and α, iterating inequality (B.1) N times we obtain

$$\|\gamma - \gamma_0\|_{C^0(K)} \le C'\,|\gamma(s) - \gamma_0(s)| + C'\,\|u - u_0\|_2$$

for some C' depending only on B and α. So if x and u are as in the hypothesis and ϵ is sufficiently small we have $\|\gamma - \gamma_0\|_{C^0(K)} \le R$. But then $\gamma\left(\left(\bar{t} - \frac{T}{2}\right) \vee a\right) \in \overline{B}$ and, by the choice of T, this implies that γ is defined on $\left[a, \left(a + \frac{T}{2}\right) \wedge b\right]$. Thus γ is defined on $[a,b]$ and, choosing an even smaller ϵ, we can guarantee the estimate $\|\gamma - \gamma_0\|_\infty < \delta$. ☐

Proposition B.10. The set $\mathcal{V}$ is relatively open in $M \times L^2([0,1],\mathbb{R}^r) \times [0,1]$. The flow $\mathrm{End}_t(x,u)$ is continuous as a map from $\mathcal{V}$ to M.

Proof. To prove the openness of $\mathcal{V}$ we fix any $(x_0, u_0, t_0) \in \mathcal{V}$ and call γ_0 the maximal solution associated to (x_0, u_0). If $t_0 < 1$ we choose some $\epsilon > 0$ such that $t_0 + \epsilon \in I_{max}(x, u)$, otherwise we set $\epsilon := 0$. It suffices to prove that, if x is sufficiently close to x_0 and $\|u - u_0\|_2$ is sufficiently small, the maximal solution associated to (x, u) is defined on $[0, t_0 + \epsilon]$. This follows immediately from the last lemma once we split the interval $[0, t_0]$ into finitely many subintervals whose images are contained in domains of local charts, and iterate its statement from the last subinterval to the first one (beginning with $\delta := 1$, for example).

This argument shows also that if $(x_n, u_n) \to (x, u)$ then the solutions associated to (x_n, u_n) converge uniformly on $[0, t_0 + \epsilon]$ to γ_0. This gives the second part of the thesis. $\square$

COROLLARY B.11 (escape from compact sets). If $I_{max}(x, u) \subsetneq [0, 1]$, then $\gamma(I_{max}(x, u))$ is not contained in any compact subset of M. Here γ is the maximal solution associated to (x, u).

Proof. Assume by contradiction that $\gamma(I_{max}(x, u)) \subseteq K$, where $K \subseteq M$ is compact. Since $K \times \{0\} \times \{0\}$ is a compact subset of $\mathcal{V}$, by Proposition B.10 we can find some $\epsilon > 0$ such that $K \times B_\epsilon \times [0, \epsilon) \subseteq \mathcal{V}$. Now let $I_{max}(x, u) = [0, T)$ and choose any $t_0 < T$ satisfying $T - t_0 < \epsilon$ and $\int_{t_0}^T |u|^2 \, d\mathcal{L}^1 < \epsilon$. Let

$$\overline{u}(t) := \begin{cases} u(t_0 + t) & \text{if } t \in [0, T - t_0] \\ 0 & \text{if } t \in (T - t_0, 1] \end{cases}$$

so that $\|\overline{u}\|_2 < \epsilon$. By our choice of ϵ, $I_{max}(\gamma(t_0), \overline{u}) \supseteq [0, T - t_0]$. Thus, denoting by $\overline{\gamma}$ the trajectory associated to $(\gamma(t_0), \overline{u})$, the new curve

$$\tilde{\gamma}(t) := \begin{cases} \gamma(t) & \text{if } t \in [0, t_0] \\ \overline{\gamma}(t - t_0) & \text{if } t \in [t_0, T] \end{cases}$$

(which solves the differential equation on $[0, T]$) contradicts the maximality of γ. $\square$

Now we prove that the flow is smooth in the spatial variable.

PROPOSITION B.12. The map $x \mapsto \mathrm{End}_t(x, u)$ is differentiable and its differential is continuous on $\mathcal{V}$.

Proof. It suffices to show the statement when $M = \mathbb{R}^n$: then in the general case we split $[0, t]$ into finitely many subintervals $[t_i, t_{i+1}]$ whose images are contained in domains of local charts and write the flow as a composition of differentiable maps (strictly speaking, since the involved vector field is time-dependent, each of these maps is the flow from time t_i to t_{i+1}; one can also write this as the flow at time $t_{i+1} - t_i$ upon translating u suitably). In what follows we use the notation $\Phi_t(x) := \mathrm{End}_t(x, u)$, omitting the dependence on u for simplicity. We also call γ the trajectory associated to (x, u), as usual. Let $W \in H^1([0, t], \mathbb{R}^{n \times n})$ solve the linearized equation

$$\dot{W}(s) = \sum_i u_i(s) dX_i(\gamma(s)) W(s)$$

and $W(0) = I$ (we think of $dX_i(\gamma(s))$ as a matrix, as well). Existence and uniqueness of the solution to this equation follows from the previous propositions, since the couple (γ, W) solves

$$\begin{cases} \dot{\gamma}(s) = \sum_i u_i(s) X_i(\gamma(s)) \\ \dot{W}(s) = \sum_i u_i(s) dX_i(\gamma(s)) W(s) \end{cases} \tag{B.2}$$

and this can be interpreted as an equation (for an $\mathbb{R}^n \times \mathbb{R}^{n\times n}$-valued curve) of the type considered in this section, for suitable vector fields in $\mathbb{R}^n \times \mathbb{R}^{n\times n}$. To prove that W is defined on the whole $[0,t]$, it suffices to show that it stays bounded and then argue as in the proof of Lemma B.9. But this follows from the next lemma.
We now prove that $d\Phi_t = W(t)$: in fact, fixing a small $\delta \in \mathbb{R}^n$ and defining

$$z(s) := \Phi_s(x+\delta) - \Phi_s(x) - W(s)\delta,$$

z solves

$$\begin{aligned}\dot{z}(s) &= \sum_i u_i(s) X_i\left(\Phi_s(x+\delta)\right) - \sum_i u_i(s) X_i\left(\Phi_s(x)\right) - \sum_i u_i(s) dX_i\left(\Phi_s(x)\right) W(s)\delta \\ &= \left(\sum_i u_i(s) dX_i\left(\Phi_s(x)\right) W(s)\right) z(s) + \sum_i u_i(s)\left(X_i\left(\Phi_s(x+\delta)\right) - X_i\left(\Phi_s(x)\right)\right. \\ &\quad \left. - dX_i\left(\Phi_s(x)\right)\left[\Phi_s(x+\delta) - \Phi_s(x)\right]\right)\end{aligned}$$

for a.e. s and $z(0) = 0$. Continuity of the flow implies that the second term can be bounded by $\|u\|_2\, o(|\delta|)$, the estimate being uniform over $[0,t]$. The next lemma then gives $|z(t)| = o(|\delta|)$ (the implied constant depends on both x and u), so $d\Phi_t = W(t)$. Finally, continuity over $\mathcal{V}$ follows again from the previous continuity result applied to the augmented problem (B.2) with initial value $(\gamma, W)(0) = (x, I)$. □

