Shortest paths for sub-Riemannian metrics on rank-two distributions

Wensheng Liu

Héctor J. Sussmann

Department of Mathematics
Rutgers University
New Brunswick, NJ 08903

wliu@hilbert.rutgers.edu
sussmann@hamilton.rutgers.edu

Table of Contents

ABSTRACT

We study length-minimizing arcs in sub-Riemannian manifolds (M, E, G) whose metric G is defined on a rank-two bracket-generating distribution E. It is well known that all length-minimizing arcs are extremals, and that these extremals are either "normal" or "abnormal." Normal extremals are locally optimal, in the sense that every sufficiently short piece of such an extremal is a minimizer. The question whether every length-minimizer is a normal extremal remained open for several years, and was recently settled by R. Montgomery, who exhibited a counterexample. But Montgomery's geometric optimality proof depends heavily on special properties of his example and still leaves open the question whether abnormal minimizers are an exceptional phenomenon or a common occurrence. We present an analytic technique for proving local optimality of a large class of abnormal extremals that we call "regular." Our technique is based on (a) a "normal form theorem," stating that, locally, a regular abnormal extremal can always be put in a special form by a suitable change of coordinates, and (b) an inequality showing that, once a trajectory is in this special form, then local optimality follows. Using this approach we prove that regular abnormal extremals are locally optimal. If E satisfies a mild additional restriction —valid in particular for all regular 2-dimensional distributions and for generic 2-dimensional distributions— then regular abnormal extremals are "typical" (in a sense made precise in the text), so our result implies that the abnormal minimizers are ubiquitous rather than exceptional. We also discuss some related issues, and in particular show, by means of an example, that a smooth abnormal extremal need not be locally optimal, even if in addition it belongs to the class —recently studied by Bryant and Hsu— of C^1-rigid curves.

Keywords: Sub-Riemannian manifolds, Geodesics, Hamiltonians, Abnormal Extremals.

1 Introduction

The structure of sub-Riemannian minimizers and of the corresponding "geodesics" has recently attracted a great deal of attention (cf. [1], [5], [6], [7], [9], [10], [11], [12], [14], [17], [18], [19], [22], [23], [27], [28]; see Remark 1 below for a discussion of the use of the word "geodesic"), due to the delicate issues that arise because of the possibility of the existence of "abnormal" length-minimizing arcs.

This phenomenon, well known in Optimal Control Theory, was not immediately recognized as possible in the more special case of sub-Riemannian geometry. For example, in 1986 it was stated, in [22], that all length-minimizing arcs for a sub-Riemannian manifold are characteristics of the associated Hamiltonian (i.e. "normal extremals," in our terminology). A proof was suggested for this result, relying on an application of the Pontryagin Maximum Principle from Optimal Control Theory. It turns out, however, that the Maximum Principle only makes it possible to draw the weaker conclusion that every minimizer is either a characteristic of the Hamiltonian (i.e. a normal extremal) or a member of another class of arcs known as "abnormal extremals." The possibility that a minimizer might be an abnormal extremal can easily be ruled out in the Riemannian case and, more generally, for the special class —introduced by R. Strichartz in [23]— of sub-Riemannian metrics defined on "strongly bracket-generating" distributions, but for general sub-Riemannian metrics there is no obvious way to go beyond the necessary conditions of the Maximum Principle and exclude abnormal extremals. This left open the question whether there can exist sub-Riemannian minimizers that are not normal extremals ("strictly abnormal minimizers," in the terminology introduced below). The suggestion that this could indeed happen had in fact been made much

Work supported in part by the National Science Foundation under NSF grant DMS92-02554.
Received by the editor November 1993, and in revised form February 1994.

earlier, in 1977, by B. Gaveau, who had studied a particular sub-Riemannian structure for which he asserted that there were pairs of points p_1, p_2 that could not be joined by a characteristic curve ([9], p. 133, Theorem 1). Since, for the system studied by Gaveau, the existence of a minimizer joining any two points is obvious, his result would have amounted to a non-constructive proof of the existence of a strictly abnormal minimizer. However, contrary to the statement of [9], the points considered by Gaveau *can* be joined by a characteristic curve, as was pointed out by R. Brockett in 1984 (cf. [6]). (Brockett established this by explicitly computing the minimizers and showing that they were characteristics. A brief discussion of the Gaveau-Brockett system is included below in Appendix A, where we provide an alternative proof of Brockett's result, by directly applying the Pontryagin Maximum Principle to show that all minimizers are characteristic curves.) Meanwhile, the claim that all minimizers are normal extremals was made again in 1988 in [1]. Subsequently, relying on this assertion, Hamenstädt stated in 1990, in [12], that "the critical points are geodesics, i.e. locally minimizing curves," and "every geodesic is a critical point; together this gives a complete description of the geodesics." (Hamenstädt's "critical points" are exactly the same as our "normal extremals.") However, the claim of [1] is also incorrect, and the "complete description of the geodesics" announced in [12] is in fact incomplete, because it omits a rather important and interesting class of minimizers that are rather different from the normal extremals.

Correct answers to these questions are in fact suggested in a rather natural way by Optimal Control Theory, which gives necessary conditions for trajectories of fairly general control systems to minimize functionals of a rather general type. The sub-Riemannian minimization problem just happens to be a special case of the much broader class of situations where the Maximum Principle of Optimal Control Theory applies. (This fact was recognized by Strichartz in [22], even though his result failed to include the abnormal extremals.) When the Maximum Principle is in fact applied to this case, the possibility that some minimizers might be "abnormal" emerges directly by mechanically writing the necessary conditions. This, of course, does not yet answer the question whether abnormal minimizers actually exist, but it provides a radically different perspective on the problem: whereas a differential geometer's natural inclination is to look for the true characterization of minimizers in the form of a modified version of the geodesic equation, based on making variations as in the classical derivation of the Euler-Lagrange equations, the control theorist's predisposition is to start from the opposite direction, applying the Maximum Principle to conclude that a minimizer is either a "normal extremal" or an "abnormal" one, inferring from this that abnormal extremal minimizers probably exist, and then setting out to prove that they do.

The problem of the existence of strictly abnormal minimizers was finally settled in 1991, when R. Montgomery, in [17], produced an example of such a minimizer, thereby showing that the intuition derived from the control-theoretic point of view was in fact the right one, even though Montgomery himself was led to his discovery by physical rather than control-theoretic considerations. Montgomery's very long and ingenious optimality proof was somewhat simplified in 1992 by I. Kupka (cf. [14]). However, neither of these proofs makes it possible to go beyond individual examples and prove, for instance, that large classes of abnormal extremals are optimal. In [16], we studied an example for which the optimality proof was much simpler. Moreover, this example had the extra virtue of being such that, after suitable changes of coordinates, one can transform fairly general situations into normal forms to which the proof of [16] applies. This was made precise in the preprint [25], where a general "optimality lemma" was announced, which basically describes the most general situation where a method similar to that of [16] can be applied to prove optimality. Using this optimality lemma plus a coordinate transformation, it was proved in [25] that, for a completely arbitrary four-dimensional sub-Riemannian manifold whose metric is defined on a regular two-dimensional distribution, there passes through each point exactly one locally simple abnormal extremal parametrized by arc-length, and all these abnormal extremals are locally optimal. (Here "simple" means "without double points.") It was also shown that, if an extra generic condition is satisfied, then these abnormal extremals are not normal. It then became apparent that abnormal minimizers are not pathological objects that can only be shown to exist in very special and elaborately constructed examples. They are in fact ubiquitous and easy to find (at least for generic cases), and can be proved optimal by relatively simple general techniques.

The purpose of this work is to present the general theory of regular abnormal extremals for manifolds M of arbitrary dimensions equipped with sub-Riemannian structures (E, G) arising from two-dimensional distributions, and to prove in particular that these abnormal extremals are locally optimal. (The fact that *normal* extremals are locally optimal is an immediate consequence of the Control Theory version of classical Hamilton-Jacobi theory, which gives sufficient conditions for optimality of "fields of extremals," as explained, for example, in [15]. An independent derivation was given by Hamenstädt in [12]. Since the proof of [12] is based on a formalism quite different from ours, and the control-theoretic proof is in our view a rather transparent illustration of the power of Optimal Control methods and the advantages of the Hamiltonian formulation, we have included the latter proof in Appendix C.)

A special effort has been made to produce a detailed, self-contained presentation. (We believe this is especially important in view of the fact that so many papers published on this subject contain incorrect statements.) We also insist on using the appropriate geometric language, which in our view is that of Optimal Control Theory, properly formulated in a Hamiltonian setting. In particular, our starting point is the description of Riemannian geodesics as the projections of the bicharacteristics of a Hamiltonian function on the cotangent bundle, rather than the usual characterization in terms of the standard geodesic equation $\nabla_{\dot\gamma}\dot\gamma = 0$ in the tangent bundle. This formulation generalizes naturally to the sub-Riemannian case, and leads to a completely intrinsic definition of "extremal." (This is to be contrasted, e.g., with the "simple differential equation for the critical points" presented in [12], which evolves in the tangent bundle but depends on the choice of a frame complementary to the distribution.) To emphasize the coordinate-free nature of the characterization of extremals, we use a coordinate-free version of the Maximum Principle. Since the distribution E need not have a global basis of sections, we rely in fact on a version that is doubly intrinsic, in that it is not only coordinate-free but also formulated without any reference to choices of sections.

Extremals are, by definition, projections on M of certain arcs in the cotangent bundle T^*M of M called *biextremals*. Biextremals can be of either of two mutually exclusive types, namely, "normal" or "abnormal." An extremal is normal or abnormal depending on whether it is a projection of a normal or an abnormal biextremal. (An extremal can be both normal and abnormal. Abnormal extremals that are not normal will be called "strictly abnormal.") Normal extremals are obviously smooth, but abnormal extremals can fail to be smooth, or even of class C^1 (cf. the example in Subsection 9.6.) The Pontryagin Maximum Principle implies that every arc that minimizes length is an extremal. As pointed out above, normal extremals locally minimize length, and it is then natural to ask whether abnormal extremals —or, perhaps, abnormal extremals that satisfy some extra conditions— also locally minimize. Two natural candidates for the "extra conditions" are smoothness and rigidity. (Rigid curves for rank 2 distributions were studied by R. Bryant and L. Hsu in [7]. The class of curves considered here is almost exactly the same as that of [7]. For a discussion of the precise relationship between the two classes, cf. Remark 2 below.) We show, however, that an abnormal extremal need not be a local minimizer, even if it is both smooth and rigid. (This is done in Appendix D, where we exhibit a C^1-rigid real-analytic abnormal extremal which is not a local minimizer.)

So, abnormal extremals can be local minimizers, as in the examples of Montgomery and Kupka, but need not be. This naturally leads to the main question studied here, namely, which of the two possibilities occurs for "typical" abnormal extremals. Our answer —for distributions E of rank two— is that local minimization is the most common situation. We make this precise by first defining the class of "regular abnormal biextremals." Our main result is then Theorem 5, which says that regular abnormal extremals (i.e. projections of regular abnormal biextremals) locally minimize length. In addition, we show that, when the distribution E satisfies a very mild extra condition, then these regular abnormal biextremals are "typical" in the following precise sense: if we use $AEC(E)$ (the "abnormal extremal carrier of E) to denote the set of nonzero members of the annihilator of $[E, E]$, then: (i) every nonconstant abnormal biextremal is contained in $AEC(E)$, (ii) there is a relatively open dense subset $RA(E)$ of $AEC(E)$ which is a submanifold and carries a one-dimensional foliation F_E whose leaves, properly parametrized, are the regular abnormal biextremals, (iii) every locally simple abnormal biextremal contained in $RA(E)$ and parametrized by arc-length is regular.

Remark 1 Arcs that minimize length are often called "geodesics." In the Riemannian case, the precise definition of a "geodesic" is that it is a curve that satisfies the necessary condition for length-minimization given by the geodesic equation or, equivalently, a curve that locally minimizes length. In the sub-Riemannian setting there is no unanimous agreement in the literature on the proper use of the word "geodesic." There is an obvious sub-Riemannian analogue of the geodesic equation, namely, the equation of non-null bicharacteristics of the Hamiltonian of the sub-Riemannian metric. This suggests the possibility of calling these bicharacteristics —or their projections on the base manifold— "geodesics." This would specialize in the Riemannian case to the usual definition of geodesics, but in the more general sub-Riemannian setting it would no longer be true that every minimizer is a geodesic. Alternatively, we could choose to define a "geodesic" to be an arc that locally minimizes length. (Some authors —e.g. R. Montgomery in [17], whose first version was pointedly entitled "Geodesics that do not satisfy the geodesic equations"— appear to be leaning in this direction, and Hamenstädt in [12] explicitly states that "geodesics" are "locally minimizing curves.") This would also reduce in the Riemannian case to the usual definition of geodesics, and would in addition have the desired property that all minimizers are geodesics. On the other hand, the new definition would have the drawback of making geodesics hard to characterize by means of differential equations. (Such a characterization has not yet been found, and may very well not exist.) *In this work we choose to avoid the word "geodesic" altogether.* Arcs that minimize length are called "minimizers." Arcs that satisfy the necessary conditions for minimization given by the Maximum Principle of Optimal Control Theory are "extremals." Extremality is a necessary condition for an arc to be a minimizer, and has an explicit characterization in terms of constrained ordinary differential equations, but in the sub-Riemannian case an extremal need not be a local minimizer, as explained above. ∎

Remark 2 In [7], Bryant and Hsu studied rigid curves for smooth distributions, and proved that, if a rank 2 distribution E on M satisfies a nonintegrability condition, then through every point p of M there passes a locally rigid curve. Our results are related to those of [7] as follows. The class of points p where our results apply is slightly larger than that of [7], because in [7] it is assumed that the distributions E and $[E,E]$ are nowhere integrable, for which it is necessary that $\dim[E,E](p) = 3$ and $\dim[E,[E,E]](p) > \dim[E,E](p)$ (cf. [7], Theorem 3.1), whereas our results apply at every point p such that $\dim[E,[E,E]](p) > \dim[E,E](p)$, even if $\dim[E,E](p) = 2$. (This includes, in particular, the generic singularities of rank 2 distributions in dimension 3.) On the other hand, for the points considered in [7], our regular abnormal extremals are exactly the "projections of Ψ-characteristics in $Q_1^\bullet$" of [7]. The main result of [7] is that these curves are locally rigid. Our results —specialized to sub-Riemannian manifolds (M,E,G) for which E satisfies the nonintegrability hypotheses of [7]— imply in particular that these curves are also local minimizers. In Appendix D we show that by using essentially the same method as in the proof of our local minimization theorem one can also establish the local rigidity of regular abnormal extremals for rank 2 distributions, thereby providing an alternative proof of the Bryant-Hsu result under our slightly more general conditions. Since the curves studied here and in [7] happen to be both locally rigid and locally minimizing, one might suspect that there is a deeper relationship between rigidity and local optimality, e.g. that perhaps the two properties are equivalent, or that one of them might imply the other. The purpose of Appendix D is to argue that this is not so, by showing that both implications are false in general.* ▮

2 Three examples

Before discussing the general theory, we illustrate the main issues by means of three examples. The first one is the classical theory of Rie-

*Our example of a nonrigid abnormal minimizer is for a rank 3 distribution.

mannian geodesics, reformulated in the Hamiltonian language that generalizes naturally to the sub-Riemannian case. The second one is the simplest possible sub-Riemannian but not Riemannian situation, corresponding to the Heisenberg Lie algebra. In this case, the "natural" extension of the Riemannian theory of geodesics turns out to work, in the sense that the length-minimizing arcs are characteristic curves of the Hamiltonian associated to the metric. Finally, the third example —first discussed in [16]— exhibits a case where the natural extension does not work: we explicitly show that a certain arc is a minimizer but is not a characteristic of the Hamiltonian.

2.1 Riemannian geodesics

If M is a Riemannian manifold, with metric tensor G, then the curves that locally minimize length and are parametrized by arc-length are exactly those that satisfy, in local coordinates, the geodesic equation $\ddot{x}^i + \sum_{jk} \Gamma^i_{jk}\dot{x}^j\dot{x}^k = 0$, where the Γ^i_{jk} are the Christoffel symbols of the Riemannian connection associated to the metric, together with the constraint $\sum_i (\dot{x}^i)^2 = 1$. Alternatively, one can avoid the Christoffel symbols by using the Legendre transformation to pass to the Hamiltonian formulation. This amounts to working on the cotangent bundle T^*M, and introducing a "momentum" covector λ whose components λ_i are given by $\lambda_i = -\sum_j g_{ij}(x)\dot{x}^j$. (This is precisely —up to a sign[†]— the familiar "lowering indices" operation.) We then define the Hamiltonian $H : T^*M \to \mathbb{R}$ by letting $H = -\frac{1}{2}\sum_{ij} g^{ij}(x)\lambda_i\lambda_j$, i.e.

$$H(x,\lambda) = -\frac{1}{2}\|\lambda\|^2_{G,x} \ , \qquad (1)$$

where $\|\lambda\|_{G,x}$ is the norm of the covector λ, regarded as a linear functional on the tangent space T_xM of M at x, endowed with the inner

[†] Our somewhat nonstandard sign conventions are chosen so as to be consistent with those commonly used in Optimal Control Theory.

product $G(x)$. The geodesic equations become $\dot{x} = \frac{\partial H}{\partial \lambda}$, $\dot{\lambda} = -\frac{\partial H}{\partial x}$, i.e. the usual Hamilton equations, whose solutions are the *bicharacteristics* of H. The constraint $\|\dot{x}\|_G = 1$ becomes the equality $\|\lambda\|_G = 1$, i.e. $H = -\frac{1}{2}$. The projections on M of the bicharacteristics are the *characteristics* of H. In other words, *Formula (1) associates to the Riemannian metric G a smooth real-valued function H on the cotangent bundle T^*M. The symplectic structure of T^*M associates to H a Hamiltonian vector field $\overrightarrow{H}$. The projections on M of the integral curves of $\overrightarrow{H}$ along which $H \equiv -\frac{1}{2}$ are the geodesics parametrized by arc-length.* More generally, the geodesics parametrized by a constant times arc-length are precisely the projections on M of arbitrary non-null bicharacteristics of H (i.e. integral curves of $\overrightarrow{H}$ along which the constant value of H is $\neq 0$).

2.2 The Heisenberg algebra case

We now discuss the simplest example of a sub-Riemannian structure which is not Riemannian. Let ω be the 1-form in $\mathbb{R}^3$ given by

$$\omega = dz - (x\,dy - y\,dx) \tag{2}$$

We wish to consider arcs γ in $\mathbb{R}^3$ that satisfy the velocity constraint $\langle \omega, \dot{\gamma} \rangle = 0$. Precisely, we let $\mathcal{C}_{a,b}$ be the set of all absolutely continuous curves $\gamma : [a,b] \to \mathbb{R}^3$ that satisfy $\langle \omega(\gamma(t)), \dot{\gamma}(t) \rangle = 0$ for almost all $t \in [a,b]$, and define $\mathcal{C} = \cup_{-\infty<a\leq b<+\infty} \mathcal{C}_{a,b}$. Equivalently, if we let $E(q) = \ker \omega(q)$ for $q \in \mathbb{R}^3$, then $E(q)$ is a two-dimensional subspace of $\mathbb{R}^3$ for each q (so E is a "two-dimensional distribution" in $\mathbb{R}^3$), and the curves γ in $\mathcal{C}$ are those that satisfy $\dot{\gamma}(t) \in E(\gamma(t))$ for almost all t.

It is not hard to see that, in spite of the constraint $\dot{\gamma}(t) \in E(\gamma(t))$, every pair q_1, q_2 of points of $\mathbb{R}^3$ can be joined by a curve in $\mathcal{C}$. (For example, the two vector fields $f = \frac{\partial}{\partial x} - y\frac{\partial}{\partial z}$ and $g = \frac{\partial}{\partial y} + x\frac{\partial}{\partial z}$ form a basis of sections of E. The Lie bracket $[f,g]$ is equal to $2\frac{\partial}{\partial z}$, so f, g and $[f,g]$ are linearly independent at each point. It then follows from

Chow's Theorem, cf. [8], that, given any two points q_1, q_2, there is a finite concatenation of integral curves of f and g that goes from q_1 to q_2.)

We define a Riemannian metric G on E by declaring (f,g) to be an orthonormal basis of sections of E. We now look for the curves in $\mathcal{C}$ that minimize length, i.e. the curves $\gamma \in \mathcal{C}$ such that, if a,b are such that $\gamma \in \mathcal{C}_{a,b}$, then $\int_a^b \|\dot\gamma(t)\|_G dt \le \int_\alpha^\beta \|\dot\gamma^*(t)\|_G dt$ for every (α,β,γ^*) such that $\gamma^* \in \mathcal{C}_{\alpha,\beta}$, $\gamma^*(\alpha) = \gamma(a)$ and $\gamma^*(\beta) = \gamma(b)$. (Here, if $v \in E(q)$, we write $\|v\|_G = \sqrt{G(q)(v,v)}$.)

If we proceed by analogy with the case of the Riemannian geodesics, it is natural to introduce the Hamiltonian H, defined on $\mathbb{R}^6$ (i.e. on $T^*\mathbb{R}^3$) by

$$H(q,\lambda) = -\frac{1}{2}\|\lambda\|^2_{G,q} \ , \tag{3}$$

where $\|\lambda\|_{G,q}$ denotes the norm of the linear functional $E(q) \ni v \to \langle\lambda, v\rangle \in \mathbb{R}$ with respect to the inner product $G(q)$, that is,

$$\|\lambda\|_{G,q} = \sup\{\langle\lambda,v\rangle \ : \ v \in E(q)\,,\ \|v\|_G \le 1\}\ . \tag{4}$$

Since $f(q)$ and $g(q)$ form an orthonormal basis of $E(q)$, we have

$$H(q,\lambda) = -\frac{1}{2}\Big(\langle\lambda, f(q)\rangle^2 + \langle\lambda, g(q)\rangle^2\Big)\ . \tag{5}$$

Writing $q = (x,y,z)$, $\lambda = (\xi,\eta,\zeta)$, we find:

$$H(x,y,z,\xi,\eta,\zeta) = -\frac{1}{2}\Big((\xi - y\zeta)^2 + (\eta + x\zeta)^2\Big)\ . \tag{6}$$

The bicharacteristic equations turn out to be

$$\begin{aligned} \dot x &= y\zeta - \xi\ ,\ \dot y = -\eta - x\zeta\ ,\ \dot z = -x(\eta + x\zeta) + y(\xi - y\zeta)\ , \\ \dot\xi &= (\eta + x\zeta)\zeta\ ,\ \dot\eta = (y\zeta - \xi)\zeta\ ,\ \dot\zeta = 0\ . \end{aligned} \tag{7}$$

Notice that the third equation simply says that $\dot z = x\dot y - y\dot x$, i.e. that our curve satisfies $\omega = 0$. The sixth equation says that ζ is a constant, and the fourth and fifth equation say that $\dot{\xi} = -\zeta\dot{y}$ and $\dot{\eta} = \zeta\dot{x}$. If the constant value of ζ is 0, then ξ and η are constant, and so the curve $t \to (x(t), y(t))$ is a straight line, along which the motion takes place with constant velocity. The function $t \to z(t)$ is then computed by solving $\dot{z} = -x\eta + y\xi$, and the resulting curve $t \to (x(t), y(t), z(t))$ is a straight line in $\mathbb{R}^3$. If $\zeta \neq 0$, then the equations (7) imply $\ddot{x} - 2\dot{y}\zeta = \ddot{y} + 2\dot{x}\zeta = 0$. So $\dot{x}(t) = A\cos 2\zeta t + B\sin 2\zeta t$ and $\dot{y}(t) = -A\sin 2\zeta t + B\cos 2\zeta t$, where A and B are constants. The condition that our curve is parametrized by arc-length then says that $A^2 + B^2 = 1$, so we can write $A = \sin\theta$, $B = \cos\theta$. So we have defined, for each initial condition $q_0 = (x_0, y_0, z_0)$, a two-parameter family of curves —depending on θ and ζ— whose projections on the (x, y)-plane are given by $x = x_0 + \frac{\cos\theta - \cos(2\zeta t + \theta)}{2\zeta}$, $y = y_0 + \frac{\sin(2\zeta t + \theta) - \sin\theta}{2\zeta}$. These projections are therefore circles, along which motion occurs with constant angular velocity. The Pontryagin Maximum Principle, applied to this case, implies that all the minimizers are curves of the kind just described, i.e. characteristic curves of the Hamiltonian H. So in this example it is also true, as in the Riemannian case, that *Formula (3) associates to the sub-Riemannian metric (E, G) a smooth real-valued function H on the cotangent bundle T^*M, and every length-minimizing arc parametrized by arc-length is a characteristic curve of H.*

2.3 An abnormal minimizer

We now give an example where there exist minimizers that are not characteristic curves of the associated Hamiltonian. We let $M = \mathbb{R}^3$, with the usual coordinates x, y, z, and take E to be the kernel of the 1-form $\omega = x^2 dy - (1 - x)dz$. Since ω never vanishes, it is clear that E is a smooth two-dimensional subbundle of the tangent bundle TM. The vector fields $f = \frac{\partial}{\partial x}$, $g = (1 - x)\frac{\partial}{\partial y} + x^2\frac{\partial}{\partial z}$, form a global basis of sections of E, and satisfy

$$[f,g] = -\frac{\partial}{\partial y} + 2x\frac{\partial}{\partial z}\,, \qquad [f,[f,g]] = 2\frac{\partial}{\partial z}\,,$$

$$[g,[f,g]] = [f,[f,[f,g]]] = [g,[f,[f,g]]] = 0\ .$$

In particular, f, g and $[f,g]$ are linearly independent everywhere except where $x = 0$ or $x = 2$, and f, g and $[f,[f,g]]$ are independent everywhere except where $x = 1$. It follows that for every $q \in \mathbb{R}^3$ the values at q of the four vector fields f, g, $[f,g]$ and $[f,[f,g]]$ span $\mathbb{R}^3$. Once again, Chow's Theorem implies that, given any two points q_1, q_2, there is a finite concatenation of integral curves of f and g that goes from q_1 to q_2. We define a Riemannian metric G on the bundle E by $G = dx^2 + \psi(x)(dy^2 + dz^2)$, where $\psi(x) = ((1-x)^2 + x^4)^{-1}$. Then f and g form an orthonormal basis of sections of E. We define the classes $\mathcal{C}_{a,b}$ and $\mathcal{C}$ exactly as in the previous section, but with the new ω and E. As before, we seek to characterize the length-minimizers for this metric and constraint.

Pursuing the approach of the previous examples, we define the Hamiltonian $H : T^*M \to \mathbb{R}$ by $H(q,\lambda) = -\frac{1}{2}\|\lambda\|^2_{G,q}$, where $\|\lambda\|_{G,q}$ denotes the norm of the linear functional $E(q) \ni v \to \langle\lambda, v\rangle \in \mathbb{R}$ with respect to the inner product $G(q)$ on $E(q)$. Since $f(q)$ and $g(q)$ form an orthonormal basis of $E(q)$, we have $H(q,\lambda) = -\frac{1}{2}\Big(\langle\lambda, f(q)\rangle^2 + \langle\lambda, g(q)\rangle^2\Big)$. Writing $q = (x,y,z)$, $\lambda = (\xi,\eta,\zeta)$, we find:

$$H(x,y,z,\xi,\eta,\zeta) = -\frac{1}{2}\Big(\xi^2 + (\eta(1-x) + \zeta x^2)^2\Big)\ . \tag{8}$$

If we let $\varphi = \eta(1-x) + \zeta x^2$, then the bicharacteristic equations turn out to be

$$\dot{x} = -\xi\ ,\ \ \dot{y} = (x-1)\varphi\ ,\ \ \dot{z} = -x^2\varphi\ ,\ \ \dot{\xi} = (2x\zeta - \eta)\varphi\ ,\ \ \dot{\eta} = 0\ ,\ \ \dot{\zeta} = 0\ , \tag{9}$$

We now fix a and b, and consider the curve $\gamma_* : [a,b] \to \mathbb{R}^3$ in $\mathcal{C}$ given by $\gamma_*(t) = (0,t,0)$. We claim that γ_* *is not a characteristic curve* *of H for any choice of a, b, but γ_* is a minimizer if $b-a$ is small enough.* To see that γ_* cannot be a characteristic, assume that it is, pick a bicharacteristic $\Gamma_* : [a,b] \to \mathbb{R}^6$ of H that projects onto γ_*, and write $\Gamma_*(t) = (x_*(t), y_*(t), z_*(t), \xi_*(t), \eta_*(t), \zeta_*(t))$. For Γ_*, the fourth equation of (9) reduces to $\dot{\xi}_* = -\eta_*^2$, since $x_*(t) \equiv 0$. The first equation of (9) then implies that $\xi_*(t) \equiv 0$, so the constant η_* vanishes. But the second equation says that $\dot{y}_* = -\eta_*$, and y_* is not constant along γ_*, so we have reached a contradiction.

The proof that γ_* minimizes length is based on the following lemma (cf. [16]).

Lemma 1 *Let $0 < \tau \le \frac{2}{3}$, and let u, $v : [0,\tau] \to \mathbb{R}$ be two measurable functions such that $|u(t)| \le 1$ and $|v(t)| \le 1$ for almost all t. Define $x : [0,\tau] \to \mathbb{R}$ by $x(t) = \int_0^t u(s)\,ds$. Assume that $x(\tau) = 0$ and $\int_0^\tau x(t)^2 v(t)dt = 0$. Then $\int_0^\tau \big(1 - x(t)\big)v(t)dt \le \tau$, and equality holds if and only if $u(t) = 0$ and $v(t) = 1$ for almost all $t \in [0,\tau]$.*

PROOF. Write $A = \int_0^\tau \big(1-x(t)\big)v(t)dt$. Let $V(t) = \int_0^t v(s)ds$. Then $-\tau \le V(\tau) \le \tau$. Let $\alpha = \tau - V(\tau)$, $\beta = \sup\{|x(t)| : t \in [0,\tau]\}$. Since $|v(t)| \le 1$, we have $\left|\int_0^\tau x(s)v(s)ds\right| \le \beta\tau$, so that

$$\begin{aligned} A &\le \int_0^\tau v(s)ds + \Big|\int_0^\tau x(s)v(s)ds\Big| \\ &\le V(\tau) + \beta\tau = V(\tau) - \tau + \tau + \beta\tau = \tau - \alpha + \beta\tau. \end{aligned}$$

Our conclusion will follow if we show that $\beta\tau \le \alpha$. If we prove that $\beta \le \frac{3}{2}\alpha$, then $\beta\tau \le \frac{3}{2}\alpha\tau$, which implies $\beta\tau \le \alpha$, since $\tau \le \frac{2}{3}$. To prove that $\beta \le \frac{3}{2}\alpha$, we show that

$$\frac{2\beta^3}{3} \le \int_0^\tau x(t)^2 dt \le \beta^2\alpha \ . \tag{10}$$

First, using the equality $\int_0^\tau x(t)^2 v(t)dt = 0$, we have

$$\begin{aligned} \int_0^\tau x(t)^2 dt &= \int_0^\tau x(t)^2\big(1-v(t)\big)dt \le \beta^2 \int_0^\tau \big(1-v(t)\big)dt \\ &= \beta^2(\tau - V(\tau)) = \beta^2\alpha \ , \end{aligned} \tag{11}$$

establishing the upper bound for $\int_0^\tau x(t)^2 dt$. To prove the lower bound, we pick a point $a \in [0,\tau]$ such that $|x(a)| = \beta$. Then clearly $a \geq \beta$ and $\tau - a \geq \beta$, since $x(0) = 0$, $x(\tau) = 0$, and $|\dot{x}(t)| \leq 1$. So the intervals $I_1 = [a-\beta, a]$, $I_2 = [a, a+\beta]$ are entirely contained in $[0,\tau]$. On each of the two intervals I_j, the function $|x|$ is bounded below by the linear function φ_j that takes the value β at a and is 0 at the other endpoint. It is clear that $\int_{I_j} \varphi_j^2(t)dt = \frac{\beta^3}{3}$. So $\int_{I_j} x(t)^2 dt \geq \frac{\beta^3}{3}$. Therefore $\int_0^\tau x(t)^2 dt \geq \int_{I_1} x(t)^2 dt + \int_{I_2} x(t)^2 dt \geq \frac{2\beta^3}{3}$, completing the proof of (10). Now from the above proof we see that $\int_0^\tau \Big(1 - x(t)\Big)v(t)dt = \tau$ would imply that $\beta = \frac{3}{2}\alpha$. This is true only when (11) is an equality, which is only possible if $x(t) \equiv \beta$. Since $x(0) = 0$, we see that $x \equiv 0$. Therefore $u = 0, v = 1$ for almost all $t \in [0,\tau]$. This concludes the proof. ∎

Proposition 1 *If $b - a \leq \frac{2}{3}$, then γ_* is a minimizer, and all the other minimizers going from $(0,a,0)$ to $(0,b,0)$ are obtained from γ_* by a reparametrization of the time interval.*

PROOF. It is clear that γ_* is parametrized by arc-length and has length $T = b - a$. Let $[\tilde{a}, \tilde{b}] \ni t \to (x(t), y(t), z(t)) = \gamma(t)$ be any curve in $\mathcal{C}$ that also goes from $(0,a,0)$ to $(0,b,0)$, and let τ be its length. We want to show that $\tau \geq T$. We can clearly assume that $\tilde{a} = 0$. The vector $\dot{\gamma}(t)$ can be written in a unique way as a linear combination $u(t)f(\gamma(t)) + v(t)g(\gamma(t))$, and the coefficients $u(t)$, $v(t)$ are measurable functions of t (since $u = \dot{x}$ and $v = \frac{(1-x)\dot{y} + x^2\dot{z}}{(1-x)^2 + x^4}$). Since f and g form an orthonormal basis of sections of E, we have $\tau = \int_0^{\tilde{b}} \sqrt{u(t)^2 + v(t)^2} dt$. Assume first that γ is parametrized by arc-length, i.e. that $u(t)^2 + v(t)^2 \equiv 1$. Then $\tau = \tilde{b}$. Therefore the components $x(t)$, $y(t)$, $z(t)$ of $\gamma(t)$ satisfy $\dot{x}(t) = u(t)$, $\dot{y}(t) = (1 - x(t))v(t)$, $\dot{z}(t) = x(t)^2 v(t)$. The functions u, v, x satisfy the hypotheses of the lemma. So, if $\tau \leq \frac{2}{3}$, we have $T = y(\tau) - y(0) = \int_0^\tau \Big(1 - x(t)\Big)v(t)dt \leq \tau$. If $T \leq \frac{2}{3}$, then either $\tau \leq \frac{2}{3}$, in which case we have shown that $T \leq \tau$, or $\tau \geq \frac{2}{3}$, in which case $T \leq \tau$ as well. Moreover, equality can only hold if $u(t) = 0$ and

$v(t) = 1$ for almost all $t \in [0, \tau]$, in which case $\gamma(t) = (0, a+t, 0)$, so $\gamma(t) \equiv \gamma_*(t+a)$.

If γ is not parametrized by arc-length, then it is easily shown that γ can be reparametrized by arc-length, i.e. there is an absolutely continuous curve $\delta : [0, \tau] \to \mathbb{R}^3$ which is parametrized by arc-length, and satisfies (i) $\dot{\delta}(t) \in E(\gamma(t))$ for almost all t, and (ii) $\gamma(t) = \delta(\sigma(t))$ for a monotonically non-decreasing absolutely continuous surjective function $\sigma : [0, \tilde{b}] \to [0, \tau]$. The first part of the proof then applies to δ, so $T \leq \tau$ if $T \leq \frac{2}{3}$. Moreover, equality can only happen if $\delta = \gamma_*$ modulo time reparametrization, in which case γ and γ_* coincide modulo time reparametrization. ▮

3 Notational conventions and definitions

The purpose of this section is to state precisely our general notational conventions regarding manifolds, charts, curves, bundles, smooth distributions, Hamilton vector fields, and other objects of interest.

3.1 Manifolds, charts, bundles, curves and arcs

Throughout this work, the word "smooth" always means "of class C^∞." "Manifold" means "finite-dimensional, Hausdorff, smooth, second-countable manifold without boundary." "Bundle" and "subbundle" always mean "smooth vector bundle" and "smooth subbundle," respectively. The *dimension*, or *rank* of a bundle is its fiber dimension.