Lemma B.13 (Gronwall's inequality). If $\sigma \in H^1([0,T],\mathbb{R}^N)$ satisfies $|\dot\sigma(s)| \le \alpha(s)\,|\sigma(s)| + \beta(s)$ a.e. for some nonnegative $\alpha, \beta \in L^2([0,T])$, we have the estimate

$$\|\sigma\|_\infty \le \exp\left(\int_0^T \alpha(s)\,ds\right)|\sigma(0)| + \int_0^T \exp\left(\int_s^T \alpha(s')\,ds'\right)\beta(s)\,ds.$$

Proof. Let $\rho(s) := |\sigma(s)|$, which belongs to $H^1([0,T])$ since it is obtained by composing σ with a Lipschitz function. We claim that $\dot\rho(s) \le |\dot\sigma(s)|$ for a.e. s: in fact, assuming that both ρ and σ are differentiable at s, if $\rho(s) = 0$ we have $\dot\rho(s) = 0$ (since s is a local minimum), while otherwise $\dot\rho(s) = \left\langle \frac{\sigma(s)}{|\sigma(s)|}, \dot\sigma(s)\right\rangle$. So we obtain

$$\dot\rho(s) \le \alpha(s)\rho(s) + \beta(s)$$

a.e., which is equivalent to

$$\frac{d}{ds}\left(\rho(s)\exp\left(-\int_0^s \alpha(s')\,ds'\right)\right) \le \exp\left(-\int_0^s \alpha(s')\,ds'\right)\beta(s).$$

Integrating both sides of the inequality we get

$$\rho(t) \le \exp\left(\int_0^t \alpha(s)\,ds\right)\rho(0) + \int_0^t \exp\left(\int_s^t \alpha(s')\,ds'\right)\beta(s)\,ds$$

and we are done since the right-hand side is increasing in t. □

Remark B.14. Since the couple $(\Phi_t(x), d\Phi_t(x))$ solves (B.2), Proposition B.12 applied to this differential equation implies that $x \mapsto d\Phi_t(x)$ is C^1 and that the second differential is continuous on $\mathcal{V}$. Iterating this argument, we obtain that $x \mapsto \mathrm{End}_t(x,u)$ is smooth and that all the spatial derivatives are continuous on $\mathcal{V}$.

Remark B.15. A similar device can be used to obtain smoothness in the variable u: let us assume (without loss of generality) $M = \mathbb{R}^n$. For any $u, v \in L^2([0,1],\mathbb{R}^r)$ we consider the augmented system

$$\begin{cases} \dot\gamma(t) = \sum_{i=1}^r u_i(t)X_i(\gamma(t)) + \sum_{i=1}^r v_i(t)sX_i(\gamma(t)) \\ \dot s = 0 \end{cases}$$

which can be considered as an ordinary differential equation of the form considered in this section, with M replaced by $M \times \mathbb{R}$, γ replaced by (γ, s) and with the $2r$ vector fields $X_1(\gamma), \ldots, X_r(\gamma), sX_1(\gamma), \ldots, sX_r(\gamma)$ (extended with a zero in the additional component). The new control is $(u_1, \ldots, u_r, v_1, \ldots, v_r)$. The solution to this system is the trajectory associated to the control $u + sv$ (notice that s is constant by the second equation). Proposition B.12 allows us to differentiate with respect to s (regarded as the initial condition for the same variable): hence the directional derivative $\delta(t) := \frac{\partial \operatorname{End}_t}{u}(x,u)[v]$ exists and the couple (γ, δ) satisfies

$$\begin{cases} \dot\gamma(t) = \sum_{i=1}^r u_i(t)X_i(\gamma(t)) \\ \dot\delta(t) = \sum_{i=1}^r u_i(t)dX_i(\gamma(t))[\delta(t)] + \sum_{i=1}^r v_i(t)X_i(\gamma(t)) \end{cases}$$

(which is obtained by writing $\begin{pmatrix} \delta \\ \sigma \end{pmatrix}(t) := W(t)\begin{pmatrix} 0 \\ 1 \end{pmatrix}$, where $W(t)$ is given by (B.2) applied to our augmented system, and noticing that $\sigma(t) \equiv 1$), with initial condition $\gamma(0) = x$, $\delta(0) = 0$. This differential equation for the couple (γ, δ) falls again in the class considered in this section, so Proposition B.10 can be applied. We deduce that $\operatorname{End}_t(x,u)$ is Gâteaux differentiable in u, with continuous differential. Hence, it is C^1 in the variable u. Iteration of this device shows that, in fact, $\operatorname{End}_t(x,u)$ is C^∞ in u.

By combining the two previous remarks, we finally obtain the following smoothness result.

Corollary B.16 (smoothness). The map $(x,u) \mapsto \operatorname{End}_t(x,u)$ is C^∞-regular on the set $\{(x,u) : (x,u,t) \in \mathcal{V}\}$, for any $t \in [0,1]$, and its differentials of any order are continuous on $\mathcal{V}$.

APPENDIX C

Some useful formulas for flows of vector fields

In this section we state and prove some formulas for time-dependent vector fields. Except for Propositions C.10 and C.13, for simplicity they will be always assumed to be smooth, which is sufficient for our needs. We will also assume $M = \mathbb{R}^n$, even if it is clear from their proofs that Propositions C.9, C.10, C.11, C.12 and C.13 make sense and hold on a generic smooth manifold, as well.

REMARK C.1. Notice that the smoothness of the flow can be deduced from the results of the previous section, since the flow of a time-dependent vector field can be viewed as that of an autonomous vector field by adding an auxiliary spatial variable: $\gamma(t)$ solves $\dot{\gamma}(t) = X_t(\gamma(t))$ iff the couple $(\gamma(t), s(t))$, with $s(t) := t$, solves

$$\begin{cases} \dot{\gamma}(t) = X_s(\gamma(t)) \\ \dot{s} = 1 \end{cases}$$

which is the equation defining the flow of the autonomous vector field $(x, s) \mapsto (X_s(x), 1)$ on $M \times \mathbb{R}$.

When time-dependent vector fields are involved, we will use a notation such as X_t, in order to emphasize that they are not autonomous.

DEFINITION C.2. Given a smooth time-dependent vector field X_t, defined on an open subset of $M \times \mathbb{R}$ (or on an open subset of $M \times I$ for some interval I), and given two times $a \leq b$, assume that

$$\begin{cases} \dot{\gamma}(t) = X_t(\gamma(t)) \\ \gamma(a) = x \end{cases}$$

has a solution defined on $[a, b]$. We define $\Phi_{a,b}(x, X_t) := \gamma(b)$. We also use the notation $\Phi_{a,b}(X_t)(x) := \Phi_{a,b}(x, X_t)$. When X is autonomous, its flow will be simply denoted by $\Phi_t(x, X)$ or $\Phi_t(X)(x)$.

REMARK C.3. A similar definition can be given when $b < a$. In this way the map $(x, a, b) \mapsto \Phi_{a,b}(x, X_t)$ is smooth and is defined on an open subset of $M \times \mathbb{R} \times \mathbb{R}$. We remark that we have the semigroup property

$$\Phi_{b,c}(X_t) \circ \Phi_{a,b}(X_t) = \Phi_{a,c}(X_t).$$

Moreover, for an autonomous vector field we clearly have $\Phi_{a,b}(X) = \Phi_{b-a}(X)$.