If M is a manifold and $k \in \{0, 1, , \ldots\} \cup \{\infty\}$, we use $C^k(M)$ to denote the set of all real-valued functions on M that are of class C^k, and $V^k(M)$ to denote the set of all vector fields of class C^k on M. If E is a bundle over M, we use $C^k(E)$ to denote the set of all sections of class C^k of E. For $p \in M$, we use T_pM, T_p^*M to denote, respectively, the tangent and cotangent spaces of M at p, and we write $T_p^\# M \stackrel{\text{def}}{=} T_p^*M - \{0\}$. We use TM, T^*M, π_M, π_M^* to denote, respectively, the tangent and cotangent bundles of M and the canonical projections from TM and T^*M onto M, and write $T^\# M$, $\pi_M^\#$ for the cotangent bundle with the zero section removed and the corresponding projection onto M. If $\lambda \in T_p^*M$, $v \in T_pM$, then $\lambda(v)$, $\langle \lambda, v\rangle$ or, simply, λv, will denote the value at v of the linear functional λ. A *chart* on an n-dimensional manifold M is a smooth diffeomorphism $\kappa : U \to \mathbb{R}^n$ from an open subset of M onto an open subset of $\mathbb{R}^n$. If $\kappa(U)$ is the cube $C^n(\rho) \stackrel{\text{def}}{=} \{(x_1, \ldots, x_n) : |x_i| < \rho \text{ for } i = 1, \ldots, n\}$ for some $\rho > 0$, then κ is called a *cubic chart of radius* ρ. If $p \in U$ and $\kappa(p) = (0, \ldots, 0)$, then we will say that κ is *centered at* p. For each chart $\kappa = (x^1, \ldots, x^n) : U \to \mathbb{R}^n$ on M there are induced charts $T\kappa : \pi_M^{-1}(U) \to \mathbb{R}^{2n}$, $T^*\kappa : (\pi_M^*)^{-1}(U) \to \mathbb{R}^{2n}$ of TM, T^*M, given by

$$
\begin{aligned}
T\kappa(x, v) &= \Big(\kappa(x), (\langle dx^1, v\rangle, \ldots, \langle dx^n, v\rangle)\Big) , \\
T^*\kappa(x, \lambda) &= \Big(\kappa(x), (\langle \lambda, \partial_1\rangle, \ldots, \langle \lambda, \partial_n\rangle)\Big) ,
\end{aligned}
$$

where $\partial_j \stackrel{\text{def}}{=} \frac{\partial}{\partial x^j}$.

A *curve* in a manifold M is a continuous map $\gamma : I \to M$ defined on a subinterval I of the real line. An *arc* in M is an absolutely continuous curve whose domain is a compact interval. ("Absolute continuity" is of course well defined in terms of local charts.) The *boundary* $\partial\gamma$ of an arc $\gamma : [a, b] \to M$ is the ordered pair $\partial\gamma = (\gamma(a), \gamma(b))$.

3.2 Hamiltonian functions, Hamilton vector fields, and bicharacteristics

If N is a symplectic manifold with symplectic 2-form Ω, and $H \in C^1(N)$, we use $\overrightarrow{H}$ to denote the *Hamilton vector field* associated to H. (By definition, $\overrightarrow{H}$ is the vector field V on N such that $\Omega(X, V) = \langle dH, X\rangle$ for every vector field X on N.) If $H \in C^k(N)$ and $k \geq 1$, then the vector field $\overrightarrow{H}$ is of class C^{k-1}. If $H, K \in C^1(N)$, then the *Poisson bracket* $\{H, K\}$ is the directional derivative of K in the direction of $\overrightarrow{H}$, i.e. $\{H, K\} = \langle dK, \overrightarrow{H}\rangle = \Omega(\overrightarrow{H}, \overrightarrow{K})$. The well known formulas $\{H, KL\} = \{H, K\}L + \{H, L\}K$, $\{H, \{K, L\}\} + \{K, \{L, H\}\} + \{L, \{H, K\}\} = 0$ (the Jacobi identity) and $\overrightarrow{HK} = H\overrightarrow{K} + K\overrightarrow{H}$ will be used throughout, as will the fact that the map $H \to \overrightarrow{H}$ is a Lie algebra homomorphism from $(C^\infty(N), \{\cdot,\cdot\})$ to $(V^\infty(N), [\cdot,\cdot])$.

If $\mathcal{H}$ is any subset of $C^1(N)$, we define, for each $x \in N$, a subset $\overrightarrow{\mathcal{H}}(x)$ of the tangent space T_xN to N at x by letting $\overrightarrow{\mathcal{H}}(x) = \{\overrightarrow{H}(x) : H \in \mathcal{H}\}$. We then define a *bicharacteristic* of $\mathcal{H}$ to be an absolutely continuous curve Γ on N such that $\dot{\Gamma}(t) \in \overrightarrow{\mathcal{H}}(\Gamma(t))$ for almost every t. Finally, define a *null bicharacteristic* of $\mathcal{H}$ to be a bicharacteristic along which all the functions $H \in \mathcal{H}$ vanish. In the special case when $\mathcal{H}$ consists of a single function H, the bicharacteristics (resp. null bicharacteristics) of $\mathcal{H}$ are called *bicharacteristics* (resp. *null bicharacteristics*) of H. Clearly, the bicharacteristics of a Hamiltonian $H \in C^k(N)$ are curves of class C^k along which H is constant. In particular, the bicharacteristics of a smooth Hamiltonian are smooth. If $\mathcal{H} \subseteq C^1(N)$ consists of more than one element, then in general the bicharacteristics of $\mathcal{H}$ need not be of class C^1, even if $\mathcal{H} \subseteq C^\infty(N)$.

If $\mathcal{H}_1$, $\mathcal{H}_2$ are linear spaces (over $\mathbb{R}$) of smooth functions on N, we call $\mathcal{H}_1$ and $\mathcal{H}_2$ *equivalent* if the $C^\infty(N)$-modules generated by $\mathcal{H}_1$ and $\mathcal{H}_2$ coincide, i.e. if every element H of $\mathcal{H}_2$ is a finite linear combination $f_1H_1 + \ldots + f_kH_k$ of elements $H_i \in \mathcal{H}_1$ with smooth coefficients f_i, and viceversa. The following trivial observation will be used repeatedly:

Proposition 2 *If $\mathcal{H}_1$, $\mathcal{H}_2$ are equivalent linear spaces of smooth functions on a symplectic manifold N, then $\mathcal{H}_1$ and $\mathcal{H}_2$ have the same null bicharacteristics.* ∎

3.3 Hamiltonian lifts and characteristics

The cotangent bundle T^*M of a manifold M has a natural symplectic structure determined by the 2-form $\Omega_M = d\omega_M$, where ω_M is the 1-

form given $\omega_M(x,\lambda)(v) = \langle \lambda, d\pi_M^*(v)\rangle$ for $v \in T_{(x,\lambda)}(T^*M)$. Relative to a coordinate chart $T^*\kappa = (x^1,\ldots,x^n,\lambda_1,\ldots,\lambda_n)$ induced by a chart $\kappa = (x^1,\ldots,x^n)$ on M, we have the formulas $\omega_M = \sum_j \lambda_j dx^j$, $\Omega_M = \sum_j d\lambda_j \wedge dx^j$, $\overrightarrow{H} = \sum_j \left(\frac{\partial H}{\partial \lambda_j}\frac{\partial}{\partial x^j} - \frac{\partial H}{\partial x^j}\frac{\partial}{\partial \lambda_j}\right)$, and $\{H,K\} = \sum_j \left(\frac{\partial H}{\partial \lambda_j}\frac{\partial K}{\partial x^j} - \frac{\partial H}{\partial x^j}\frac{\partial K}{\partial \lambda_j}\right)$.

If M is a smooth manifold and $\mathcal{H} \subseteq C^1(T^*M)$, then a *characteristic* (resp. *null characteristic*) of $\mathcal{H}$ is a curve γ on M which is of the form $\pi_M^* \circ \Gamma$ for some bicharacteristic (resp. null bicharacteristic) Γ of $\mathcal{H}$. If Γ can be chosen to be entirely contained in $T^\# M$, then γ will be called a *nontrivial* characteristic (resp. nontrivial null characteristic).

To each vector field X on M we associate the function $H_X : T^*M \to \mathbb{R}$ given by $H_X(q,\lambda) = \langle \lambda, X(q)\rangle$ for $\lambda \in T_q^*M$. Then H_X is of class C^k if and only if X is. Moreover, $d\pi_M^*(\overrightarrow{H_X}(x,\lambda)) = X(x)$ for all $(x,\lambda) \in T^*M$. In particular *the characteristics of H_X are exactly the integral curves of X*. The identity $\{H_X, H_Y\} = H_{[X,Y]}$ holds for $X, Y \in V^1(M)$, and therefore the map $X \to H_X$ is a Lie algebra homomorphism from $(V^\infty(M), [\cdot,\cdot])$ to $(C^\infty(N), \{\cdot,\cdot\})$.

If $X \in V^1(M)$, then the vector field $\overrightarrow{H_X}$ is the *Hamiltonian lift* of X.

3.4 Distributions, admissible curves, orbits

Let M be a smooth manifold. A mapping E that assigns to every $x \in M$ a linear subspace $E(x)$ of TM (resp. of T^*M) will be called a *singular distribution* (resp. *codistribution*) on M. If E is actually a smooth subbundle of TM (resp. of T^*M) then E will be called a *smooth distribution* (resp. *codistribution*) on M. If E is a singular distribution (resp. codistribution), we use $E^\perp$ to denote the annihilator of E, so $E^\perp$ is a singular codistribution (resp. distribution), and $E^\perp$ is smooth if and only if E is. Clearly, $(E^\perp)^\perp = E$.

If E is a smooth distribution on M, an *E-admissible curve* (or, simply, an E*-curve*) is a locally absolutely continuous curve γ in M such that $\dot\gamma(t) \in E(\gamma(t))$ for almost all t in the domain of γ. If γ is an arc —i.e. if the domain of γ is compact— then γ is an E*-admissible arc* or, simply, an E*-arc*.

Remark 3 Admissible curves are sometimes called "horizontal curves" in the literature, because one of the most common ways of constructing interesting smooth distributions E is by letting E be the family of horizontal subspaces of a connection. ∎

If $p, q \in M$, we write $p \sim_E q$ (and say that q *is* E*-reachable from* p) if there is an E-arc γ such that $\partial\gamma = (p, q)$. We say that M is E*-connected* if $p \sim_E q$ for all $p, q \in M$. It is clear that $\sim_E$ is an equivalence relation on M. The equivalence classes modulo $\sim_E$ are the E*-orbits*. It can be proved (see, e.g., [24])) that every E-orbit S has a unique topology and differentiable structure with respect to which S is a smooth connected submanifold of M such that $E(x) \subseteq T_xM$ for all $x \in S$. Then the restriction $E\lceil S$ of E to S is a smooth distribution on S, and S is $E\lceil S$-connected.

3.5 Nonholonomic distributions; regularity

If E is a smooth distribution on M, and $U \subseteq M$ is open, we use $\mathrm{Sec}(E, U)$ to denote the set of all smooth sections of E defined on U, and write $\mathrm{Sec}(E) = \cup_U \mathrm{Sec}(E, U)$. For each positive integer k we let $\mathrm{Sec}^k(E, U)$ denote the set of all vector fields X on U such that X is a linear combination with smooth coefficients of iterated brackets of degree $\leq k$ of members of $\mathrm{Sec}(E, U)$, and write $\mathrm{Sec}^k(E) = \cup_U \mathrm{Sec}^k(E, U)$. For $p \in M$, we let $E^k(p)$ denote the set $\{X(p) : X \in \mathrm{Sec}^k(E)\}$. We write $\mu_E^k(p) = \dim E^k(p)$ for $k \geq 1$, and $\mu_E^0(p) = 0$, and let $\nu_E^k(p) = \mu_E^k(p) - \mu_E^{k-1}(p)$ for $k \geq 1$. We let $E^\infty(p) = \cup_{k=1}^\infty E^k(p)$. The first positive integer k such that $\mu_E^k(p) = \dim E^\infty(p)$ is known as the *step number* of E at p, and we write $k = \mathrm{Step}(E, p)$. The sequence of integers $(\nu_E^1(p), \nu_E^2(p), \ldots, \nu_E^k(p))$ is the *type* of E at p. Any finite sequence $(\nu_E^1(p), \nu_E^2(p), \ldots, \nu_E^j(p))$ with $1 \leq j \leq k$ will be said to be a *partial type* of E at p. (For example, E is of partial type $(2,1,1,1)$ at p iff the spaces $E(p)$, $E^2(p)$, $E^3(p)$ and $E^4(p)$ are of dimensions 2, 3, 4 and 5, respectively, and E is of type $(2,1,1,1)$ at p if in addition $E^4(p) = E^\infty(p)$.)

Definition 1 Let k be a positive integer. A smooth distribution E is called *k-regular* if for every integer j such that $1 \leq j \leq k$ the number $\mu_E^k(p)$ is independent of p. If E is k-regular for all k, then we call E *regular*. For a k-regular distribution E and a $j \in \{1, \ldots, k\}$ we use μ_E^j (resp. ν_E^j) to denote the common value of the numbers $\mu_E^j(p)$ (resp. $\nu_E^j(p)$) and we say that E is *of partial type* $(\nu_E^1, \ldots, \nu_E^k)$. If E is regular, we write $\mathrm{Step}(E)$ for the common value of the numbers $\mathrm{Step}(E, p)$ and, if $k = \mathrm{Step}(E)$, we say that E is *of type* $(\nu_E^1, \ldots, \nu_E^k)$. ∎

Definition 2 A smooth distribution E on a manifold M is *nonholonomic*, or *bracket-generating*, if $E^\infty(p) = T_pM$ for all $p \in M$. If E is both regular and nonholonomic, then the number $\mathrm{Step}(E)$ is known as the *degree of nonholonomy* of E. If k is a positive integer, a smooth distribution E such that $E^k(p) = T_pM$ for all $p \in M$ will be called *k-generating.* ∎

Definition 3 A smooth distribution E on M is said to be *strongly bracket-generating* (SBG) at a point $p \in M$ if $E(p) + [X, \mathrm{Sec}(E)](p) = T_pM$ for every $X \in \mathrm{Sec}(E)$ such that $X(p)$ is defined and $\neq 0$.

Remark 4 The most obvious examples of SBG distributions E are those that are 1-generating, i.e. such that $E = TM$. SBG distributions other than these are rather exceptional. For example, if $m < n$ and $n = \dim M$, then a necessary condition for a regular m-dimensional

E on M to be SBG is that m be even, E be 2-generating, and $n < 2m$. (Examples of other, even more restrictive, necessary conditions are given in [7].)

The most commonly encountered situation where $E \neq TM$ but E is SBG is the *contact* case, i.e. that of a regular bracket-generating distribution of type $(2, 1)$. ∎

4 Abnormal extremals

To every smooth distribution E on a smooth manifold M we associate a set $\mathcal{H}_E$ of Hamiltonian functions $T^*M \to \mathbb{R}$ by letting $\mathcal{H}_E$ be the set of all functions H_X, as X ranges over all smooth sections of E.

Definition 4 Let M be a smooth manifold, and let E be a smooth distribution on M. An *abnormal biextremal* (resp. *abnormal extremal*) of E is a locally absolutely continuous curve in $T^\# M$ (resp. in M) which is a null bicharacteristic (resp. a nontrivial null characteristic) of $\mathcal{H}_E$. ∎

With the above definition, the abnormal extremals of E are exactly the projections on M of the abnormal biextremals. In view of the identity $d\pi_M^*(\overrightarrow{H_X}(x, \lambda)) = X(x)$, it is clear that every abnormal extremal of E is an E-curve.

It follows from Definition 4 that every abnormal biextremal is contained in the set of common zeros of the family of functions $\mathcal{H}_E$, which is precisely the annihilator $E^\perp$ of E. The local structure of $E^\perp$ is described by the following trivial result, stated here for future use.

Proposition 3 *Let E be an m-dimensional smooth distribution in an n-dimensional manifold M. Then $E^\perp$ is a smooth $(2n-m)$-dimensional subbundle of the cotangent bundle T^*M. If V is a set of the form $V = (\pi_M^*)^{-1}(U)$, where U is the domain of a basis $f_1, \ldots, f_m$ of sections of E, then $E^\perp \cap V$ is the subset of V defined by the equations $H_1 = \ldots =$*

$H_m = 0$, where $H_i = H_{f_i}$, i.e. $H_i(x, \lambda) = \langle \lambda, f_i(x) \rangle$. The differentials $dH_i(\xi)$ are linearly independent at every point $\xi \in V$. Moreover, if $\xi \in E^\perp \cap V$, then the vectors $\overrightarrow{H_i}(\xi)$ form a basis of $\overrightarrow{\mathcal{H}_E}(\xi)$, and $d\pi_M^*(\xi)$ maps $\overrightarrow{\mathcal{H}_E}(\xi)$ isomorphically onto $E(\pi_M^*(\xi))$. ∎

Remark 5 *If $\xi \notin E^\perp$, then $\overrightarrow{\mathcal{H}_E}(\xi)$ has dimension $n+m$, and is in fact equal to the set of all $v \in T_\xi(T^*M)$ such that $d\pi_M^*(\xi)(v) \in E(\pi_M^*(\xi))$.* Indeed, write $x = \pi_M^*(\xi)$, and let v be such that $d\pi_M^*(\xi)(v) \in E(x)$. Let f, g be smooth sections of E such that $d\pi_M^*(\xi)(v) = f(x)$ and $a = H_g(\xi) \neq 0$. Let $\tilde{v} = \overrightarrow{H_f}(\xi)$. Then the vector $v - \tilde{v}$ is tangent to the fiber T_x^*M, so it can be identified with a covector $w \in T_x^*M$. Let $z = -\frac{w}{a}$. Define $X = f + \psi g$, where ψ is a smooth function on M that vanishes at x and is such that $d\psi(x) = z$. Then $\overrightarrow{H_X}(\xi) = \tilde{v} + H_g(\xi)\overrightarrow{\psi}(\xi)$. It is easy to see that the vector $\overrightarrow{\psi}(\xi)$ is tangent to T_x^*M, and equal to $-z$. But then $\overrightarrow{H_X}(\xi) = v$. ∎

Remark 6 Proposition 3 and Remark 5 say that $\overrightarrow{\mathcal{H}_E}$ is a singular distribution on T^*M, whose dimension is equal to m on $E^\perp$ and to $n+m$ on $T^*M - E^\perp$. (As before, we are letting $m = \dim E$, $n = \dim M$.) If we pull $\overrightarrow{\mathcal{H}_E}$ back to $E^\perp$, we get an m-dimensional smooth vector bundle $\overrightarrow{\mathcal{H}_E}\lceil E^\perp$ on $E^\perp$. Since $d\pi_M^*(\xi)$ maps $\overrightarrow{\mathcal{H}_E}(\xi)$ isomorphically onto $E(\pi_M^*(\xi))$ for every $\xi \in E^\perp$, we see that *the bundle $\overrightarrow{\mathcal{H}_E}\lceil E^\perp$ can be canonically identified with the pullback $\left(\pi_M^*\lceil E^\perp\right)^* E$ of E to $E^\perp$ via the restriction to $E^\perp$ of the projection π_M^**. This means, in particular, that *every Riemannian metric on E gives rise to a Riemannian metric on $\overrightarrow{\mathcal{H}_E}\lceil E^\perp$*, a fact that will be important later. ∎

Remark 7 The abnormal biextremals, as defined above, are exactly the *characteristic curves* of $E^\perp$ that lie in $T^\# M$. Recall that, if S is a submanifold of a symplectic manifold (N, Ω), the *characteristic singular distribution* induced by Ω on S is the map $S \ni s \to D(s) = K(s) \cap T_s S$, where $K(s)$ is the orthogonal complement of T_sS relative to the 2-form Ω, i.e. $K(s) = \{v \in T_sN : \Omega(v,w) = 0 \text{ for all } w \in T_sS\}$. A *characteristic curve of* S is a D-curve in S, i.e. a curve $\Gamma : I \to S$ such that $\dot\Gamma(t) \in D(\Gamma(t))$ for almost all $t \in I$. If S is defined by equations $H_1(s) = \ldots = H_m(s) = 0$, where the m smooth functions have linearly independent differentials at every point of S, then the tangent space T_sS is the set of all vectors $v \in T_sN$ that satisfy $\langle dH_i(s), v\rangle = 0$ for $i = 1,\ldots,m$. Equivalently, $v \in T_sS$ if and only if $\Omega(v, \overrightarrow{H_i}(s)) = 0$ for $i = 1,\ldots,m$. So $K(s)$ is the linear span of the vectors $\overrightarrow{H_i}(s)$, $i = 1,\ldots,m$. If $\mathcal{H}$ denotes the linear span (over $\mathbb{R}$) of the functions $H_1,\ldots,H_m$, then a characteristic curve of S is precisely a curve Γ along which the H_i vanish and whose tangent vector $\dot\Gamma(t)$ at almost every point t belongs to $\overrightarrow{\mathcal{H}}(\Gamma(t))$.

In our case, if we take $N = T^*M$, $\Omega = \Omega_M$, $S = E^\perp$, and a curve $\Gamma : I \to T^*M$ is contained in a set of the form $V = (\pi_M^*)^{-1}(U)$, where U is the domain of a basis $f_1,\ldots,f_m$ of sections of E, then Γ is a characteristic curve of $E^\perp$ if and only if the functions H_i ($\stackrel{\text{def}}{=} H_{f_i}$) vanish along Γ and the vector $\dot\Gamma(t)$ is in the linear span of the $\overrightarrow{H_i}(\Gamma(t))$ for almost every t. In other words, Γ has to be a null bicharacteristic of the $\mathbb{R}$-linear span of $H_1,\ldots,H_m$. Now, this linear span is easily seen to be equivalent to $\mathcal{H}_E$ on V, and then Proposition 2 implies that the null bicharacteristics of $\mathcal{H}_E$ on V are the same as those of the linear span of the H_i. ▮

Remark 8 Let E be a smooth distribution on a manifold M. Let $p \in M$. Fix a Riemannian metric G on E. Let $\mathcal{D}_E(p)$ denote the set of all E-arcs $\gamma : [0,1] \to M$ that have finite energy (i.e. are such that $\int_0^1 \|\dot\gamma(t)\|_G^2 dt < \infty$) and satisfy $\gamma(0) = p$. Then this set is independent of the choice of the metric G, and Bismut has shown in [3] how to define a differentiable structure on $\mathcal{D}_E(p)$ such that $\mathcal{D}_E(p)$ is a Hilbert manifold. There is a well defined map, namely the *endpoint map*, $\mathcal{E}_p$:

$\mathcal{D}_E(p) \to M$ that assigns to each $\gamma \in \mathcal{D}_E(p)$ the end point of γ, i.e. $\mathcal{E}_p(\gamma) = \gamma(1)$. Then $\mathcal{E}_p$ is a smooth map. An E-arc $\gamma : [0,1] \to M$ is called *singular* if it is a singular point of $\mathcal{E}_{\gamma(0)}$, i.e. if the differential of $\mathcal{E}_{\gamma(0)}$ at γ is not surjective. The differential of $\mathcal{E}_p$ is easily calculated, and the calculation yields the result that *the abnormal extremals are precisely the singular arcs.* ∎

We conclude this section by singling out two simple cases when the abnormal extremals are trivial.

Proposition 4 *Let E be a smooth distribution on the smooth manifold M. Then (i) if $E = TM$ then there are no abnormal extremals of E; (ii) if E is strongly bracket-generating but $E \neq TM$ then the abnormal extremals of E are exactly the constant curves.*

PROOF. Assume $\Gamma : I \to T^{\#}M$ is an abnormal biextremal of E, and write $\Gamma(t) = (\gamma(t), \lambda(t))$. If Y is any smooth section of E, then H_Y vanishes along Γ. If $E = TM$, this implies that all the functions H_Y, for all smooth vector fields Y on M, vanish along Γ, so $\lambda(t) = 0$ for all t, contradicting the fact that $\Gamma(t) \in T^{\#}M$. If E is SBG, let $t \in I$ be such that $\dot{\Gamma}(t)$ exists and is equal to $\overrightarrow{H_X}(\Gamma(t))$ for some smooth section X of E. Then $\{H_X, H_Y\}(\Gamma(t)) = 0$ and $H_Y(\Gamma(t)) = 0$ for every smooth section Y of E, so $\lambda(t)$ annihilates $E(\gamma(t)) + [X, \mathrm{Sec}(E)](\gamma(t))$. Since $\lambda(t) \neq 0$, it follows that $X(\gamma(t)) = 0$, i.e. $\dot{\gamma}(t) = 0$. So $\dot{\gamma} = 0$ a.e., and then γ is constant. ∎

5 Sub-Riemannian manifolds, length minimizers and extremals

A *Riemannian metric* on a bundle E is a smooth section $p \to G_p$ of the bundle $E^* \otimes E^*$ such that for each $p \in M$ the bilinear form $E(p) \times E(p) \ni (v, w) \to G_p(v, w) \in \mathbb{R}$ is symmetric and strictly positive definite. It is clear that one can always construct a Riemannian metric on any subbundle E of TM by just taking a Riemannian metric on TM and restricting it to E. Conversely, every Riemannian metric on a smooth distribution E on M arises in this way, as can be shown by using partitions of unity.

Definition 5 A *sub-Riemannian structure* on a manifold M is a pair (E, G) such that E is a smooth distribution on M and G is a Riemannian metric on E. A *sub-Riemannian manifold* is a triple (M, E, G) such that M is a manifold and (E, G) is a sub-Riemannian structure on M.

5.1 The sub-Riemannian distance, length minimization and time optimality

If (M, E, G) is a sub-Riemannian manifold, and $p \in M$, $v, w \in E(p)$, we use $\langle v, w\rangle_G$ to denote the number $G_p(v, w)$. If $p \in M, v \in E(p)$, the *length* $\|v\|_G$ of v is the number $\langle v, v\rangle_G^{1/2} = G_p(v, v)^{1/2}$. The *length* $\|\gamma\|_G$ of an E-arc $\gamma : [a, b] \to M$ is the integral $\int_a^b \|\dot\gamma(t)\|_G dt$. We agree to define $\|\gamma\|_G$ to be $+\infty$ if γ is not an E-arc.

An E-arc γ is *parametrized by arc-length* if $\|\dot\gamma(t)\|_G = 1$ for almost all t in the domain of γ. If $\gamma : [a, b] \to M$ is E-admissible, then we can define another arc $\hat\gamma$, called the *arc-length reparametrization of* γ, as follows. We first define a real-valued function τ on $[a, b]$ by $\tau(t) = \int_a^t \|\dot\gamma(s)\|_G ds$. Then τ is a monotonically nondecreasing function on $[a, b]$ with range $[0, \|\gamma\|_G]$. Moreover, if $t_1 < t_2$ but $\tau(t_1) = \tau(t_2)$, then $\dot\gamma(t) = 0$ for almost all $t \in [t_1, t_2]$, so $\gamma(t_2) = \gamma(t_1)$. We can therefore define $\hat\gamma : [0, \|\gamma\|_G] \to M$ by letting $\hat\gamma(s) = \gamma(t)$ if $\tau(t) = s$. Then it can be shown that $\hat\gamma$ is E-admissible, parametrized by arc-length, and satisfies $\hat\gamma \circ \tau = \gamma$.

If E is a smooth distribution on a smooth manifold M, equipped with a Riemannian metric G, then we define $d_G : M \times M \to \mathbb{R} \cup \{\infty\}$ by

$d_G(p,q) = \inf\{\|\gamma\|_G : \partial\gamma = (p,q)\}$. The number $d_G(p,q)$ is the *distance* from p to q. Clearly, $d_G(p,q) < \infty$ iff $p \sim_E q$. If M is connected and E is bracket-generating, then it is well known that $d_G : M \times M \to \mathbb{R}$ is a metric that induces the topology of M. In the general case we can consider, for every E-orbit S of E, the restriction d_G^S of d_G to $S \times S$, which turns out to be a metric on S. If the restriction $E\lceil S$ is bracket-generating as a distribution on S, then the metric d_G^S induces the topology corresponding to the canonical differentiable structure of S. (The condition that $E\lceil S$ is bracket-generating is always satisfied, for every E-orbit S, if E is a C^ω distribution on a C^ω manifold M, but can fail when E is just of class C^∞. In that case, it is easy to construct examples where the topology induced by d_G^S is strictly stronger than the one arising from the canonical differentiable structure of S.)

The function d_G is known as the *sub-Riemannian distance* associated to (M,E,G).

Definition 6 Let (M,E,G) be a sub-Riemannian manifold. An E-arc $\gamma : [a,b] \to M$ is called a *minimizer* if $d_G(\gamma(a),\gamma(b)) = \|\gamma\|_G$. ∎

We let $\mathcal{M}(M,E,G)$ be the class of all minimizers, and $\mathcal{M}_{PAL}(M,E,G)$ be the class of those minimizers that are parametrized by arc-length. We then know that every arc $\gamma \in \mathcal{M}(M,E,G)$ is equivalent modulo reparametrization to a $\hat{\gamma} \in \mathcal{M}_{PAL}(M,E,G)$, so the problem of completely characterizing all minimizers is equivalent to that of characterizing all the minimizers that are parametrized by arc-length. Let $\Sigma(M,E,G)$ be the class of all E-arcs δ that satisfy $\|\dot{\delta}(t)\|_G \le 1$ for almost all t. It is clear that every E-arc which is parametrized by arc-length is in $\Sigma(M,E,G)$. For an arc $\gamma \in \Sigma(M,E,G)$, we will use $T(\gamma)$ to denote the *time* (or *duration*) of γ, i.e. the number $T(\gamma) = b - a$, where $[a,b] = \mathrm{Domain}(\gamma)$.

Definition 7 If $\gamma \in \Sigma(M,E,G)$, we call γ *time-minimizing*, or *time-optimal*, if $T(\gamma) \le T(\tilde{\gamma})$ whenever $\tilde{\gamma} \in \Sigma(M,E,G)$, $\partial\tilde{\gamma} = \partial\gamma$. ∎

Definition 8 If $\gamma \in \Sigma(M, E, G)$, we call γ *uniquely time-minimizing*, or *uniquely time-optimal*, if $T(\gamma) \leq T(\tilde{\gamma})$ for every $\tilde{\gamma} \in \Sigma(M, E, G)$ such that $\partial\tilde{\gamma} = \partial\gamma$, and equality only holds if $\tilde{\gamma} = \gamma$. ▮

Definition 9 If $\gamma : [a, b] \to M$ belongs to $\Sigma(M, E, G)$, we call γ *locally time-optimal* (resp. *locally uniquely time-optimal*), if there exists a $\delta > 0$ such that the restriction of γ to every subinterval of $[a, b]$ of length $\leq \delta$ is time-optimal (resp. uniquely time-optimal). ▮

We let $\mathcal{TO}(M, E, G)$ be the class of all time-optimal arcs in $\Sigma(M, E, G)$. We then have

$$\mathcal{M}_{PAL}(M, E, G) = \mathcal{TO}(M, E, G) \ . \tag{12}$$

Indeed, if $\gamma \in \mathcal{M}_{PAL}(M, E, G)$, then $T(\gamma) = \|\gamma\|_G$. Suppose there was an arc $\gamma^* \in \Sigma(M, E, G)$ such that $\partial\gamma^* = \partial\gamma$ but $T(\gamma^*) < T(\gamma)$. Since $\|\dot{\gamma}^*(t)\|_G \leq 1$ for almost all t, we have $\|\gamma^*\|_G \leq T(\gamma^*)$, and therefore $\|\gamma^*\|_G < \|\gamma\|_G$, contradicting the assumption that γ was a minimizer. Conversely, suppose that $\gamma \in \mathcal{TO}(M, E, G)$. If γ was not parametrized by arc-length, then we would have $\|\dot{\gamma}(t)\|_G \leq 1$ for almost all t (because $\gamma \in \Sigma(M, E, G)$) and $\|\dot{\gamma}(t)\|_G < 1$ for t in a set of positive measure. Therefore $\|\gamma\|_G < T(\gamma)$. If $\hat{\gamma}$ is the arc-length reparametrization of γ, then $\hat{\gamma} \in \Sigma(M, E, G)$, $\partial\hat{\gamma} = \partial\gamma$, and $T(\hat{\gamma}) = \|\hat{\gamma}\|_G = \|\gamma\|_G < T(\gamma)$, contradicting the time-optimality of γ. So γ is parametrized by arc-length. If γ was not a minimizer, there would exist a γ^* which is parametrized by arc-length and satisfies $\partial\gamma^* = \partial\gamma$ and $\|\gamma^*\|_G < \|\gamma\|_G$. But this implies that $T(\gamma^*) < T(\gamma)$, since both γ and γ^* are parametrized by arc-length. Since $\gamma^* \in \Sigma(M, E, G)$, γ is not time-optimal, and we have reached a contradiction, completing the proof of (12).

5.2 Extremals

Let (M, E, G) be a sub-Riemannian manifold. In view of (12), the problem of characterizing the minimizers for (M, E, G) is equivalent to that of characterizing the time-optimal arcs. It turns out that $\Sigma(M, E, G)$ is the set of trajectories of a control system, and this fact can be used to obtain a necessary condition for a $\gamma \in \Sigma(M, E, G)$ to be time-optimal by applying the "Maximum Principle" of Optimal Control Theory. The condition says that γ has to be an "extremal," so we begin by defining what it means for a curve to be an extremal.

If $(p, \lambda) \in T^*M$, then the restriction $\lambda \lceil E(p)$ of λ to the subspace $E(p)$ of T_pM has a well-defined norm, since $E(p)$ is an inner product space. We will use $\|\lambda\|_G$ to denote this norm.

Definition 10 The function $H : T^*M \to \mathbb{R}$ given by

$$H(x, \lambda) = -\frac{1}{2}\|\lambda\|_G^2 \tag{13}$$

is the *Hamiltonian* of the sub-Riemannian structure (E, G). ∎

Definition 11 A *normal biextremal* of a sub-Riemannian structure (E, G) is a curve Γ in T^*M such that (i) Γ is a bicharacteristic of the Hamiltonian H of (E, G), and (ii) H does not vanish along Γ. A *normal extremal* of (E, G) is a curve in M which is the projection of a normal biextremal. ∎

Definition 12 A *biextremal* of a sub-Riemannian structure (E, G) is a curve which is either a normal biextremal of (E, G) or an abnormal biextremal of E. An *extremal* of (E, G) is a curve which is either a normal extremal of (E, G) or an abnormal extremal of E. ∎

With the above definitions, the normal extremals (resp. abnormal extremals, extremals) are exactly the projections of the normal biextremals (resp. abnormal biextremals, biextremals). Biextremals are always contained in $T^\# M$, and extremals in M.

The Hamiltonian H of (E,G) is a smooth function on T^*M. Indeed, if U is any open subset of M on which there exists an orthonormal basis $(f_1,\ldots,f_m)$ of smooth sections of E, then

$$H(x,\lambda) = -\frac{1}{2}\sum_{k=1}^{m}\langle\lambda, f_k(x)\rangle^2 \quad \text{for} \quad (x,\lambda)\in(\pi_M^*)^{-1}(U)\,, \tag{14}$$

which in particular proves the smoothness of H. Moreover, if V is an open relatively compact subset of U, and $\varphi : M \to \mathbb{R}$ is a smooth function such that $\varphi \equiv 1$ on V and $\mathrm{support}(\varphi) \subseteq U$, then the vector fields $g_k \stackrel{\mathrm{def}}{=} \varphi f_k$ are global smooth sections of E. If $\Gamma : I \to T^{\#}M$ is an abnormal biextremal of E, then the functions H_{g_k} have to vanish along Γ. Therefore $H(\Gamma(t)) = 0$ if $\Gamma(t) \in (\pi_M^*)^{-1}(V)$, since $H = -\frac{1}{2}\sum_k H_{g_k}^2$ on $(\pi_M^*)^{-1}(V)$. Since every point of T^*M belongs to $(\pi_M^*)^{-1}(V)$ for some relatively compact subset V of the domain of a basis of sections of E, we conclude that H vanishes along Γ. Since, by definition, the value of H along a normal biextremal is $\neq 0$, we conclude that Γ is not a normal biextremal. Therefore *a biextremal is either normal or abnormal, and the two possibilities are mutually exclusive.* The analogous statement for extremals is not true, since it is very easy to give examples of curves γ, Γ_1, Γ_2 such that Γ_1 is a normal biextremal, Γ_2 is an abnormal biextremal, and $\gamma = \pi \circ \Gamma_1 = \pi \circ \Gamma_2$. This makes it necessary to introduce a further distinction:

Definition 13 A *strictly abnormal extremal* of a sub-Riemannian manifold (M,E,G) is an abnormal extremal of E that is not a normal extremal of (E,G). ▮

The smoothness of H clearly implies that *the normal biextremals and the normal extremals of a sub-Riemannian structure* (E,G) *are smooth.* The analogous statement for abnormal extremals is false, as can be shown by simple examples (cf. Subsection 9.6 below).