DEFINITION C.4. Given a diffeomorphism $\phi : U \to U'$ between two open subsets U, U' of $\mathbb{R}^n$ (or of a smooth manifold) and a smooth vector field X defined on U', we define its *pullback* $\phi^* X := (d\phi)^{-1}[X]$, i.e.

$$\phi^* X(x) := (d\phi_x)^{-1}[X(\phi(x))],$$

which is a smooth vector field on U. Conversely, given a smooth vector field X on U, we define its *pushforward* $\phi_* X := d\phi[X]$, i.e.

$$\phi_* X(y) := d\phi_{\phi^{-1}(y)}[X(\phi^{-1}(y))],$$

which is a smooth vector field on U'.

Before stating the next proposition, it is useful to introduce the following notation. For any $t_1, \dots, t_k \in [0,T]$ we define the vector field $Y_{t_1,\dots,t_k}$ recursively as follows: for any $k \geq 2$ we set

$$Y_{t_1,\dots,t_k} := dY_{t_1,\dots,t_{k-1}}[Y_{t_k}].$$

So, for instance, $Y_{t_1,t_2}(x) = dY_{t_1}(x)[Y_{t_2}(x)]$ for any x.

PROPOSITION C.5 (Volterra's expansion). For any smooth, nonautonomous vector field $Y_t : U \times [0,T] \to \mathbb{R}^n$ and any x such that $\Phi_{0,T}(x, Y_t)$ is defined we have

$$\Phi_{0,T}(x, Y_t) = x + \sum_{j=1}^{k} \int_0^T \int_0^{t_1} \cdots \int_0^{t_{j-1}} Y_{t_1,\dots,t_j}(x)\, dt_j \, \cdots \, dt_1 + O\left(\max_{t \in [0,T]} \|Y_t\|_{C^k(K)}^{k+1} \right),$$

for any $k \geq 0$. Here K is any compact set containing the whole trajectory $\tau \mapsto \Phi_{0,\tau}(x, Y_t)$ (for $\tau \in [0,T]$) and, in evaluating the C^k norm, only the spatial derivatives of Y_t are taken into account. The implied constant in the error depends only on k.

Proof. Let us call $\gamma(\tau) := \Phi_{0,\tau}(x, Y_t)$ for any $\tau \in [0,T]$. We know that

$$\Phi_{0,\tau}(x, Y_t) = x + \int_0^\tau Y_{t_1}(\gamma(t_1))\, dt_1.$$

This same formula, with τ replaced by t_1, shows that

$$Y_{t_1}(\gamma(t_1)) = Y_{t_1}(x) + \int_0^{t_1} dY_{t_1}[Y_{t_2}](\gamma(t_2))\, dt_2.$$

So we get

$$\Phi_{0,T}(x, Y_t) = x + \int_0^T Y_{t_1}(\gamma(t_1))\, dt_1 = x + \int_0^T Y_{t_1}(x)\, dt_1 + \int_0^T \int_0^{t_1} Y_{t_1,t_2}(\gamma(t))\, dt_2\, dt_1.$$

Iterating this computation, we arrive at

$$\begin{aligned} \Phi_{0,T}(x, Y_t) =& x + \sum_{j=1}^{k} \int_0^T \int_0^{t_1} \cdots \int_0^{t_{j-1}} Y_{t_1,\dots,t_j}(x)\, dt_j \, \cdots \, dt_1 \\ &+ \int_0^T \int_0^{t_1} \cdots \int_0^{t_k} Y_{t_1,\dots,t_{k+1}}(\gamma(t))\, dt_{k+1} \, \cdots \, dt_1. \end{aligned}$$

The thesis follows from the fact that, by applying repeatedly Leibniz's rule,

$$\left| Y_{t_1,\dots,t_{k+1}}(\gamma(t)) \right| \leq C \max_{t \in [0,T]} \|Y_t\|_{C^k(K)}^{k+1}.$$

$\square$

We now obtain an expansion for the pullback of an autonomous vector field by the flow of another autonomous vector field. Let us define inductively $\mathrm{ad}(Z)^k Y := \left[Z, \mathrm{ad}(X)^{k-1} Y\right]$ and $\mathrm{ad}(Z)^0 Y := Y$.

PROPOSITION C.6 (Ad expansion). We have

$$\Phi_t(Z)^* Y(x) = \sum_{j=0}^{k} \frac{t^j}{j!} \mathrm{ad}(Z)^j Y(x) + O\left(t^{k+1}\right).$$

More precisely,

$$\left\| \Phi_t(Z)^* Y - \sum_{j=0}^{k} \frac{t^j}{j!} \operatorname{ad}(Z)^j Y \right\|_{C^m(K')} = O\left(t^{k+1} \|Y\|_{C^{m+k+1}(K'')}\right)$$

for any compact set K' where $\Phi_t(Z)$ is defined and any compact neighbourhood K'' of $\{\Phi_s(Z)(x) \mid x \in K, s \in [0,t]\}$. The implied constant depends only on k, m and Z.

Proof. Let us recall the formula

$$\frac{d}{dh}\Phi_h(Z)^* Y = [Z, Y],$$

which is a simple corollary of Proposition C.10 below. For any $s \in [0,t]$, using the identity $\Phi_{s+h}(Z) = \Phi_h(Z) \circ \Phi_s(Z)$, we obtain

$$\frac{d}{ds}\Phi_s(Z)^* Y = \frac{d}{dh}\Phi_s(Z)^* \Phi_h(Z)^* Y \bigg|_{h=0} = \Phi_s(Z)^*[Z, Y].$$

Iterating this formula, we deduce

$$\frac{d^j}{ds^j}\Phi_s(Z)^* Y = \operatorname{ad}(X)^j Y.$$

Now the integral form of Taylor's formula tells us that

$$\Phi_t(Z)^* Y = \sum_{j=0}^{k} \frac{t^j}{j!} \operatorname{ad}(X)^j Y + \int_0^t \frac{(t-s)^k}{k!} \Phi_s(Z)^* \operatorname{ad}(Z)^{k+1}(Y)\, ds,$$

which is enough to conclude (using the fact that $\left\|\operatorname{ad}(Z)^j Y\right\|_{C^m(K'')} = O\left(\|Y\|_{C^{m+j}(K'')}\right)$). $\square$

By combining the two propositions just proved, we deduce two corollaries which are used in Section 4.4.

Corollary C.7. Fix two smooth nonautonomous vector fields Y_t and Z_t and let $\alpha, \beta \in C^\infty([0,T])$. If $\|\alpha\|_\infty$ is sufficiently small (depending on β), then

$$\begin{aligned}\Phi_{0,T}\left(x, \beta(t)\Phi_{\alpha(t)}(Z_t)^* Y_t\right) =& \Phi_{0,T}\left(x, \beta(t)\left(Y_t + \alpha(t)[Z_t, Y_t] + \frac{\alpha^2(t)}{2}[Z_t, [Z_t, Y_t]]\right)\right) \\ &+ O(\|\beta\|_\infty \|\alpha\|_\infty^3)\end{aligned}$$

locally uniformly in x. The implied constant depends only on Y_t and Z_t.