5.3 The relationship between minimality and extremality

The main result relating minimality and extremality is the following theorem, whose proof is given in Appendix B.

Theorem 1 *Let (M, E, G) be a sub-Riemannian manifold, and let $\gamma : [a, b] \to M$ be a length-minimizer parametrized by arc-length. Then γ is an extremal.*

Theorem 1 only gives useful information if the class of extremals is not too large, and in the extreme case when all trajectories are extremals it gives no information at all. This case can actually occur: it is easy to show that *if S is any orbit of positive codimension, then all E-arcs contained in S are abnormal extremals.* In particular, if all E-orbits have positive codimension, then all E-arcs are abnormal extremals and Theorem 1, as stated, gives no information. However, even in that case one can extract useful information out of Theorem 1 by applying it to the sub-Riemannian manifold $(S, E\lceil S, G_S)$, where G_S is the restriction of G to $E\lceil S$. (Indeed, it is clear that a minimizer γ for (M, E, G) which is contained in S is a minimizer for $(S, E\lceil S, G_S)$.) This yields a stronger conclusion:

Corollary 1 *If (M, E, G) is a sub-Riemannian manifold, $\gamma : [a, b] \to M$ is a length-minimizer parametrized by arc-length, and S is the E-orbit containing γ, then γ is an extremal for $(S, E\lceil S, G_S)$, i.e. γ is the projection of a biextremal $t \to \Gamma(t) = (\gamma(t), \lambda(t))$ of (M, E, G) such that $\lambda(t)$ is nonzero as a covector on S, i.e. $\lambda(t)\lceil T_{\gamma(t)}S \neq 0$ for all t.*

Remark 9 Clearly, it is always better to study minimizers by applying Corollary 1 rather than Theorem 1. It turns out that this *always* yields nontrivial information, because it is easy to prove that, on an E-connected sub-Riemannian manifold (M, E, G), any two points of M

can be joined by an E-arc which is not an extremal, so not all E-arcs are extremals. A much stronger conclusion holds if E is bracket-generating: the set of E-arcs that are not extremals is generic (cf. [21] for a proof in the real-analytic case) and, even more strongly, the set of extremals is of infinite codimension in the space of E-arcs. In particular, for real-analytic sub-Riemannian manifolds (M, E, G), the situation is as follows: for every restriction $(S, E\lceil S, G_S)$ to an E-orbit S, the distribution $E\lceil S$ is bracket-generating, and then the class of extremals that satisfy the stronger conclusion of Corollary 1 is a very small subset of the set of all E-arcs.

Many authors include the requirement that E be bracket-generating in the definition of a sub-Riemannian manifold. We have not done so, but this will mean that many questions about extremals —such as, for instance, the question whether there can exist nonsmooth abnormal extremals— will only be interesting in the bracket-generating case. ∎

The following well-known fact gives a partial converse of Theorem 1 for normal extremals:

Theorem 2 *Let (M, E, G) be a sub-Riemannian manifold. Then every normal extremal is locally optimal.*

Theorem 2 is a special case of the general fact that trajectories that can be embedded in a smooth field of extremals are locally optimal, cf. e.g. [15], [4]. An independent derivation was given by Hamenstädt in [12]. (A self-contained version of the classical control-theoretic proof is included, for completeness, in Appendix C.)

In view of Theorems 1 and 2, it is natural to ask whether abnormal extremals are also locally optimal. The answer turns out to be negative: in Appendix D we give an example of a real-analytic, C^1-rigid abnormal extremal which is not locally optimal, for a bracket-generating sub-Riemannian structure.

Finally, one could ask whether minimizers that are not normal exist at all, and how common they are. The first question was given an affirmative answer by Montgomery in [17]. (A much simpler example has been presented above in Subsection 2.3.) The second one will be answered —for rank two distributions— in the following two sections, where we will define a large class of abnormal extremals for which an analogue of Theorem 2 holds.

6 Regular abnormal extremals for rank-two distributions

We now make a more detailed study of the abnormal extremals in sub-Riemannian manifolds (M, E, G) such that $\dim E = 2$. We first note that if $\dim M = 2$ then (M, E, G) is in fact a Riemannian manifold, so there are no abnormal extremals. So from now on we will assume that $\dim M = n \geq 3$ and $\dim E = 2$.

6.1 The regular abnormal foliation of a rank-two distribution

We begin by studying the abnormal biextremals associated to a two-dimensional distribution E. We will show that all these arcs lie in a set $AEC(E)$ —the "abnormal extremal carrier" of E— and that there is a subset $RA(E)$ of $AEC(E)$ —the "regular abnormal set"— such that (i) $RA(E)$ is a submanifold of T^*M, (ii) $RA(E)$ admits a canonical foliation F_E with one-dimensional leaves, (iii) the abnormal biextremals contained in $RA(E)$ are precisely the arcs in $RA(E)$ that are entirely contained in one of the leaves. Moreover, we will show that (under a mild extra condition on E) the set $RA(E)$ is open and dense in $AEC(E)$, thereby justifying the claim that the abnormal biextremals that are contained in $RA(E)$ are the typical abnormal biextremals.

Recall that $E^\perp$ denotes the annihilator of E. Then $E^\perp$ is a submanifold of T^*M of codimension two, and it is clear from Definition 4 that every abnormal biextremal has to be contained in $E^\perp \cap T^\# M$. So it may appear natural to define a "typical abnormal extremal" to be one that goes through a "generic point of $E^\perp \cap T^\# M$." This, however, would only produce a rather trivial and uninteresting class of abnormal biextremals, namely, constant curves in $E^\perp \cap T^\# M$. (It follows easily from Definition 4 that every constant trajectory $t \to (p(t), \lambda(t)) \in E^\perp \cap T^\# M$ is an abnormal biextremal.) It turns out that through most points of $E^\perp \cap T^\# M$ there pass no other abnormal biextremals, because along any nonconstant abnormal biextremal a further condition has to hold, namely, the vanishing of the Poisson bracket $\{H_f, H_g\}$ of the Hamiltonians H_f, H_g of a pair of independent sections of E:

Proposition 5 *If E is a smooth two-dimensional distribution on a manifold M, then every nonconstant abnormal biextremal of E is contained in $(E^2)^\perp \cap T^\# M$.*

PROOF. Let Γ be a nonconstant abnormal biextremal. It follows from Definition 4 that Γ is contained in $T^\# M$. We now prove that Γ is also contained in $(E^2)^\perp$. It clearly suffices to prove this under the extra assumption that the domain of Γ is a bounded open interval (a, b). Let A be the set of those $t \in (a, b)$ such that $\dot\Gamma(t)$ exists and belongs to $\mathcal{H}_E(\Gamma(t))$. Then $(a, b) - A$ has measure zero. Let I denote the set of those $t \in (a, b)$ such that Γ is constant on the interval $(t - \varepsilon, t + \varepsilon)$ for some $\varepsilon > 0$. Then I is an open subset of (a, b), and $I \neq (a, b)$. Let $t \in (a, b)$ and let $\Gamma(t) = (p, \lambda)$. Let U be a neighborhood of p in M such that there are smooth global sections f,g of E that are independent at every point of U. Let $V = (\pi_M^*)^{-1}(U)$. Let $\varepsilon > 0$ be such that $\Gamma(s) \in V$ for $s \in (t - \varepsilon, t + \varepsilon)$. If X is any section of E, then there are smooth functions α, β on U such that $X = \alpha f + \beta g$ on U. Then $\overrightarrow{H_X} = \alpha \overrightarrow{H_f} + \beta \overrightarrow{H_g} + H_f \overrightarrow{\alpha} + H_g \overrightarrow{\beta}$ on V. In particular, since H_f and H_g vanish along Γ, we have $\overrightarrow{H_X}(\Gamma(s)) = (\alpha\overrightarrow{H_f} + \beta\overrightarrow{H_g})(\Gamma(s))$ whenever $\Gamma(s) \in V$. For every $s \in (t-\varepsilon, t+\varepsilon) \cap A$, we have $\dot{\Gamma}(s) \in \overrightarrow{\mathcal{H}_E}(\Gamma(s))$, which means that there exists a smooth section X_s of E such that $\dot{\Gamma}(s) = \overrightarrow{H_{X_s}}(\Gamma(s))$. But then $\overrightarrow{H_{X_s}}(\Gamma(s))$ is a linear combination of $\overrightarrow{H_f}(\Gamma(s))$ and $\overrightarrow{H_g}(\Gamma(s))$. Write $\overrightarrow{H_{X_s}}(\Gamma(s)) = \varphi(s)\overrightarrow{H_f}(\Gamma(s)) + \psi(s)\overrightarrow{H_g}(\Gamma(s))$. Since H_f vanishes along Γ, the Poisson bracket $\{H_{X_s}, H_f\}$ must vanish at $\Gamma(s)$. If $\psi(s) \neq 0$, this implies that $\{H_g, H_f\}(\Gamma(s)) = 0$. If $\varphi(s) \neq 0$, then the same conclusion follows using the fact that H_g vanishes along Γ. So $\{H_g, H_f\}(\Gamma(s)) = 0$ for every $s \in A \cap (t-\varepsilon, t+\varepsilon)$ such that $\dot{\Gamma}(s) \neq 0$.

Now assume that $t \notin I$. Then there cannot exist a $\delta > 0$ such that $\dot{\Gamma}(s) = 0$ for all $s \in A \cap (t-\delta, t+\delta)$. So every interval $(t-\delta, t+\delta)$ contains an $s \in A$ for which $\dot{\Gamma}(s) \neq 0$, and then $\{H_g, H_f\}(\Gamma(s)) = 0$. Since Γ and $\{H_g, H_f\}$ are continuous, we conclude that $\{H_g, H_f\}(\Gamma(t)) = 0$. Since $\{H_g, H_f\} = H_{[f,g]}$, we have $H_{[f,g]}(\Gamma(t)) = 0$. So $\langle\lambda, f(p)\rangle = \langle\lambda, g(p)\rangle = \langle\lambda, [f,g](p)\rangle = 0$. Since $f(p)$, $g(p)$ and $[f,g](p)$ linearly span $E^2(p)$, we have shown that $(p,\lambda) \in (E^2)^\perp$.

So far, we have established that $\Gamma(t) \in (E^2)^\perp$ for all $t \in (a,b)$ such that $t \notin I$. Suppose now that $t \in I$. Let J be the connected component of I that contains t. Then J is an open subinterval of (a,b), and $J \neq (a,b)$. So there is a point τ in the closure of J relative to (a,b) such that $\tau \notin J$, and therefore $\tau \notin I$. Since Γ is constant on J, we have $\Gamma(t) = \Gamma(\tau)$. But, since $\tau \notin I$, we already know that $\Gamma(\tau) \in (E^2)^\perp$. So $\Gamma(t) \in (E^2)^\perp$. ∎

Remark 10 As pointed out before, every *constant* trajectory in $E^\perp$ is an abnormal biextremal. Naturally, such an abnormal biextremal need not be contained in $(E^2)^\perp$. ∎

Having shown that all the nonconstant abnormal biextremals must be contained in $(E^2)^\perp \cap T^\# M$, it is natural to call the set

$$AEC(E) \stackrel{\text{def}}{=} (E^2)^\perp \cap T^\# M$$

the *abnormal extremal carrier* of E, provided that we show that all, or almost all, the points in $AEC(E)$ belong to nonconstant abnormal biextremals. More precisely, we will define a subset $RA(E)$ of $AEC(E)$ which, under a mild restriction on E, turns out to be open and dense in $AEC(E)$. This subset will be a submanifold of $E^{\perp}$ of codimension one, and will admit a one-dimensional foliation F_E —defined by a line subbundle L_E of the tangent bundle $T\,RA(E)$— with the property that every arc contained in a leaf of F_E is an abnormal biextremal of E.

To define $RA(E)$, L_E and F_E we must first study the local structure of the set $(E^2)^{\perp}$. Pick an open subset U of M on which there exists a basis (f, g) of smooth sections of E, and write $V = (\pi_M^*)^{-1}(U)$. Then $(E^2)^{\perp} \cap V$ is the subset of $E^{\perp} \cap V$ defined by the equation $\{H_f, H_g\} = 0$. On the other hand, the tangent space to $E^{\perp}$ at a point $\xi \in E^{\perp} \cap V$ is the set of all vectors $v \in T_\xi T^*M$ that satisfy $\langle dH_f(\xi), v\rangle = \langle dH_g(\xi), v\rangle = 0$. So $T_\xi E^{\perp}$ is precisely the orthogonal subspace —relative to the form Ω_M— of the two-dimensional subspace $\overrightarrow{\mathcal{H}_E}(\xi)$. The restriction of $\Omega_M(\xi)$ to $\overrightarrow{\mathcal{H}_E}(\xi)$ is nonsingular if $\{H_f, H_g\}(\xi) \neq 0$, and vanishes identically if $\{H_f, H_g\}(\xi) = 0$. So $\xi \in (E^2)^{\perp}$ if and only if $\overrightarrow{\mathcal{H}_E}(\xi) \subseteq T_\xi E^{\perp}$.

Definition 14 The set $RA(E) \stackrel{\text{def}}{=} (E^2)^{\perp} - (E^3)^{\perp}$ is the *regular abnormal set* of E. A point $(p, \lambda) \in T^*M$ is a *regular abnormal point* (RAP) for E if $\lambda \in E^2(p)^{\perp} - E^3(p)^{\perp}$, i.e. if $(p, \lambda) \in RA(E)$. ∎

Suppose $\xi = (p, \lambda) \in RA(E)$, and let U, V, f, g be as before. The subset $(E^2)^{\perp}$ of the submanifold $E^{\perp}$ is defined, near ξ, by the equation $\{H_f, H_g\} = 0$. Since $\xi \in (E^2)^{\perp}$, the space $\overrightarrow{\mathcal{H}_E}(\xi)$ is a subset of $T_\xi E^{\perp}$. So the vectors $\overrightarrow{H_f}(\xi)$ and $\overrightarrow{H_g}(\xi)$ are tangent to $E^{\perp}$ at ξ. On the other hand, since $\xi \notin (E^3)^{\perp}$, at least one of the numbers $\langle \lambda, [f, [f, g]](p)\rangle$, $\langle \lambda, [g, [f, g]](p)\rangle$ is $\neq 0$. This says that the directional derivative of $\{H_f, H_g\}$ in at least one of the directions $\overrightarrow{H_f}(\xi)$, $\overrightarrow{H_g}(\xi)$ is $\neq 0$. Therefore the differential at ξ of the restriction to $E^{\perp}$ of

$\{H_f, H_g\}$ is $\neq 0$. This implies that, in a neighborhood of ξ, the set $S = \{\eta \in E^\perp : \{H_f, H_g\}(\eta) = 0\}$ is a submanifold of $E^\perp$ of codimension one. Moreover, it also implies that the space $\overrightarrow{\mathcal{H}_E}(\xi)$ —which, as we know, is a subset of the tangent space $T_\xi E^\perp$— is not entirely contained in the tangent space $T_\xi S$, so $\overrightarrow{\mathcal{H}_E}(\xi) \cap T_\xi S$ is one-dimensional. Finally, it is clear that S is, near ξ, the same as $(E^2)^\perp$, and also the same as $(E^2)^\perp - (E^3)^\perp$, since $\xi \in (E^2)^\perp - (E^3)^\perp$ and $(E^3)^\perp$ is closed. We have therefore proved that $RA(E) \overset{\text{def}}{=} (E^2)^\perp - (E^3)^\perp$ is a submanifold of $E^\perp$ of codimension one, and that the two-dimensional space $\overrightarrow{\mathcal{H}_E}(\xi)$ is tangent to $E^\perp$ but not to $RA(E)$ at every point of $RA(E)$. This makes it possible to define a line bundle L_E on $RA(E)$ by letting

$$L_E(\xi) = \overrightarrow{\mathcal{H}_E}(\xi) \cap T_\xi RA(E)\,, \tag{15}$$

so $L_E(\xi)$ is a 1-dimensional subspace of $T_\xi RA(E)$. Then L_E is a smooth line subbundle of the tangent bundle of $RA(E)$. It is clear that $L_E(\xi) \subseteq \overrightarrow{\mathcal{H}_E}(\xi)$ for every $\xi \in RA(E)$. So every absolutely continuous curve $\Gamma : [a, b] \to RA(E)$ that satisfies $\dot{\Gamma}(t) \in L_E(\Gamma(t))$ for almost all t is a bicharacteristic of $\mathcal{H}_E$, and is obviously null, since $RA(E) \subseteq E^\perp$. Therefore every such Γ is an abnormal biextremal of E. Conversely, every abnormal biextremal Γ of E which goes through a point $\xi \in RA(E)$ at a time t must satisfy $\Gamma(s) \in RA(E)$ for all $s \in (t-\varepsilon, t+\varepsilon)$, and $\dot{\Gamma}(s) \in L_E(\Gamma(s))$ for almost all $s \in (t-\varepsilon, t+\varepsilon)$, if $\varepsilon > 0$ is small enough. (This is obviously true if Γ is constant. If Γ is nonconstant, Proposition 5 implies first of all that Γ is contained in $(E^2)^\perp$, and then $\Gamma(s)$ must belong to $RA(E)$ for s near t, since $RA(E)$ is relatively open in $(E^2)^\perp$. But then, for almost all s, $\dot{\Gamma}(s)$ must belong to $\overrightarrow{\mathcal{H}_E}(\Gamma(s))$ and be tangent to $RA(E)$, so $\dot{\Gamma}(s) \in L_E(\Gamma(s))$.)

Clearly, L_E is an involutive distribution on E. We use F_E to denote the set of leaves of L_E, and refer to F_E as *the regular abnormal foliation* induced by E. We summarize the above into the following

Proposition 6 *If E is a smooth two-dimensional distribution on a manifold M such that $\dim M \geq 3$, then the set $RA(E) \stackrel{\mathrm{def}}{=} (E^2)^\perp - (E^3)^\perp$ of regular abnormal points of E is a smooth submanifold of codimension three of $T^\# M$, and a smooth submanifold of codimension one of $E^\perp$. On $RA(E)$ there is a one-dimensional subbundle L_E of $T\,RA(E)$ (defined by letting $L_E(\xi) \stackrel{\mathrm{def}}{=} \overrightarrow{\mathcal{H}_E}(\xi) \cap T_\xi RA(E)$) that gives rise to a foliation F_E of $RA(E)$, with the property that an arc Γ in $RA(E)$ is an abnormal biextremal of E if and only if it is contained in a leaf of F_E.* ∎

It will be useful to have an explicit description of a section of L_E. Let $\xi = (p, \lambda) \in RA(E)$. Let (f, g) be, as before, a basis of sections of E near p. Then a linear combination $v = a\overrightarrow{H_f}(\xi) + b\overrightarrow{H_g}(\xi)$ is in $L_E(\xi)$ if and only if $v\{H_f, H_g\} = 0$, that is, if and only if

$$a\{H_f, \{H_f, H_g\}\}(\xi) + b\{H_g, \{H_f, H_g\}\}(\xi) = 0\,.$$

The fact that $\lambda \in RA(E)$ implies that

$$\{H_f, \{H_f, H_g\}\}(\xi)^2 + \{H_g, \{H_f, H_g\}\}(\xi)^2 \neq 0$$

So we can get a nonzero element $V_{f,g}(\xi) \in L_E(\xi)$ by choosing

$$a = \{H_g, \{H_f, H_g\}\}(\xi)\,, \quad b = -\{H_f, \{H_f, H_g\}\}(\xi)\,.$$

Then

$$V_{f,g}(\xi) = \{H_g, \{H_f, H_g\}\}(\xi)\overrightarrow{H_f}(\xi) - \{H_f, \{H_f, H_g\}\}(\xi)\,, \tag{16}$$

and we have proved:

Proposition 7 *If $\xi = (p, \lambda) \in RA(E)$ and (f, g) is a basis of sections of E in a neighborhood of p, then $L_E(\xi)$ is spanned by the vector $V_{f,g}(\xi)$ given by (16).* ∎

An important corollary of Proposition 7 is the following observation:

Corollary 2 *If E is a smooth two-dimensional distribution on a manifold M, then the line subbundle L_E of the tangent bundle of $RA(E)$ has a canonical orientation.*

PROOF. We define the orientation to be the one that makes the vector $V_{f,g}(\xi)$ positive, where (f,g) is an arbitrary basis of sections near p. Notice that, if two bases (f,g), (F,G) of sections are positively related, then we can deform (F,G) continuously to (f,g) near p, and produce a smooth one-parameter family of bases (f_s,g_s) such that $(f_0,g_0) = (f,g)$ and $(f_1,g_1) = (F,G)$. Since $V_{f_s,g_s}(\xi)$ never vanishes, we see that $V_{F,G}(\xi)$ and $V_{f,g}(\xi)$ are positive multiples of each other. On the other hand, direct inspection of the formula for $V_{f,g}(\xi)$ shows that $V_{f,g}(\xi) = V_{g,f}(\xi)$. So the orientation determined by $V_{f,g}(\xi)$ does not change when we reverse the orientation of the basis (f,g) either. ∎

We now turn to the question of the *genericity* of the regular abnormal points within the abnormal extremal carrier $AEC(E)$. We show that the RAP's form a relatively open dense subset of $AEC(E)$, provided that E satisfies a mild condition that holds, in particular, for generic distributions and for all regular bracket-generating distributions. To state the condition, we first define, for each finite sequence $(d_1,\ldots,d_m)$ of positive integers, and each smooth distribution E on a manifold M, a set $S_E(d_1,\ldots,d_m)$ by letting

$$S_E(d_1,\ldots,d_m) = \{p \in M : \dim E^k(p) = d_k \ \text{ for } \ k = 1,\ldots,m\}. \quad (17)$$

Definition 15 A two-dimensional distribution E on a manifold M is *weakly regular* if

$$S_E(2,2,2) \subseteq \mathrm{Clos}\Big(S_E(2,2) - S_E(2,2,2)\Big). \quad (18)$$

Proposition 8 *If E is a two-dimensional weakly regular smooth distribution on a manifold M, then $RA(E)$ is a relatively open dense subset of $(E^2)^\perp \cap T^\# M$.*

PROOF. It is clear that $RA(E)$ is relatively open in $(E^2)^\perp$. To prove that it is dense in $(E^2)^\perp \cap T^\# M$, we pick a $(p,\lambda) \in (E^2)^\perp \cap T^\# M$, and show that (p,λ) is a limit of a sequence $\{(p_j,\lambda_j)\}$ of points in $(E^2)^\perp - (E^3)^\perp$.

We consider first the case when $\dim E^3(p) \geq 3$. If $E^3(p) \neq E^2(p)$, we can just take all the p_j equal to p, and approximate λ by a sequence $\{\lambda_j\}$ of functionals in $E^2(p)^\perp - E^3(p)^\perp$. If $E^3(p) = E^2(p)$, then necessarily $\dim E^2(p) = 3$, and $\dim M > 3$ because $\lambda \neq 0$. Then $\dim E^2(x) = 3$ for x in a neighborhood of p. On the other hand, if $\dim E^3(x)$ was equal to 3 for x in some neighborhood U of p, it would follow that E^2 is involutive near p, so that $E^\infty(x) = E^2(x) \neq T_x M$ for x near p, contradicting the bracket-generating property of E. So there exist points $p_j \in U$, converging to p, such that $\dim E^3(p_j) > 3$. Since $\dim E^2(p_j) = 3$, there exist $\tilde\lambda_j \in T^*_{p_j} M$ such that $\tilde\lambda_j \to \lambda$ and $\tilde\lambda_j \in E^2(p_j)^\perp$. Then we can approximate $\tilde\lambda_j$ by $\lambda_j \in E^2(p_j)^\perp - E^3(p_j)^\perp$, obtaining a sequence $\{(p_j,\lambda_j)\}$ of RAP's that converges to (p,λ).

We now turn to the remaining case, namely, when $\dim E^3(p) = 2$. Then $p \in S_E(2,2,2)$, and our hypothesis says that p is the limit of a sequence $\{p_j\}$ such that $p_j \in S_E(2,2)$ but $p_j \notin S_E(2,2,2)$. Then $\lambda = \lim \tilde\lambda_j$, $\tilde\lambda_j \in E^2(p_j)^\perp$, and $\tilde\lambda_j$ can be approximated by $\lambda_j \in E^2(p_j)^\perp - E^3(p_j)^\perp$, since $E^3(p_j) \neq E^2(p_j)$. So, once again, (p,λ) is the limit of a sequence of points in $RA(E)$. ∎

Finally, we show that weak regularity is a generic property:

Proposition 9 *On manifolds of dimension ≥ 3: (i) all regular two-dimensional distributions are weakly regular, (ii) weak regularity is a generic property of two-dimensional distributions.*

PROOF. In the regular case, $S_E(2,2,2)$ is empty. For generic two-dimensional distributions on an n-dimensional manifold, $S_E(2,2,2)$ is a submanifold of codimension $\kappa = \frac{1}{2}n(n-1)(n-2)$. If $n \geq 4$, then $\kappa > n$, so $S_E(2,2,2)$ is empty. If $n = 3$, then $\kappa = 3$, so $S_E(2,2,2)$ is a discrete set. However, $S_E(2,2)$ is a smooth two-dimensional surface, so every $p \in S_E(2,2,2)$ is a limit of points in $S_E(2,2)$ that do not belong to $S_E(2,2,2)$, so (18) is true in this case as well. ∎

6.2 Regular abnormal biextremals of a sub-Riemannian manifold

We now assume that (M,E,G) is a sub-Riemannian manifold such that $\dim E = 2$ and $\dim M \geq 3$. In that case, as explained in Remark 6, the metric G on E induces a metric on the bundle $\overrightarrow{\mathcal{H}_E}\lceil E^\perp$, since this bundle is canonically identified with the pullback of E to $E^\perp$ via $\pi_M^*\lceil E^\perp$. This metric can be restricted to $\overrightarrow{\mathcal{H}_E}\lceil RA(E)$, and then to the subbundle L_E of $\overrightarrow{\mathcal{H}_E}\lceil RA(E)$. Moreover, using the canonical orientation of L_E described in Corollary 2, we can define a vector field $\mathcal{X}_E$, by choosing $\mathcal{X}_E(\xi)$, for each $\xi \in RA(E)$, to be the vector in $L_E(\xi)$ that has length one and is positively oriented. If (f,g) is an orthonormal basis of sections of E near $\pi_M^*(\xi)$, then

$$\mathcal{X}_E(\xi) = \frac{\{H_g,\{H_f,H_g\}\}(\xi)\overrightarrow{H_f}(\xi) - \{H_f,\{H_f,H_g\}\}(\xi)\overrightarrow{H_g}(\xi)}{\sqrt{\{H_f,\{H_f,H_g\}\}(\xi)^2 + \{H_g,\{H_f,H_g\}\}(\xi)^2}}. \tag{19}$$

A smooth curve Γ in $RA(E)$ which is an integral curve of $\mathcal{X}_E$ or of $-\mathcal{X}_E$ will be called a *regular abnormal biextremal* of the sub-Riemannian manifold (M,E,G). A curve $\gamma : [a,b] \to M$ which is a projection of a regular abnormal biextremal will be called a *regular abnormal extremal* of (M,E,G). Naturally, through every point of $RA(E)$ there pass exactly two *maximal regular abnormal biextremals*, namely, the maximal integral curves of $\mathcal{X}_E$ and of $-\mathcal{X}_E$. If U is the domain of a coordinate chart κ and of an orthonormal basis (f,g) of sections of E, then the equation of the integral trajectories of $\mathcal{X}_E$ that are contained in $(\pi_M^*)^{-1}(U) \cap RA(E)$ is

$$\dot{x} = \frac{\langle\lambda,[g,[f,g]](x)\rangle f(x) - \langle\lambda,[f,[f,g]](x)\rangle g(x)}{\sqrt{\langle\lambda,[g,[f,g]](x)\rangle^2 + \langle\lambda,[f,[f,g]](x)\rangle^2}}, \tag{20}$$

$$\dot{\lambda} = \frac{-\langle\lambda,[g,[f,g]](x)\rangle\lambda\frac{\partial f}{\partial x}(x)+\langle\lambda,[f,[f,g]](x)\rangle\lambda\frac{\partial g}{\partial x}(x)}{\sqrt{\langle\lambda,[g,[f,g]](x)\rangle^2+\langle\lambda,[f,[f,g]](x)\rangle^2}}, \quad (21)$$

but it is worth emphasizing that the flow on $RA(E)$ defined by these equations is intrinsic, and depends only on the distribution E and the metric G. We also point out that an arc $t \to (x(t), \lambda(t))$ is a solution of (20) and (21), i.e. an integral curve of $\mathcal{X}_E$, if and only if $t \to (x(t), -\lambda(t))$ is an integral curve of $-\mathcal{X}_E$. Therefore, if γ is a regular abnormal extremal that happens to be a projection of an integral curve of $-\mathcal{X}_E$, then we can realize γ as a projection of an integral curve of $\mathcal{X}_E$ by just reversing the sign of the adjoint vector λ. So we have shown:

Corollary 3 *If (M, E, G) is a sub-Riemannian manifold such that* $\dim E = 2$*, then a curve γ in M is a regular abnormal extremal if and only if it is a projection of an integral curve of $\mathcal{X}_E$. In particular, through every point $p \in M$ such that $E^3(p) \neq E^2(p)$ there passes a regular abnormal extremal.* ∎

In our study of optimal abnormal extremals arising from abnormal biextremals in $RA(E)$, we will restrict our attention to regular abnormal extremals. This is possible because, to begin with, every E-arc is equivalent up to time reparametrization to a trajectory parametrized by arc-length, so we can limit ourselves to considering abnormal extremals parametrized by arc-length. If γ is such an extremal, and $\gamma = \pi_M^* \circ \Gamma$, where $\Gamma : [a, b] \to RA(E)$ is an abnormal biextremal, then Γ is an L_E-arc in $RA(E)$, also parametrized by arc-length. Moreover, if we define a curve $\gamma : I \to M$ to be *simple* if it has no double points (i.e. if γ is a one-to-one map), and *locally simple* if I can be covered by relatively open subintervals on which γ is simple, then it is clear that an arc $\gamma : [a, b] \to M$ cannot minimize length unless it is simple. *A fortiori*, if γ minimizes length and $\gamma = \pi_M^* \circ \Gamma$, then Γ has to be locally simple. The following trivial remark then implies that Γ has to be a regular abnormal biextremal.

Proposition 10 *Let S be a smooth manifold, and let L be a smooth 1-dimensional subbundle of TS, endowed with a Riemannian metric. Let $\Gamma : I \to M$ be an L-curve parametrized by arc-length. Then the following three conditions are equivalent: (i) Γ is smooth, (ii) Γ is locally simple, (iii) Γ is locally an integral curve of a smooth section of L of length one.* ∎

Corollary 4 *If $\Gamma : [a, b] \to RA(E)$ is an abnormal biextremal arc such that $\gamma = \pi_M^* \circ \Gamma$ minimizes length and is parametrized by arc-length, then Γ is a regular abnormal biextremal.* ∎

7 Local optimality of regular abnormal extremals

Our strategy for proving that regular abnormal extremals are locally optimal will consist of showing that, locally, the equations of a regular abnormal extremal can be put in a "normal form" that is particularly suited for proving optimality. We first introduce a new definition.

Definition 16 Let (M, E, G) be a sub-Riemannian manifold such that $\dim E = 2$. An E-arc $\gamma : [a, b] \to M$, parametrized by arc-length, is a *Nice Abnormal Extremal* (NAE) if there exist U, Y, ω such that the following conditions are satisfied:

(1) U is an open subset of M such that $\gamma([a, b]) \subset U$;

(2) Y is a smooth section of E with domain U such that $\|Y(p)\|_G = 1$ for all $p \in U$;

(3) γ is an integral curve of Y, i.e. $\dot\gamma(t) = Y(\gamma(t))$ for almost all $t \in [a, b]$ (that is, γ is of class C^1 and $\dot\gamma(t) = Y(\gamma(t))$ for all $t \in [a, b]$);

(4) ω is a smooth 1-form on U, such that (i) $\omega(\gamma(t))$ annihilates $E(\gamma(t))$ for all $t \in [a,b]$, (ii) for every $p \in U$, the linear functional $\omega(p)$ does not belong to the annihilator of $E^3(p)$, and (iii) the Lie derivative $L_Y\omega$ of ω with respect to Y vanishes identically along γ. ▮

The following theorem describes the precise relationship between regular and nice abnormal extremals.

Theorem 3 *Every NAE is a regular abnormal extremal. Conversely, if $\gamma : [a,b] \to M$ is a regular abnormal extremal then γ is locally a NAE, in the sense that for every $t \in [a,b]$ there exists an $\varepsilon > 0$ such that the restriction of γ to $[t-\varepsilon, t+\varepsilon] \cap [a,b]$ is a NAE.*

PROOF. Let $\gamma : [a,b] \to M$ be a NAE, and let U, Y, ω be such that (1), (2), (3), (4) hold. Let $\lambda(t) = \omega(\gamma(t))$, so λ is a field of covectors along γ. Let $\Gamma(t) = (\gamma(t), \lambda(t))$, so Γ is a smooth curve in T^*M. Since $\lambda(t) \in E(\gamma(t))^\perp$ for all $t \in [a,b]$, the curve Γ is in fact contained in $E^\perp$. If Z is any smooth vector field on U, then $Y\langle\omega, Z\rangle = \langle L_Y\omega, Z\rangle + \langle\omega, [Y,Z]\rangle$. Therefore $Y\langle\omega, Z\rangle = \langle\omega, [Y,Z]\rangle$ along γ. In particular, if Z is a section of E, then $\langle\omega, Z\rangle$ vanishes along γ, so $Y\langle\omega, Z\rangle \equiv 0$ along γ, and then $\langle\omega, [Y,Z]\rangle$ vanishes along γ as well. This implies that $\lambda(t) \in E^2(\gamma(t))^\perp$ for each $t \in [a,b]$. So Γ is in fact contained in $(E^2)^\perp$, and Condition (4.ii) implies that Γ is contained in $RA(E)$.

If Z is an arbitrary smooth vector field on U, then the fact that the equation $Y\langle\omega, Z\rangle = \langle\omega, [Y,Z]\rangle$ holds along γ says that

$$\frac{d}{dt}\langle\lambda(t), Z(\gamma(t))\rangle = \langle\lambda(t), [Y,Z](\gamma(t))\rangle\,, \tag{22}$$

that is,

$$\frac{d}{dt}H_Z(\Gamma(t)) = H_{[Y,Z]}(\Gamma(t)) = \{H_Y, H_Z\}(\Gamma(t)) = \overrightarrow{H_Y}H_Z(\Gamma(t))\,. \tag{23}$$

In other words, the directional derivative of H_Z at $\Gamma(t)$ in the direction of $\dot\Gamma(t)$ is equal to the derivative in the direction of $\overrightarrow{H_Y}(\Gamma(t))$, for every Z. It is easy to see that, if $\xi \in T^{\#}M$, then the set $\{dH_Z(\xi) : Z \in V^\infty(U)\}$ is the whole cotangent space $T^*_\xi(T^*M)$. (In coordinates, we have $\overrightarrow{H_Z}(x,\lambda) = (Z(x), -\lambda\frac{\partial Z}{\partial x}(x))$. If $\lambda \neq 0$, then we can clearly choose Z such that $Z(x) = v$ and $-\lambda\frac{\partial Z}{\partial x}(x) = w$ for any given v, w.) Hence $\langle \mu, \dot\Gamma(t)\rangle = \langle \mu, \overrightarrow{H_Y}(\Gamma(t))\rangle$ for every $\mu \in T^*_{\Gamma(t)}(T^*M)$. So $\dot\Gamma(t) = \overrightarrow{H_Y}(\Gamma(t))$. Therefore $\dot\Gamma(t) \in L_E(\Gamma(t))$. This proves that Γ is a regular abnormal biextremal.