Corollary C.8. Fix three smooth nonautonomous vector fields Y_t, Z_t, Z_t' and let $\alpha, \alpha', \beta \in C^\infty([0,T])$. If $\|\alpha\|_\infty, \|\alpha'\|_\infty$ are sufficiently small (depending on β), then

$$\begin{aligned}&\Phi_{0,T}\left(x, \beta(t)\Phi_{\alpha(t)}(Z_t)^* \Phi_{\alpha'(t)}(Z_t')^* Y_t - \beta(t) Y_t\right) \\ &= \Phi_{0,T}\Bigg(x, \beta(t)\bigg(\alpha(t)[Z_t, Y_t] + \alpha'(t)[Z_t', Y_t] + \frac{\alpha^2(t)}{2}[Z_t, [Z_t, Y_t]] + \alpha(t)\alpha'(t)[Z_t, [Z_t', Y_t]] \\ &\qquad + \frac{\alpha'^2(t)}{2}[Z_t', [Z_t', Y_t]]\bigg)\Bigg) + O(\|\beta\|_\infty (\|\alpha\|_\infty^3 + \|\alpha'\|_\infty^3))\end{aligned}$$

locally uniformly in x. The implied constant depends only on Y_t, Z_t and Z_t'.

PROPOSITION C.9. Given two smooth autonomous vector fields $Y, Z : U \to \mathbb{R}^n$, we have the identities

$$\frac{d}{dt}\Phi_t(Z)^*Y = \Phi_t(Z)^*[Z,Y],$$
$$\frac{d}{dt}\Phi_{T-t}(Z)_*Y = \Phi_{T-t}(Z)_*[Z,Y]$$

on the domains of definition of $\Phi_t(Z)$ and $\Phi_{t-T}(Z)$, respectively. In particular,

$$\Phi_t(Y)^*Y = \Phi_{T-t}(Y)_*Y = Y.$$

Proof. By the semigroup property we have

$$\frac{d}{dt}\Phi_t(Z)^*Y = \Phi_t(Z)^*\left(\frac{d}{dh}\Phi_h(Z)^*Yh = 0\right) = \Phi_t(Z)^*[Z,Y].$$

The second identity is obtained similarly using the fact that

$$\Phi_{T-(t+h)}(Z)_*Y = \Phi_{T-(t+h)}(Z)_*\Phi_h(Z)_*\Phi_h(Z)^*Y = \Phi_{T-t}(Z)_*\Phi_h(Z)^*Y.$$

□

PROPOSITION C.10. Let $u \in L^2([0,T],\mathbb{R}^r)$ and $X_1,\dots,X_r$ smooth vector fields as in the previous section. Let Y be a smooth autonomous vector field. On the domain of definition of $\Phi_{0,T}\langle u(t),X\rangle$, for any Lebesgue point τ of u the following holds:

$$\left.\frac{d}{dh}\Phi_{\tau,\tau+h}\left(\langle u(\tau),X\rangle\right)^* Y\right|_{h=0} = \left[\langle u(\tau),X\rangle, Y\right].$$

Proof. Fix x such that $\Phi_{0,T}(x,\langle u(t),X\rangle)$ is defined. Call γ the corresponding integral curve and let $J_h := d\Phi_{\tau,\tau+h}\left(\langle u(t),X\rangle\right)$. As we saw in the proof of Proposition B.12, J_h solves

$$\dot J_h = \langle u(\tau+h), dX(\gamma(\tau+h))\rangle J_h$$

for a.e. h. Recall now that the map $A \mapsto A^{-1}$ (from $GL_n(\mathbb{R})$ to itself) is smooth and its differential at A maps $B \in \mathbb{R}^{n\times n}$ to $-A^{-1}BA^{-1}$. Hence,

$$\begin{aligned}\frac{d}{dh}\left(J_h^{-1}Y(\gamma(\tau+h))\right) =& -J_h^{-1}\dot J_h J_h^{-1}Y(\gamma(t+h)) + J_h^{-1}dY\left[\dot\gamma(\tau+h)\right]\\ =& J_h^{-1}\left(-\langle u(\tau+h),dX\rangle[Y] + dY\left[\langle u(\tau+h),X\rangle\right]\right)\end{aligned}$$

for a.e. h (in the last line all vector fields and their differentials are evaluated at $\gamma(\tau+h)$). Since $dY[X_i] - dX_i[Y] = [X_i,Y]$, taking into account that τ is a Lebesgue point we finally obtain

$$\begin{aligned}J_h^{-1}Y(\gamma(\tau+h)) &= Y(\gamma(\tau)) + \int_0^h \sum_{i=1}^r u_i(\tau+h')J_{h'}^{-1}[X_i,Y](\gamma(\tau+h'))\,dh'\\ &= Y(\gamma(\tau)) + \sum_{i=1}^r\left(\int_0^h u_i(\tau+h')\,dh'\right)[X_i,Y] + o(h)\\ &= Y(\gamma(\tau)) + h\sum_{i=1}^r u_i(\tau)[X_i,Y] + o(h)\end{aligned}$$

(as $h \mapsto J_h^{-1}Y(\gamma(\tau+h))$ is H^1). □

Let us now state and prove two important *variation formulas*, which tell us how the flow behaves when adding a perturbation to a given vector field Y_t. Both of them allow to factor out the flow of Y_t in the result, in a suitable sense. As the proof shows, they make sense and hold also on a generic smooth manifold.

Proposition C.11. Given two smooth nonautonomous vector fields Y_t and Z_t, assume that $\Phi_{0,T}(x, Y_t)$, $\Phi_{0,T}(x, Y_t + Z_t)$ are both defined and that $\Phi_{0,t}(x, Y_\tau + Z_\tau)$ belongs to the domain of $\Phi_{0,t}(Y_\tau)^{-1}$ for any $t \in [0, T]$. Then

$$\Phi_{0,T}(Y_t + Z_t)(x) = \Phi_{0,T}(Y_t) \circ \Phi_{0,T}\left(\Phi_{0,t}(Y_\tau)^* Z_t\right)(x).$$

The fact that the right-hand side is defined is part of the thesis.

Proof. Let us define

$$\widehat{\gamma}(t) := \Phi_{0,t}(Y_\tau)^{-1} \circ \Phi_{0,t}(Y_\tau + Z_\tau)(x).$$

This definition can be rewritten as

$$\Phi_{0,t}(Y_\tau)\left(\widehat{\gamma}(t)\right) = \Phi_{0,t}(Y_\tau + Z_\tau)(x). \tag{C.1}$$

Differentiation in time yields

$$d\Phi_{0,t}(Y_\tau)\left[\dot{\widehat{\gamma}}(t)\right] + Y_t\left(\Phi_{0,t}(Y_\tau)\left(\widehat{\gamma}(t)\right)\right) = (Y_t + Z_t)\left(\Phi_{0,t}(Y_\tau)\left(\widehat{\gamma}(t)\right)\right)$$

(since the left-hand side of (C.1) can be seen as the composition of $(s, s') \mapsto \Phi_{0,s}(Y_\tau)\left(\widehat{\gamma}(s')\right)$ with the diagonal map $t \mapsto (t, t)$). We finally get

$$\dot{\widehat{\gamma}}(t) = \left(\Phi_{0,t}(Y_\tau)^* Z_t\right)\left(\widehat{\gamma}(t)\right),$$

which gives the thesis, since $\widehat{\gamma}(0) = x$. □

Proposition C.12. Given two smooth nonautonomous vector fields Y_t and Z_t, assume that $\Phi_{0,T}(x, Y_t)$, $\Phi_{0,T}(x, Y_t + Z_t)$ are both defined and that $\Phi_{0,t}(x, Y_\tau + Z_\tau)$ belongs to the domain of $\Phi_{t,T}(Y_\tau)$ for any $t \in [0, T]$. Then

$$\Phi_{0,T}(Y_t + Z_t)(x) = \Phi_{0,T}\left(\Phi_{t,T}(Y_\tau)_* Z_t\right) \circ \Phi_{0,T}(Y_t)(x).$$

The fact that the right-hand side is defined is part of the thesis.