We now prove the converse. Let $\gamma : [a,b] \to M$ be a regular abnormal extremal, and let $\Gamma : [a,b] \to RA(E)$ be a regular abnormal biextremal such that $\pi^*_M \circ \Gamma = \gamma$. Write $\Gamma(t) = (\gamma(t), \lambda(t))$. Since the result we are trying to prove is local, and γ is smooth and such that $\|\dot\gamma(t)\|_G = 1$ for all t, we may assume that M is an open subset of $\mathbb{R}^n$, E admits a global basis (f, g) of sections which is orthonormal with respect to the metric G, and $\dot\gamma(t) = (\alpha f + \beta g)(\gamma(t))$, where α and β are smooth functions on M that satisfy $\alpha^2 + \beta^2 \equiv 1$. We then define the vector field Y on M by $Y = \alpha f + \beta g$. Then Y is a smooth section of E on M such that $\|Y(x)\|_G = 1$ for all $x \in M$. We also define a smooth vector field W on T^*M by

$$W(x,\lambda) = \alpha(x)\overrightarrow{H_f}(x,\lambda) + \beta(x)\overrightarrow{H_g}(x,\lambda).$$

Since Γ is a regular abnormal biextremal, we have

$$\dot\Gamma(t) = u(t)\overrightarrow{H_f}(\Gamma(t)) + v(t)\overrightarrow{H_g}(\Gamma(t))$$

and

$$\dot\gamma(t) = d\pi^*_M(\Gamma(t)) \cdot \dot\Gamma(t) = u(t)f(\gamma(t)) + v(t)g(\gamma(t))$$

so that $u(t) = \alpha(\gamma(t))$ and $v(t) = \beta(\gamma(t))$. Therefore

$$\dot\Gamma(t) = \alpha(\gamma(t))\overrightarrow{H_f}(\Gamma(t)) + \beta(\gamma(t))\overrightarrow{H_g}(\Gamma(t)) = W(\Gamma(t)).$$

So Γ is an integral curve of W.

We now pick a $\bar{t} \in [a,b]$, write $\bar{p} = \gamma(\bar{t})$, and let S be a hypersurface through $\bar{p}$ such that $Y(\bar{p}) \notin T_{\bar{p}}S$. We let $\bar{\omega}$ be a smooth section of T^*M on S such that $\bar{\omega}(\bar{p}) = \lambda(\bar{t})$. For each $p \in S$, let Γ_p be the integral curve of W that goes through $(p, \bar{\omega}(p))$ at time $\bar{t}$. Write $\Gamma_p(t) = (\gamma_p(t), \lambda_p(t))$. By shrinking S, if necessary, we may assume that there is an $\varepsilon > 0$ such that all the curves Γ_p, for $p \in S$, are defined on the interval $(\bar{t}-\varepsilon, \bar{t}+\varepsilon)$, and the map $(p,t) \rightarrow \gamma_p(t)$ is a diffeomorphism from $S \times (\bar{t}-\varepsilon, \bar{t}+\varepsilon)$ onto an open subset U of M. We can then define a 1-form ω on U by letting $\omega(\gamma_p(t)) = \lambda_p(t)$. It is clear that $\Gamma_{\bar{p}} \equiv \Gamma$ on $J = [a,b] \cap (\bar{t}-\varepsilon, \bar{t}+\varepsilon)$. So $\omega(\gamma(t)) = \lambda(t)$ for $t \in J$. In particular, $\omega(\gamma(t)) \in E(\gamma(t))^{\perp}$ for all $t \in J$, and $\omega(\bar{p}) \notin E^3(\bar{p})^{\perp}$. So, after shrinking S and ε, if necessary, we can assume that $\omega(p) \notin E^3(p)^{\perp}$ for all $p \in U$.

To conclude our proof, we have to show that $L_Y\omega(\gamma(t)) = 0$ for all $t \in J$. Now, if Z is a smooth vector field on U, then

$$\langle L_Y\omega, Z\rangle = Y\langle \omega, Z\rangle - \langle \omega, [Y,Z]\rangle . \tag{24}$$

So our conclusion will clearly follow if we prove that the derivative of $\langle \lambda(t), Z(\gamma(t))\rangle$ is $\langle \lambda(t), [Y,Z](\gamma(t))\rangle$. This is equivalent to showing that the derivative of H_Z along Γ is $H_{[Y,Z]}$, i.e. $\{H_Y, H_Z\}$. This will follow if we prove that $\dot{\Gamma}(t) = \overrightarrow{H_Y}(\Gamma(t))$. Since $\dot{\Gamma}(t) = W(\Gamma(t))$, it suffices to prove that $\overrightarrow{H_Y}$ and W agree on Γ. Now $W = \alpha\overrightarrow{H_f} + \beta\overrightarrow{H_g}$, while $Y = \alpha f + \beta g$, so that $H_Y = \alpha H_f + \beta H_g$, and then $\overrightarrow{H_Y} = \alpha\overrightarrow{H_f} + \beta\overrightarrow{H_g} + H_f\overrightarrow{\alpha} + H_g\overrightarrow{\beta}$. Therefore $\overrightarrow{H_Y} = W + H_f\overrightarrow{\alpha} + H_g\overrightarrow{\beta}$. Since H_f and H_g vanish along Γ, we conclude that $\overrightarrow{H_Y} \equiv W$ on Γ, and the proof is complete. ∎

7.1 The main inequality

The following lemma was originally announced in the preprint [25]. (The concept of a "control system" is defined in Appendix B.)

Lemma 2 *Let $n \geq 3$, and let U be an open subset of $\mathbb{R}^n$. Let φ, $\psi_1, \psi_2, \psi_4, \ldots, \psi_n$, $\eta_1, \eta_3, \ldots, \eta_n$ be smooth functions on U, and let Σ be the control system*

$$\begin{aligned} \dot{x}_1 &= u\varphi(x) + vx_1\psi_1(x)\,, \\ \dot{x}_2 &= (1 + x_1\psi_2(x))v\,, \\ \dot{x}_j &= vx_1\psi_j(x) \text{ for } j = 3, \ldots, n\,, \end{aligned}$$

where

$$\psi_3 = x_1\eta_1 + x_3\eta_3 + \ldots + x_n\eta_n \,, \tag{25}$$

and the controls u, v are subject to the constraint $u^2 + v^2 \leq 1$. Let $x^* : [a,b] \to U$ be a trajectory of Σ, corresponding to the control functions $u(t) \equiv 0$, $v(t) \equiv 1$, and starting at a point $\bar{x} = x^*(a) = (0, \bar{x}_2, 0, \ldots, 0)$ (so that $x^*(t) = (0, \bar{x}_2 + t - a, 0, \ldots, 0)$ for $a \leq t \leq b$). Assume that $\eta_1(x^*(t)) \neq 0$ for $a \leq t \leq b$. Then x^* is locally uniquely time-optimal for Σ.

Remark 11 The important point in Formula (25) is that the right-hand side does not contain a term of the form $x_2\eta_2$. This fact will play a crucial role in the proof of Lemma 2. Consequently, in our proof of the local optimality of regular abnormal extremals, the crucial step will involve changing coordinates to put our equations in the form of Lemma 2, *with ψ_3 of the form (25)*, for which it will be necessary to prove that the expression for ψ_3 does not contain a term of the form $x_2\eta_2$. ∎

PROOF. Let $K = \{x^*(t) : a \leq t \leq b\}$, so K is a compact subset of U. Let $V \subseteq U$ be open, and such that $K \subseteq V$ and the closure $\bar{V}$ of V is compact and contained in U. Then there exists a constant $C_1 > 0$ such that $\|\dot{x}(t)\| \leq C_1$ whenever $x(\cdot)$ is a trajectory of Σ which is contained in V. Let

$$\delta = \min\left\{\|x - y\| \,:\, x \in K, y \in \mathbb{R}^n - V\right\}\,. \tag{26}$$

Then $\delta > 0$. Let $\hat{\tau}_1$ be such that $0 < \hat{\tau}_1 < \frac{\delta}{C_1}$. Then, if $x(\cdot) : [t_1, t_2] \to U$ is a trajectory of Σ that goes through a point of K and is such that $t_2 - t_1 \leq \hat{\tau}_1$, it follows that $x(\cdot)$ is entirely contained in V. We let

$$\begin{aligned} C_2 &= \sup\left\{|\psi_i(x)| \,:\, x \in V\,,\, i = 1, \ldots, n\right\}\,, \\ C_3 &= \sup\left\{\frac{|\eta_i(x)|}{|\eta_1(x^*(t))|} \,:\, a \leq t \leq b\,,\, x \in V, i = 1, 3, \ldots n\right\}\,, \\ \rho &= \inf\left\{|\eta_1(x^*(t))| \,:\, a \leq t \leq b\right\}\,. \end{aligned}$$

Let $\hat{\tau}_2 > 0$ be such that, whenever $x, y \in V$ and $\|x - y\| \leq C_1\hat{\tau}_2$, then it follows that $|\eta_1(x) - \eta_1(y)| \leq \frac{\rho}{4}$. Let

$$\hat{\tau}_3 = \Big(4(n-2)C_2C_3\Big)^{-1}\,, \tag{27}$$

and pick a $\hat{\tau}_4 > 0$ such that $\hat{\tau}_4 < (3C_1C_2)^{-1}$. Let $\hat{\tau} = \min(\hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_3, \hat{\tau}_4)$.

Now let $a \leq t_1 \leq t_2 \leq b$ be such that $\tau = t_2 - t_1 \leq \hat{\tau}$. Let γ be the restriction of x^* to the interval $[t_1, t_2]$. We will show that γ is uniquely time-optimal. Assume that $\xi : [s_1, s_2] \to U$ is another trajectory of Σ that goes from $\gamma(t_1)$ to $\gamma(t_2)$ in time $\sigma = s_2 - s_1$ and corresponds to control functions u_ξ, v_ξ defined on $[s_1, s_2]$ such that $u_\xi(s)^2 + v_\xi(s)^2 \leq 1$ for all $s \in [s_1, s_2]$. We will show that $\sigma \geq \tau$, with equality holding if and only if $\xi(s) = \gamma(s - s_1 + t_1)$ for $s_1 \leq s \leq s_2$.

Let $\xi(s) = (\xi_1(s), \xi_2(s), \xi_3(s), \zeta(s))$, where $\zeta : [s_1, s_2] \to \mathbb{R}^{n-3}$. Let us assume that $\sigma \leq \tau$. Since $\tau \leq \hat{\tau}$, and ξ goes through a point of K, it follows (since $\hat{\tau} \leq \hat{\tau}_1$) that ξ is entirely contained in V, and the bound $\|\dot{\xi}(s)\| \leq C_1$ holds for almost all s. In particular, we have $|\dot{\xi}_1(s)| \leq C_1$ for almost all s.

Let $h(s) = \int_{s_1}^{s} v_\xi(r)dr$. Then $|h(s)| \leq s - s_1$ for all s, so $-\sigma \leq h(s_2) \leq \sigma$. Let $\alpha = \sigma - h(s_2)$, $\beta = \sup\left\{|\xi_1(s)| \,:\, s_1 \leq s \leq s_2\right\}$. We then have

$$\begin{aligned} \tau &= x_2^*(t_2) - x_2^*(t_1) = \xi_2(s_2) - \xi_2(s_1) \\ &= \int_{s_1}^{s_2} \Big(1 + \xi_1(s)\psi_2(\xi(s))\Big)v_\xi(s)ds \end{aligned}$$

$$\leq \int_{s_1}^{s_2} v_\xi(s)ds + \beta\sigma C_2 = h(s_2) + \beta\sigma C_2 = \sigma - \alpha + \beta\sigma C_2 .$$

On the other hand,

$$\int_{s_1}^{s_2} \xi_1(s)^2 ds = \int_{s_1}^{s_2} \xi_1(s)^2(1 - v_\xi(s))ds + \int_{s_1}^{s_2} \xi_1(s)^2 v_\xi(s)ds \ . \tag{28}$$

The first integral of the right-hand side is bounded by $\beta^2\alpha$. The second integral is equal to

$$\int_{s_1}^{s_2} \xi_1(s)^2 v_\xi(s)\frac{\eta_1(\xi(s_1))}{\eta_1(\xi(s_1))}ds \ ,$$

i.e. to

$$\int_{s_1}^{s_2} \xi_1(s)^2 v_\xi(s)\frac{\eta_1(\xi(s_1)) - \eta_1(\xi(s))}{\eta_1(\xi(s_1))}ds + \int_{s_1}^{s_2} \xi_1(s)^2 v_\xi(s)\frac{\eta_1(\xi(s))}{\eta_1(\xi(s_1))}ds \ .$$

But

$$\begin{aligned} 0 &= \frac{\xi_3(s_2) - \xi_3(s_1)}{\eta_1(\xi(s_1))} \\ &= \int_{s_1}^{s_2} \xi_1(s)^2 v_\xi(s)\frac{\eta_1(\xi(s))}{\eta_1(\xi(s_1))}ds + \sum_{i=3}^{n}\int_{s_1}^{s_2} \xi_1(s)\xi_i(s)v_\xi(s)\frac{\eta_i(\xi(s))}{\eta_1(\xi(s_1))}ds \ . \end{aligned}$$

So

$$\begin{aligned} \int_{s_1}^{s_2} \xi_1(s)^2 ds \leq \beta^2\alpha \quad &+ \quad \int_{s_1}^{s_2} \xi_1(s)^2 v_\xi(s)\frac{\eta_1(\xi(s_1)) - \eta_1(\xi(s))}{\eta_1(\xi(s_1))}ds \\ &- \quad \sum_{i=3}^{n}\int_{s_1}^{s_2} \xi_1(s)\xi_i(s)v_\xi(s)\frac{\eta_i(\xi(s))}{\eta_1(\xi(s_1))}ds \ . \end{aligned}$$

For $i \geq 3$, the functions ξ_i satisfy $\dot{\xi_i} = v_\xi\xi_1\psi_i(\xi)$. Therefore, since $\xi_i(s_1) = 0$, we have

$$|\xi_i(s)| \leq C_2\sigma^{\frac{1}{2}}\Big(\int_{s_1}^{s_2} \xi_1(s)^2 ds\Big)^{\frac{1}{2}} \ . \tag{29}$$

Then

$$\left|\int_{s_1}^{s_2} \xi_1(s)\xi_i(s)v_\xi(s)\frac{\eta_i(\xi(s))}{\eta_1(\xi(s_1))}ds\right| \leq C_2C_3\sigma\int_{s_1}^{s_2} \xi_1(s)^2 ds \ . \tag{30}$$

Also,

$$\left|\int_{s_1}^{s_2} \xi_1(s)^2 v_\xi(s) \frac{\eta_1(\xi(s_1)) - \eta_1(\xi(s))}{\eta_1(\xi(s_1))} ds\right| \le \frac{1}{4}\int_{s_1}^{s_2} \xi_1(s)^2 ds \ , \qquad (31)$$

since $|\eta_1(\xi(s_1))| \ge \rho$ and $|\eta_1(\xi(s_1)) - \eta_1(\xi(s))| \le \frac{\rho}{4}$, because $\|\xi(s_1) - \xi(s)\| \le C_1\hat{\tau}_2$ (since $s - s_1 \le \sigma \le \hat{\tau}_2$ and $\|\dot{\xi}\| \le C_1$). So we get the bound

$$\int_{s_1}^{s_2} \xi_1(s)^2 ds \le \beta^2\alpha + \left(\frac{1}{4} + (n-2)C_2C_3\sigma\right)\int_{s_1}^{s_2} \xi_1(s)^2 ds \ . \qquad (32)$$

Since $\sigma \le \hat{\tau}_3 \le \left(4(n-2)C_2C_3\right)^{-1}$, we have the bound $\int_{s_1}^{s_2} \xi_1(s)^2 ds \le \beta^2\alpha + \frac{1}{2}\int_{s_1}^{s_2} \xi_1(s)^2 ds$, from which it follows that $\int_{s_1}^{s_2} \xi_1(s)^2 ds \le 2\beta^2\alpha$. Now let $\tilde{s} \in [s_1, s_2]$ be such that $|\xi_1(\tilde{s})| = \beta$. Since $\xi_1(s_1) = x_1^*(t_1) = 0$, $\xi_1(s_2) = x_1^*(t_2) = 0$, and $|\dot{\xi}_1| \le C_1$, we have $\tilde{s} - s_1 \ge \frac{\beta}{C_1}$ and $s_2 - \tilde{s} \ge \frac{\beta}{C_1}$. Hence the intervals $I_1 = [\tilde{s} - \frac{\beta}{C_1}, \tilde{s}]$ and $I_2 = [\tilde{s}, \tilde{s} + \frac{\beta}{C_1}]$ are entirely contained in $[s_1, s_2]$. On each of these intervals I_j, $|\xi_1(s)|$ is bounded below by the linear function λ_j which is equal to β at $\tilde{s}$ and to zero at the other endpoint. Clearly, the integral of λ_j^2 over I_j is exactly $\frac{\beta^3}{3C_1}$. So $\int_{s_1}^{s_2} \xi_1(s)^2 ds \ge \int_{I_1} \xi_1(s)^2 ds + \int_{I_2} \xi_1(s)^2 ds \ge \frac{2\beta^3}{3C_1}$. Combining the upper and lower bounds for $\int_{s_1}^{s_2} \xi_1(s)^2 ds$ we get $\frac{2\beta^3}{3C_1} \le 2\beta^2\alpha$. Therefore $\beta \le 3C_1\alpha$. But then

$$\tau \le \sigma - \alpha + \beta\sigma C_2 \le \sigma - \alpha + 3C_1C_2\sigma\alpha = \sigma + \left(3C_1C_2\sigma - 1\right)\alpha \ . \quad (33)$$

Since $\sigma \le \tau \le \hat{\tau}_4 < (3C_1C_2)^{-1}$, we have $3C_1C_2\sigma - 1 \le 0$. Therefore $\tau \le \sigma$.