Proof. We define

$$\widehat{\gamma}(t) := \Phi_{t,T}(Y_\tau) \circ \Phi_{0,t}(Y_\tau + Z_\tau)(x),$$

which can be rewritten as

$$\Phi_{t,T}(Y_\tau)^{-1}\left(\widehat{\gamma}(t)\right) = \Phi_{0,t}(Y_\tau + Z_\tau)(x).$$

Differentiating as in the previous proof and keeping in mind that $\Phi_{t,T}(Y_\tau)^{-1} = \Phi_{T,t}(Y_\tau)$ (see also Remark C.3), we get

$$d\Phi_{t,T}(Y_\tau)^{-1}\left[\dot{\widehat{\gamma}}(t)\right] + Y_t\left(\Phi_{t,T}(Y_\tau)^{-1}\left(\widehat{\gamma}(t)\right)\right) = (Y_t + Z_t)\left(\Phi_{t,T}(Y_\tau)^{-1}\left(\widehat{\gamma}(t)\right)\right).$$

This implies

$$\dot{\widehat{\gamma}}(t) = \left(\Phi_{t,T}(Y_\tau)_* Z_t\left(\widehat{\gamma}(t)\right)\right)$$

and, as $\widehat{\gamma}(0) = \Phi_{0,T}(Y_\tau)(x)$ and $\widehat{\gamma}(T) = \Phi_{0,T}(Y_t + Z_t)(x)$, we are done. □

We conclude with a simple observation.

Proposition C.13. If $\gamma : [0, T] \to \mathbb{R}^n$ solves $\dot{\gamma}(t) = \dot{v}(t) Y(\gamma(t))$ for some smooth autonomous vector field Y and some $v \in H^1([0, T])$ with $v(0) = 0$, then $\gamma(0)$ belongs to the domain of $\Phi_{v(T)}(Y)$ and $\gamma(T) = \Phi_T(\gamma(0), Y)$. So we have the formula

$$\Phi_{0,T}(\dot{v}(t) Y) = \Phi_{v(T)}(Y).$$

Proof. Let $a := \inf\{s : \Phi_s(\gamma(0), Y) \text{ is defined}\}$ and define b similarly with inf replaced by sup, so that $a < 0 < b$. Assume by contradiction that $v(t) \notin (a, b)$ for some t and let $\bar{t} := \min\{t : v(t) \notin (a, b)\}$. Let us assume for instance that $v(\bar{t}) = b$. For a.e. $t \in [0, \bar{t})$ we have

$$\frac{d}{dt}\Phi_{v(t)}(\gamma(0), Y) = \dot{v}(t)Y\left(\Phi_{v(t)}(\gamma(0), Y)\right),$$

so $\Phi_{v(t)}(\gamma(0), Y) = \gamma(t)$. Since $v(t) \to b$ as $t \uparrow \bar{t}$, from Proposition B.11 we deduce that $\Phi_b(\gamma(0), Y)$ is defined, as well. This is a contradiction, since then $\Phi_{b+\epsilon}(\gamma(0), Y)$ is defined as well, for all sufficiently small ϵ. This shows that $v([0, T]) \subseteq (a, b)$. Thus the above equation makes sense and holds on all $[0, T]$, proving the thesis. $\square$

APPENDIX D

The Baker-Campbell-Hausdorff formula

This section is devoted to the proof of the celebrated Baker-Campbell-Hausdorff formula for any Lie group $\mathbb{G}$, which expresses the product $\exp(X)\exp(Y)$ as the exponential of some $Z \in \mathfrak{g}$, whose explicit formula depends only on the Lie algebra structure of $\mathfrak{g}$ (this is in accordance with the general fact that a simply connected Lie group is determined by its Lie algebra, up to isomorphism). We will also show that, for nilpotent Lie groups, it holds without the requirement that X, Y are small. At the same time, we will obtain that $\exp : \mathfrak{g} \to \mathbb{G}$ is a diffeomorphism if $\mathbb{G}$ is also simply connected.

Given a Lie group $\mathbb{G}$, we will denote by L_a and R_a the left and right multiplication by a fixed element a, so that $L_a(x) := ax$ and $R_a(x) := xa$. Notice that $L_{ab} = L_a \circ L_b$ and $R_{ab} = R_b \circ R_a$. We use the notation C_a for the conjugation by a, namely $C_a(x) := axa^{-1}$. When writing the differential of one of these maps, we will often omit the point where it is calculated if it is clear from the context (e.g. when applying the chain rule).

Let us call $\mathfrak{g} := T_e\mathbb{G}$ the Lie algebra associated to $\mathbb{G}$. We recall that the *adjoint representation* of $\mathbb{G}$, $\mathrm{Ad} : \mathbb{G} \to GL(\mathfrak{g})$, is given by $\mathrm{Ad}(g) := d(C_g)_e$ and is a group homomorphism. When V is a finite-dimensional real vector space, the Lie algebra associated to $GL(V)$ can be canonically identified with the space of linear endomorphisms of V and is usually denoted by $\mathfrak{gl}(V)$. We also recall that $\mathrm{ad} : \mathfrak{g} \to \mathfrak{gl}(\mathfrak{g})$ is defined as $\mathrm{ad}(X) := d\,\mathrm{Ad}_e$ and satisfies the identity $\mathrm{ad}(X)[Y] = [X, Y]$.

For any $X \in \mathfrak{g}$ we define $\exp(X) := \Phi_1(X^L)$, where X^L is the left-invariant vector field associated to X and $\phi_t(X^L)$ is the corresponding flow. Equivalently, $\exp(X) = \gamma_X(1)$, where $\gamma_X : \mathbb{R} \to \mathbb{G}$ is the unique Lie group homomorphism satisfying $\dot\gamma(0) = X$. The map $\exp : \mathfrak{g} \to \mathbb{G}$ is smooth and is the so-called *exponential map* of $\mathbb{G}$.

Remark D.1. If $f : \mathbb{G} \to \mathbb{G}'$ is a Lie group homomorphism, we have the well-known relation $\exp_{\mathbb{G}'} \circ df_e = f \circ \exp_{\mathbb{G}}$. Applying this with $\mathbb{G}' := GL(\mathfrak{g})$, $f := \mathrm{Ad}$ and recalling that $\exp_{GL(\mathfrak{g})}(A) = e^A$, we deduce that

$$e^{\mathrm{ad}(X)} = \mathrm{Ad}(\exp(X))$$

for any $X \in \mathfrak{g}$.

Definition D.2. Given a smooth manifold M and a smooth map $f : M \to \mathbb{G}$, its *(right) logarithmic derivative* at $x \in M$ is the linear map $\delta f(x) : T_xM \to \mathfrak{g}$ defined as

$$\delta f(x) := dR_{f(x)^{-1}} \circ df_x.$$

Lemma D.3. The logarithmic derivative satisfies this identity, for any $f, g \in C^\infty(M, G)$:

$$\delta(fg)(x) = \delta f(x) + \mathrm{Ad}(f(x))\delta g(x).$$

Proof. The left-hand side equals

$$\begin{aligned} dR_{g(x)^{-1}f(x)^{-1}} \circ d(fg)_x &= dR_{f(x)^{-1}} \circ dR_{g(x)^{-1}} \circ \left(dR_{g(x)} \circ df_x + dL_{f(x)} \circ dg_x\right) \\ &= \delta f(x) + \mathrm{Ad}(f(x))\delta g(x), \end{aligned}$$

since (noticing that $L_a R_b = R_b L_a$ for any $a, b \in G$, by associativity of the group multiplication)

$$dR_{f(x)^{-1}} \circ dR_{g(x)^{-1}} \circ dL_{f(x)} \circ dg_x = dR_{f(x)^{-1}} \circ dL_{f(x)} \circ dR_{g(x)^{-1}} \circ dg_x = dC_{f(x)} \circ \delta g(x).$$

□

In what follows we will canonically identify $T_X\mathfrak{g}$ with $\mathfrak{g}$, for any $X \in \mathfrak{g}$.

Lemma D.4. Let $g(z) := \frac{e^z - 1}{z} = \sum_{p=0}^{\infty} \frac{z^p}{(p+1)!}$ (for $z \in \mathbb{C}$); we define $g(A)$ for any $A \in \mathfrak{gl}(\mathfrak{g})$ by the analogous formula $\sum_{p=0}^{\infty} \frac{A^p}{(p+1)!}$. The following identity holds:

$$\delta \exp(X) = g(\operatorname{ad} X).$$

Proof. Let $\alpha(t) := t\delta\exp(tX) \in \mathfrak{gl}(\mathfrak{g})$. By Lemma D.3 and the chain rule we have

$$\begin{aligned}
\alpha(s+t) &= (s+t)\delta\exp((s+t)X) \\
&= \delta\exp((s+t)\cdot)(X) \\
&= \delta\exp(s\cdot)(X) + \operatorname{Ad}(\exp(sX))\,\delta\exp(t\cdot)(X) \\
&= s\delta\exp(sX) + t\operatorname{Ad}(\exp(sX))\,\delta\exp(tX) \\
&= \alpha(s) + \operatorname{Ad}(\exp(sX))\,\alpha(t).
\end{aligned}$$

Hence, differentiating in s and evaluating at $s = 0$,

$$\dot{\alpha}(t) = \dot{\alpha}(0) + \operatorname{ad}(X)\alpha(t).$$

Notice that $\dot{\alpha}(0) = \delta\exp(0) = \mathrm{id}_{\mathfrak{g}}$, so the above differential equation can be rewritten as

$$\dot{\alpha}(t) = I + \operatorname{ad}(X)\alpha(t).$$

The initial condition $\alpha(0) = 0$ uniquely determines the solution. Since

$$t \mapsto \sum_{p=0}^{\infty} \frac{s^{p+1}}{(p+1)!} \operatorname{ad}(X)^p$$

is a solution with the same initial condition, we deduce

$$\delta\exp(X) = \alpha(1) = \sum_{p=0}^{\infty} \frac{\operatorname{ad}(X)^p}{(p+1)!} = g(\operatorname{ad} X).$$

□

Theorem D.5. Let $f(z) := \frac{\log z}{z-1} = \sum_{p=0}^{\infty} \frac{(-1)^p}{p+1}(z-1)^p$ (for any complex $z \in B_1(1)$). For any sufficiently small $X, Y \in \mathfrak{g}$ we have

$$\exp(X)\exp(Y) = \exp\left(C(X,Y)\right), \quad C(X,Y) := Y + \int_0^1 f(e^{t\operatorname{ad} X}e^{\operatorname{ad} Y})X\,dt.$$

Here $f(A) := \sum_{p=0}^{\infty} \frac{(-1)^p}{p+1}(A - I)^p$ for any $A \in \mathfrak{gl}(\mathfrak{g})$ with $\|A - I\| < 1$.

Proof. Let U be a neighbourhood of 0 in $\mathfrak{g}$ such that $\exp|_U$ is a diffeomorphism onto $V := \exp(U)$. Let $X, Y \in \mathfrak{g}$ be so small that $\left\|e^{t\operatorname{ad} X}e^{\operatorname{ad} Y} - I\right\| \le \frac{1}{2}$ and $\exp(tX)\exp(Y) \in V$ for all $t \in [0,1]$. We define

$$\beta(t) := \exp^{-1}\left(\exp(tX)\exp(Y)\right)$$

for $t \in [0,1]$. Now we compute $dR_{\exp(-\beta(t))}\left[\frac{d}{dt}\exp(\beta(t))\right]$ in two ways:

$$dR_{\exp(-\beta(t))}\left[\frac{d}{dt}\exp(\beta(t))\right] = \delta\exp(\beta(t))\dot{\beta}(t) = g(\operatorname{ad}\beta(t))\dot{\beta}(t),$$

thanks to Lemma D.4; moreover,

$$dR_{\exp(-\beta(t))}\left[\frac{d}{dt}\exp(\beta(t))\right] = dR_{\exp(-Y)\exp(-tX)}\left[\frac{d}{dt}\left(\exp(tX)\exp(Y)\right)\right]$$
$$= dR_{\exp(-tX)} \circ dR_{\exp(-Y)} \circ dR_{\exp(Y)}\left[\frac{d}{dt}\exp(tX)\right] = X,$$

since $\frac{d}{dt}\exp(tX) = dR_{\exp(tX)}[X]$ (as $\exp((t+s)X) = \exp(sX)\exp(tX)$). Thus we get

$$X = g(\operatorname{ad}\beta(t))\dot\beta(t). \tag{D.1}$$

But, recalling Remark D.1,

$$e^{\operatorname{ad}\beta(t)} = \operatorname{Ad}\left(\exp(\beta(t))\right) = \operatorname{Ad}(\exp(tX))\operatorname{Ad}(\exp(Y)) = e^{t\operatorname{ad}X}e^{\operatorname{ad}Y},$$

so we arrive at

$$\operatorname{ad}\beta(t) = \log(e^{t\operatorname{ad}X}e^{\operatorname{ad}Y}). \tag{D.2}$$

Here, again, we define $\log(A) := \sum_{p=0}^{\infty}\frac{(-1)^p}{p+1}(A-I)^{p+1}$ for all $A \in \mathfrak{gl}(\mathfrak{g})$ with $\|A-I\| < 1$, in analogy with the power series expansion of $\log z$ near 1 (to be precise, in deriving the last identity, we should also guarantee that $\operatorname{ad}\beta(t)$ is sufficiently small, so that $\log(e^{\operatorname{ad}\beta(t)})$ is defined and equals $\operatorname{ad}\beta(t)$). Now we notice that $g(\log z)f(z) = 1$, so this holds also when z is replaced by any such A. In particular, multiplying both sides of (D.1) by $f(e^{t\operatorname{ad}X}e^{\operatorname{ad}Y})$ and using (D.2), we get

$$\dot\beta(t) = f(e^{t\operatorname{ad}X}e^{\operatorname{ad}Y})X.$$

Since $\beta(0) = Y$, we finally have

$$\exp^{-1}\left(\exp(X)\exp(Y)\right) = \beta(1) = \beta(0) + \int_0^1 \dot\beta(t)\,dt = Y + \int_0^1 f(e^{t\operatorname{ad}X}e^{\operatorname{ad}Y})X\,dt.$$

□

Corollary D.6 (Baker-Campbell-Hausdorff formula). If $X, Y \in \mathfrak{g}$ are sufficiently small, then

$$C(X,Y) = X + Y + \sum_{p=1}^{\infty}\frac{(-1)^p}{p+1}\sum_{\substack{k_1,\dots,k_p\ge 0\\ \ell_1,\dots,\ell_p\ge 0\\ k_i+\ell_i\ge 1}}\frac{(\operatorname{ad}X)^{k_1}(\operatorname{ad}Y)^{\ell_1}\cdots(\operatorname{ad}X)^{k_p}(\operatorname{ad}Y)^{\ell_p}}{(k_1+\cdots+k_p+1)k_1!\cdots k_p!\,\ell_1!\cdots\ell_p!}X$$

and this double series converges absolutely.

Proof. It suffices to notice that

$$f(e^{t\operatorname{ad}X}e^{\operatorname{ad}Y}) = X + \sum_{p=1}^{\infty}\frac{(-1)^p}{p+1}\left(\sum_{\substack{k,\ell\ge 0\\ k+\ell\ge 1}}\frac{t^k}{k!\,\ell!}(\operatorname{ad}X)^k(\operatorname{ad}Y)^\ell\right)^p$$
$$= X + \sum_{p=1}^{\infty}\frac{(-1)^p}{p+1}\sum_{\substack{k_1,\dots,k_p\ge 0\\ \ell_1,\dots,\ell_p\ge 0\\ k_i+\ell_i\ge 1}}\frac{(\operatorname{ad}X)^{k_1}(\operatorname{ad}Y)^{\ell_1}\cdots(\operatorname{ad}X)^{k_p}(\operatorname{ad}Y)^{\ell_p}}{k_1!\cdots k_p!\,\ell_1!\cdots\ell_p!}t^{k_1+\cdots+k_p}X.$$

The last equality is justified provided that we know that the last inner sum converges absolutely. Moreover, the double summation can be interchanged with the integral provided that the double series converges absolutely when $t = 1$; if this is the case, the second part of the thesis follows as well. Now, assuming $|X| \le 1$,

$$\begin{aligned}
&\sum_{p=1}^{\infty} \frac{1}{p+1} \sum_{\substack{k_1,\dots,k_p \ge 0 \\ \ell_1,\dots,\ell_p \ge 0 \\ k_i+\ell_i \ge 1}} \left| \frac{(\operatorname{ad} X)^{k_1} (\operatorname{ad} Y)^{\ell_1} \cdots (\operatorname{ad} X)^{k_p} (\operatorname{ad} Y)^{\ell_p}}{k_1! \cdots k_p!\, \ell_1! \cdots \ell_p!} t^{k_1+\dots+k_p} X \right| \\
&\le \sum_{p=1}^{\infty} \frac{1}{p+1} \sum_{\substack{k_1,\dots,k_p \ge 0 \\ \ell_1,\dots,\ell_p \ge 0 \\ k_i+\ell_i \ge 1}} \frac{\|\operatorname{ad} X\|^{k_1+\dots+k_p} \|\operatorname{ad} Y\|^{\ell_1+\dots+\ell_p}}{k_1! \cdots k_p!\, \ell_1! \cdots \ell_p!} t^{k_1+\dots+k_p} \\
&= \sum_{p=1}^{\infty} \frac{1}{n+1} \left(e^{\|\operatorname{ad} X\|} e^{\|\operatorname{ad} Y\|} - 1 \right)^p < +\infty,
\end{aligned}$$

once we also assume that $e^{\|\operatorname{ad} X\|} e^{\|\operatorname{ad} Y\|} \le 1 + \frac{1}{2}$. □

If we sort the terms in the above formula according to their degree (thinking them as $\mathfrak{g}$-valued homogeneous polynomials in X, Y, or simply counting the total number of X's and Y's which appear in each term), the first terms in the expansion are given by

$$C(X,Y) = X + Y + \frac{1}{2}[X,Y] + \frac{1}{12}\left([X,[X,Y]] + [Y,[Y,X]]\right) + \dots,$$

the remainder being formed by terms whose degree is at least four.

If $\mathbb{G}$ is a nilpotent Lie group (i.e. a Lie group whose associated Lie algebra is nilpotent), the double summation which appears in the Baker-Campbell-Hausdorff formula is in fact a finite sum: by definition, there exists some $N \ge 1$ such that

$$(\operatorname{ad} Z_1)(\operatorname{ad} Z_2) \cdots (\operatorname{ad} Z_N) = 0$$

for any $Z_1, \dots, Z_N \in \mathfrak{g}$, so in the formula one can restrict the double sum to the terms where $k_1 + \dots + k_p + \ell_1 + \dots + \ell_p < N$. Since $k_i + \ell_i \ge 1$, this condition implies $p < N$, as well as $k_i, \ell_i < N$, proving that finitely many terms are involved.

Thus, the right-hand side of the Baker-Campbell-Hausdorff formula defines a polynomial function $P : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$. This means that, choosing any basis $X_1, \dots, X_n$ of $\mathfrak{g}$ and writing $X = \sum_i \alpha_i X_i$, $\sum_i \beta_j X_j$, $P(X,Y) = \sum_k P_k(X,Y) X_k$, P_k is a polynomial in the variables α_i, β_j.

Proposition D.7. If $\mathbb{G}$ is a nilpotent group and $P : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ is defined as above, then $\exp(X)\exp(Y) = \exp(P(X,Y))$ for all $X, Y \in \mathfrak{g}$. Moreover, if $\mathbb{G}$ is simply connected, then $\exp : \mathfrak{g} \to \mathbb{G}$ is a diffeomorphism.

Proof. The function $P : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ is a binary operation that makes $\mathfrak{g}$ a Lie group (whose identity element is 0): in fact we have $P(X,0) = P(0,X) = X$ (recall that $(\operatorname{ad} X)^k X = 0$ if $k > 0$) and $P(X,-X) = P(-X,X) = 0$. This binary operation is also associative: when X, Y, Z are small, we have

$$\begin{aligned}
\exp(P(X,P(Y,Z))) &= \exp(X)\exp(P(Y,Z)) \\
&= \exp(X)\exp(Y)\exp(Z) \\
&= \exp(P(X,Y))\exp(Z) \\
&= \exp(P(P(X,Y),Z)),
\end{aligned}$$

so that (since exp is invertible near 0)

$$P(X,P(Y,Z)) = P(P(X,Y),Z).$$

Since both sides of the above equation are polynomial functions, we deduce that it has to hold everywhere. This proves that $\mathfrak{g}$ becomes a Lie group.

In what follows, in order to make the exposition more transparent, we will write explicitly the canonical identification $i : \mathfrak{g} \to T_0\mathfrak{g}$, whenever it occurs. Since $P(sX, tX) = (s+t)X$, the map $t \mapsto tX$ is a one-parameter semigroup of $\mathfrak{g}$, so (denoting by $\exp_{\mathfrak{g}}$ the exponential map in the group $\mathfrak{g}$)

$$\exp_{\mathfrak{g}}(i(X)) = X.$$

Here $\exp_{\mathfrak{g}}$ denotes the exponential map in the group $\mathfrak{g}$. $i : \mathfrak{g} \to T_0\mathfrak{g}$ is also an isomorphism of Lie algebras: indeed,

$$\mathrm{Ad}(X)[i(Y)] = \frac{d}{dt} P(P(X, tY), -X)\Big|_{t=0},$$

$$\mathrm{ad}(i(X))[i(Y)] = \frac{d}{ds}\,\mathrm{Ad}(sX)[i(Y)]\Big|_{s=0} = \frac{d^2}{ds\,dt} P(P(sX, tY), -sX)\Big|_{s,t=0} = i([X, Y])$$

and the left-hand side of the last identity equals the Lie bracket in the Lie algebra $T_0\mathfrak{g}$ of the group $\mathfrak{g}$. Now it is well-known that, since $\mathfrak{g}$ is a simply connected Lie group and $i^{-1} : T_0\mathfrak{g} \to \mathfrak{g} = T_e\mathbb{G}$ is a Lie algebra isomorphism, there exists a unique Lie group homomorphism $h : \mathfrak{g} \to \mathbb{G}$ satisfying $dh_0 = i^{-1}$. But

$$\exp_{\mathbb{G}} \circ dh_0 = h \circ \exp_{\mathfrak{g}}$$

and, as we saw above, $\exp_{\mathfrak{g}} = i^{-1} = dh_0$. Thus $h = \exp_{\mathbb{G}}$ and we finally have

$$\exp_{\mathbb{G}}(P(X, Y)) = h(P(X, Y)) = h(X)h(Y) = \exp_{\mathbb{G}}(X)\exp_G(Y).$$

If $\mathbb{G}$ is simply connected as well, applying the aforementioned fact to $i : T_e\mathbb{G} = \mathfrak{g} \to T_0\mathfrak{g}$, we obtain a homomorphism $h' : \mathbb{G} \to \mathfrak{g}$ with $dh'_e = i$. Now $h'h : \mathfrak{g} \to \mathfrak{g}$ is a group homomorphism with $d(h'h)_0 = \mathrm{id}$, so by uniqueness we obtain $h'h = \mathrm{id}_{\mathfrak{g}}$. Similarly $hh' = \mathrm{id}_{\mathbb{G}}$, proving that $h = \exp_{\mathbb{G}}$ is a diffeomorphism. $\square$

Bibliography

[AS95] A. A. Agrachev and A. V. Sarychev. "Strong minimality of abnormal geodesics for 2-distributions". In: *Journal of Dynamical and Control Systems* 1.2 (1995), pp. 139–176. doi: 10.1007/BF02254637.

[AT04] Luigi Ambrosio and Paolo Tilli. *Topics on analysis in metric spaces*. Oxford University Press, 2004. url: https://global.oup.com/academic/product/topics-on-analysis-in-metric-spaces-9780198529385.

[Bel94] A. Bellaïche. "The tangent space in sub-riemannian geometry". In: *Journal of Mathematical Sciences* 83.4 (1994), pp. 461–476. issn: 1573-8795. doi: 10.1007/BF02589761.

[GK95] Chr Golé and R. Karidi. "A note on Carnot geodesics in nilpotent Lie groups". In: *Journal of Dynamical and Control Systems* 1.4 (1995), pp. 535–549. doi: 10.1007/BF02255895.

[HL16] Eero Hakavuori and Enrico Le Donne. "Non-minimality of corners in sub-riemannian geometry". In: *Inventiones mathematicae* (2016), pp. 1–12. doi: 10.1007/s00222-016-0661-9.

[Jea14] Frédéric Jean. *Control of Nonholonomic Systems: from Sub-Riemannian Geometry to Motion Planning*. SpringerBriefs in Mathematics. Cham: Springer International Publishing, 2014. doi: 10.1007/978-3-319-08690-3_1.

[LLMV13] E. Le Donne et al. "Extremal polynomials in stratified groups". In: *ArXiv e-prints* (July 2013). arXiv: 1307.5235 [math.DG].

[LLMV15] E. Le Donne et al. "Corners in non-equiregular sub-Riemannian manifolds". In: *ESAIM: COCV* 21.3 (2015), pp. 625–634. doi: 10.1051/cocv/2014041.

[LM08] Gian Paolo Leonardi and Roberto Monti. "End-Point Equations and Regularity of Sub-Riemannian Geodesics". In: *Geometric and Functional Analysis* 18.2 (2008), pp. 552–582. doi: 10.1007/s00039-008-0662-y.

[LS95] H. J. Sussmann and Wensheng Liu. "Shortest paths for sub-Riemannian metrics on rank 2 distributions". In: *Mem. Amer. Math. Soc* 564 (1995), p. 104. doi: 10.1090/memo/0564.

[Mit85] John Mitchell. "On Carnot-Carathéodory metrics". In: *Journal of Differential Geometry* 21.1 (1985), pp. 35–45. url: http://projecteuclid.org/euclid.jdg/1214439462.

[Mon14] Roberto Monti. "The regularity problem for sub-Riemannian geodesics". In: *Geometric Control Theory and Sub-Riemannian Geometry*. Ed. by Gianna Stefani et al. Cham: Springer International Publishing, 2014, pp. 313–332. isbn: 978-3-319-02132-4. doi: 10.1007/978-3-319-02132-4_18.

[Mon94] Richard Montgomery. "Abnormal Minimizers". In: *SIAM Journal on Control and Optimization* 32.6 (1994), pp. 1605–1620. doi: 10.1137/S0363012993244945.

[NSW85] Alexander Nagel, Elias M. Stein, and Stephen Wainger. "Balls and metrics defined by vector fields I: Basic properties". In: *Acta Mathematica* 155.1 (1985), pp. 103–147. issn: 1871-2509. doi: 10.1007/BF02392539.

[Str86] Robert S. Strichartz. "Sub-Riemannian geometry". In: *J. Differential Geom.* 24.2 (1986), pp. 221–263. URL: http://projecteuclid.org/euclid.jdg/1214440436.

[Str89] Robert S. Strichartz. "Corrections to: "Sub-Riemannian geometry"". In: *J. Differential Geom.* 30.2 (1989), pp. 595–596. URL: http://projecteuclid.org/euclid.jdg/1214443604.

[Sus14] H. J. Sussmann. "A regularity theorem for minimizers of real-analytic subriemannian metrics". In: *53rd IEEE Conference on Decision and Control*. Dec. 2014, pp. 4801–4806. DOI: 10.1109/CDC.2014.7040138.

[TY13] Kanghai Tan and Xiaoping Yang. "Subriemannian geodesics of Carnot groups of step 3". In: *ESAIM: Control, Optimisation and Calculus of Variations* 19 (1 Jan. 2013), pp. 274–287. ISSN: 1262-3377. DOI: 10.1051/cocv/2012006.

[Vit14] Davide Vittone. "The regularity problem for sub-Riemannian geodesics". In: *Geometric Measure Theory and Real Analysis*. Ed. by Luigi Ambrosio. Pisa: Scuola Normale Superiore, 2014, pp. 193–226. ISBN: 978-88-7642-523-3. DOI: 10.1007/978-88-7642-523-3_4.