So we have shown that $\sigma \le \tau$ implies $\tau \le \sigma$. Therefore σ cannot be $< \tau$. So γ is optimal, as stated. Moreover, the inequality $\tau \le \sigma$ is in fact strict unless $\alpha = 0$ (since $3C_1C_2\sigma < 1$). So $\sigma = \tau$ can only happen if $\alpha = 0$, i.e. if $v_\xi = 1$ almost everywhere. But then $u_\xi = 0$ almost everywhere, and $\xi(s) = \gamma(s - s_1 + t_1)$. So γ is uniquely optimal. ∎

7.2 The normal form theorem

We now show that a nice abnormal extremal can be put in the form given by Lemma 2. First, we need to introduce some definitions and notations, and prove a lemma on invariant distributions.

Definition 17 Let M be a manifold, and let $\mathcal{D}$ be a smooth distribution on M. Let X be a smooth section of $\mathcal{D}$. If $\gamma : [a,b] \to M$ is a curve in M, we say that $\mathcal{D}$ is *invariant* under X along γ if for every vector field $Y \in \mathrm{Sec}(\mathcal{D})$ the vector $[X,Y](\gamma(t))$ belongs to $\mathcal{D}(\gamma(t))$ for all $t \in [a,b]$. ▮

Whenever $\kappa = (x_1, \ldots, x_n)$ is a cubic chart with domain U, we will automatically identify the domain U with the cube $C^n(\rho)$, so all smooth functions and vector fields on U will be viewed as functions or vector fields on $C^n(\rho)$.

Lemma 3 *Let M be a manifold of dimension $n \geq 3$. Let $\mathcal{D}$ be a smooth distribution of dimension $k < n$ on M. Let $p \in M$, and let Y be a smooth section of $\mathcal{D}$. Assume that $\alpha > 0$, $\gamma : [-\alpha, \alpha] \to M$ is an integral curve of Y passing through p at time 0, and $\mathcal{D}$ is invariant under Y along γ. Then for every $X \in \mathrm{Sec}(\mathcal{D})$ such that $X(p), Y(p)$ are linearly independent there exist a $\rho \in (0, \alpha)$ and a cubic chart $\kappa = (x_1, \ldots, x_n)$ of radius ρ, centered at p, with respect to which*

$$X = \frac{\partial}{\partial x_1} \text{ and } Y = \frac{\partial}{\partial x_2} + x_1 \sum_{k=1}^{n} \psi_k(x) \frac{\partial}{\partial x_k}, \tag{34}$$

where the ψ_k are smooth real-valued functions on $C^n(\rho)$, and ψ_3 is of the form (25). Moreover, if $[X,[X,Y]](p) \notin \mathcal{D}(p)$, then κ can be taken such that $\frac{\partial \psi_3}{\partial x_1}(0) \neq 0$.

Remark 12 If ρ, κ are as in the conclusion of Lemma 3, then it is clear that the curve $\tilde{\gamma} : t \to (0, t, 0, \ldots, 0)$, $t \in (-\rho, \rho)$, is an integral trajectory of Y. So $\tilde{\gamma}$ is the restriction of γ to $(-\rho, \rho)$. In other words:

relative to the chart κ, the restriction of γ to $(-\rho, \rho)$ is contained in the domain of κ and is given by $\gamma(t) = (0, t, 0, \ldots, 0)$. ∎

Before we prove Lemma 3, we briefly review the exponential formalism for flows. If X is a vector field, we use exponential notation for the flow of X, and have the maps e^{tX} act on the right, so $t \to pe^{tX}$ denotes the integral curve of X that passes through p at time $t = 0$. It is also convenient to write the vector fields as acting on points on the right, i.e. to use pX as an alternative notation for $X(p)$. (The reader may just view this as no more than a notational convention. The deep reason for this convention is that vector fields are first-order differential operators on functions, and act on functions on the left. Points are dual objects to functions —Dirac Delta distributions, to be precise— and therefore vector fields should act on points on the right. In addition, having the vector fields and their exponentials act on the right has the extra advantage that Lie-theoretic formulae such as Campbell-Hausdorff come out right if one chooses to define the Lie bracket via $[X, Y] \stackrel{\text{def}}{=} XY - YX$, whereas, if the left action notation were used, consistency would require that the bracket $[X, Y]$ be defined as $YX - XY$.)

If $\theta : M \to N$ is a smooth map, and v is a tangent vector to M at p, then $v\theta$ will denote the tangent vector at $\theta(p)$ which is usually denoted by $\theta_* v$, or $d\theta(p)(v)$. With these notations, for any two smooth vector fields X, Y on a smooth manifold M, and any t such that e^{-tY} is defined, $pe^{-tY} X e^{tY}$ is a tangent vector at p, and the following formula is true:

$$\frac{d}{dt}\Big(pe^{-tY} X e^{tY}\Big) = pe^{-tY}[X, Y]e^{tY} \ . \tag{35}$$

Lemma 4 *Let Y be a smooth vector field on a manifold M. Let $p \in M$ be a point and let J be an interval such that $0 \in J$ and pe^{tY} is defined for all $t \in J$. Let $\gamma : J \to M$ be defined by $\gamma(t) = pe^{tY}$. Let $\mathcal{D}$ be a smooth k-dimensional distribution on M which is invariant under Y along γ. Then, for any pair of points $q_i = pe^{t_i Y}$, $t_i \in J$, $i = 1, 2$, it* *follows that*

$$ve^{(t_1-t_2)Y} \in \mathcal{D}(q_1) \quad \text{whenever} \quad v \in \mathcal{D}(q_2)\ . \tag{36}$$

PROOF. First, pick an arbitrary point $\bar{p} = \gamma(\bar{t})$ on γ, and choose a basis $(X_1, \ldots, X_k)$ of sections of $\mathcal{D}$ on a neighborhood U of $\bar{p}$. Then, since $\mathcal{D}$ is invariant under Y along γ, we can write

$$(\bar{p}e^{-tY})[X_i, Y] = \sum_j \beta_{ij}(t)(\bar{p}e^{-tY})X_j$$

for each t such that $\bar{p}e^{-tY} \in U$, where the β_{ij} are smooth functions. So, if we let $v_i(t) = \bar{p}e^{-tY}X_ie^{tY}$, and let t vary on an interval $I = (-\varepsilon, \varepsilon)$, chosen so that $\bar{p}e^{-tY} \in U$ for $t \in I$, we can conclude that the $T_{\bar{p}}M$-valued vector functions v_i satisfy the differential equations

$$\dot{v}_i(t) = \sum_j \beta_{ij}(t)v_j(t)\,. \tag{37}$$

Since $v_i(0) \in \mathcal{D}(\bar{p})$, we can conclude that $v_i(t) \in \mathcal{D}(\bar{p})$ for all $t \in I$. (Indeed, if λ is any linear functional on $T_{\bar{p}}M$ that vanishes on $\mathcal{D}(\bar{p})$, then the scalar functions $\sigma_i(t) = \langle \lambda, v_i(t) \rangle$ satisfy the linear system $\dot{\sigma}_i = \sum_j \beta_{ij}\sigma_j$. Since the σ_i vanish for $t = 0$, they vanish for all t, and therefore $\langle \lambda, v_i(t) \rangle = 0$ for all $t \in I$.) This shows that $\bar{p}e^{-tY}X_ie^{tY} \in \mathcal{D}(\bar{p})$ for $i = 1, \ldots, k$, $t \in I$. If we let $q = \bar{p}e^{-tY}$, and observe that the vectors qX_i form a basis of $\mathcal{D}(q)$, we have established that $ve^{tY} \in \mathcal{D}(\bar{p})$ whenever $v \in \mathcal{D}(q)$, if $q = \bar{p}e^{-tY}$, and t belongs to an interval $I(\bar{p})$ that may depend on $\bar{p}$. This in particular applies to p. We now show that $ve^{(t_1-t_2)Y} \in \mathcal{D}(\gamma(t_1))$ whenever $v \in \mathcal{D}(\gamma(t_2))$, with no restriction on t other than the requirement that $t_1, t_2 \in J$. To see this, write $t_1 \sim t_2$ if $ve^{(t_1-t_2)Y} \in \mathcal{D}(\gamma(t_1))$ whenever v is in $\mathcal{D}(\gamma(t_2))$, i.e. if the linear transformation $\mu_{t_1,t_2} : T_{\gamma(t_2)}M \to T_{\gamma(t_1)}M$ given by $\mu_{t_1,t_2}(v) = ve^{(t_1-t_2)Y}$ actually maps $\mathcal{D}(\gamma(t_2))$ into $\mathcal{D}(\gamma(t_1))$. It is clear that $t_1 \sim t_2$ implies $t_2 \sim t_1$, since the spaces $\mathcal{D}(\gamma(t_1))$ and $\mathcal{D}(\gamma(t_2))$ have the same dimension. Also, trivially, $t_1 \sim t_1$, and the transitivity property $\Big((t_1 \sim$

$t_2) \wedge (t_2 \sim t_3)\big) \Rightarrow (t_1 \sim t_3)$ holds as well. So $\sim$ is an equivalence relation on J. If C is an equivalence class, and $t \in C$, then we know that there is an open interval L containing 0 such that, if $q = \gamma(t)e^{-sY}$, $s \in L$, then $ve^{sY} \in \mathcal{D}(\gamma(t))$ for $v \in \mathcal{D}(q)$. But this says that $t \sim t - s$ for $s \in L$. So t is an interior point of C. Therefore C is open. Since all the equivalence classes modulo $\sim$ are open, and J is connected, we conclude that J itself is an equivalence class, i.e. that $t_1 \sim t_2$ for all t_1, t_2 in J, which is precisely the desired conclusion. ∎

PROOF OF LEMMA 3. After replacing M with a neighborhood of p, we may assume that $\mathcal{D}$ is spanned by a k-tuple $X_1, X_2, X_4, \ldots, X_{k+1}$ of smooth vector fields, where $X_1 = X, X_2 = Y$, and that there are smooth vector fields $X_3, X_{k+2}, \ldots, X_n$ such that $\{X_k(x), k = 1, \ldots, n\}$ is a basis of T_xM for every $x \in M$. Let Φ be the map defined by

$$\Phi(x_1, \ldots, x_n) = pe^{x_nX_n} \cdots e^{x_1X_1} \ .$$

Since $X_1, \ldots, X_n$ are linearly independent at p, Φ is well defined on a neighborhood of the origin in $\mathbb{R}^n$, and maps some cube $C^n(\rho)$ diffeomorphically onto a neighborhood U of p in M. We choose ρ such that in addition $\rho \leq \alpha$. The inverse map Φ^{-1} defines a chart, by means of which we identify U with $C^n(\rho)$, so that p becomes the point 0. Clearly, X_1 just equals $\frac{\partial}{\partial x_1}$. Moreover, $X_2(x)$ is equal to $\frac{\partial}{\partial x_2}$ at all points x such that $x_1 = 0$. So $X_2 = \frac{\partial}{\partial x_2} + x_1 \sum_{i=1}^n \psi_i \frac{\partial}{\partial x_i}$, where the ψ_i are smooth functions on $C^n(\rho)$. We now replace M by U and γ by its restriction to the interval $(-\rho, \rho)$.

We now have to show that ψ_3 is of the form (25). For this we need only show that $\psi_3 = 0$ when $x_1 = x_3 = x_4 = \cdots = x_n = 0$. Let $\Pi = \{(0, x_2, 0, \ldots, 0) : |x_2| < \rho\}$. Since $[X_1, X_2] = \sum_{i=1}^n (\psi_i + x_1 \frac{\partial \psi_i}{\partial x_1}) \frac{\partial}{\partial x_i}$, it suffices to show that the coefficient of $\frac{\partial}{\partial x_3}$ in the above expression for $[X_1, X_2]$ is identically zero on Π, i.e. that, on Π, $[X_1, X_2]$ is a linear combination of the vector fields $\frac{\partial}{\partial x_i}$, $i \neq 3$. To show this, we first observe that

$$pe^{x_2X_2}\frac{\partial}{\partial x_i} = pX_ie^{x_2X_2} \;\text{ for }\; i \geq 2 \; . \tag{38}$$

Let $I = \{1, 2, 4, \ldots, k+1\}$. Then, if $i \in I$, the vector field X_i is a section of $\mathcal{D}$. Assume now that $i \in I$ and $i \geq 2$. It is clear that the integral curve of X_2 passing through p in Ω is exactly the line segment Π in $C^n(\rho)$, i.e. the curve γ. Therefore $\mathcal{D}$ is invariant under X_2 along Π. We conclude that $qe^{-x_2X_2}X_ie^{x_2X_2}$ belongs to $\mathcal{D}(q)$ for every $q \in \Pi$, $i \in I$. Letting $q = pe^{x_2X_2}$, $i > 1$, we find that $pX_ie^{x_2X_2} \in \mathcal{D}(pe^{x_2X_2})$, i.e. that $\frac{\partial}{\partial x_i}(q)$ is in $\mathcal{D}(q)$ for all $q \in \Pi$, $i \in I$, $i \geq 2$. The conclusion is also true for $i = 1$, since $X_1 = \frac{\partial}{\partial x_1}$. So the k vectors $\frac{\partial}{\partial x_i}(q)$, $i \in I$, belong to $\mathcal{D}(q)$ for all $q \in \Pi$. These k vectors are obviously linearly independent, and therefore form a basis of $\mathcal{D}(q)$, since $\mathcal{D}$ is k-dimensional. Since X_1 is a section of $\mathcal{D}$, and $\mathcal{D}$ is X_2-invariant along Π, the vector $[X_1, X_2](q)$ belongs to $\mathcal{D}(q)$ for $q \in \Pi$. This implies in particular that $[X_1, X_2](q)$ is a linear combination of the $\frac{\partial}{\partial x_i}(q)$ for $i \in I$. Since $3 \notin I$, this establishes the desired conclusion about $[X_1, X_2]$.

Finally, we remark that, if $[X, [X, Y]](p) \notin \mathcal{D}(p)$, then we can choose X_3 to be $[X, [X, Y]]$. With this choice,

$$X_3 = [X_1, [X_1, X_2]] = \sum_{i=1}^{n}\Big(2\frac{\partial \psi_i}{\partial x_1} + x_1\frac{\partial^2 \psi_i}{\partial x_1^2}\Big)\frac{\partial}{\partial x_i} \; . \tag{39}$$

Since $X_3(p) = \frac{\partial}{\partial x_3}(0)$, we conclude that $\frac{\partial \psi_3}{\partial x_1}(0) = \frac{1}{2} \neq 0$. ∎

We are now ready to state the Normal Form Theorem. Recall that $\Sigma(M, E, G)$ is the set of all E-arcs γ in M such that $\dot\gamma(t)$ has G-length ≤ 1 for almost all t in the domain of γ. If U is an open subset of M, we can restrict E to U, obtaining a two-dimensional distribution $E\lceil U$ on U, and then we can define the restriction $G\lceil(E\lceil U)$ of G to $E\lceil U$. We will simply write $\Sigma(U, E, G)$ rather than $\Sigma(U, E\lceil U, G\lceil(E\lceil U))$. If U is such that there is an orthonormal basis (X, Y) of sections of $E\lceil U$, then $\Sigma(U, E, G)$ is exactly the set of trajectories of the control system $\dot x = uX(x) + vY(x)\,,\; u^2 + v^2 \leq 1$.

Theorem 4 *Let (M, E, G) be a sub-Riemannian manifold such that* $\dim M \geq 3$ *and* $\dim E = 2$. *Let $\gamma : [a, b] \to M$ be a regular abnormal extremal. Then for every $\bar{t} \in [a, b]$ there exist a $\rho > 0$, an open neighborhood V of $\gamma(\bar{t})$, a cubic coordinate chart κ of radius ρ, centered at $\gamma(\bar{t})$, with domain V, and an orthonormal basis (X, Y) of sections of $E\lceil V$, such that*

- *(i) X and Y are given, relative to the chart κ, by (34), where the ψ_k are smooth functions such that ψ_3 is of the form (25) and $\frac{\partial \psi_3}{\partial x_1}(0) \neq 0$.*
- *(ii) there is an $\varepsilon > 0$ such that $\gamma([\bar{t} - \varepsilon, \bar{t} + \varepsilon] \cap [a, b]) \subseteq V$ and $\gamma(s) = (0, s - \bar{t}, 0, \ldots, 0)$ for $s \in [\bar{t} - \varepsilon, \bar{t} + \varepsilon] \cap [a, b]$.*

Conversely, if γ is an arc contained in the domain V of a cubic chart κ, (X, Y) is an orthonormal basis of sections of E on V such that (i) holds, and γ is of the form $s \to (0, s, 0, \ldots, 0)$, then (a) γ is an abnormal extremal of E, and (b) γ is strictly abnormal if and only if the function ψ_2 cannot be expressed along γ as a linear combination of $\psi_4, \ldots, \psi_n$ with constant coefficients.

PROOF. If γ is the projection of a regular abnormal biextremal $\Gamma : [a, b] \to RA(E)$, then we can extend Γ to a regular abnormal biextremal Γ_{ext} defined on a larger interval $[a - \delta, b + \delta]$, and let $\gamma_{ext} = \pi_M^* \circ \Gamma_{ext}$, so γ_{ext} is a regular abnormal extremal defined on $[a - \delta, b + \delta]$. We may then replace γ by γ_{ext}. This means that we can assume that $\bar{t}$ is an interior point of $[a, b]$. Also, without loss of generality, we may assume that $\bar{t} = 0$. We pick $\alpha > 0$ such that $a \leq -\alpha$, $\alpha \leq b$, and replace γ by its restriction to $[-\alpha, \alpha]$. In addition, α will also be required to satisfy two more conditions —labeled (C1), (C2)— to be stated later. Let $p = \gamma(0)$.

In view of Theorem 3, we may assume that

(C1) γ is a nice abnormal extremal.

Let U, Y, ω be such that the four conditions of Definition 16 are satisfied. We may also assume that

(C2) $\gamma([-\alpha,\alpha]) \subseteq U$,

and that there is a smooth section X of E such that (X,Y) is an orthonormal basis of $\mathrm{Sec}(E,U)$.

Let $\mathcal{D} = \ker\omega$. Since ω is nowhere vanishing on U, $\mathcal{D}$ is a distribution of dimension $n-1$ on U. Since $L_Y\omega \equiv 0$ along γ, we have

$$\omega([Y,Z]) = Y\Big(\omega(Z)\Big) - (L_Y\omega)(Z) = Y\Big(\omega(Z)\Big) \tag{40}$$

along γ for any smooth vector field Z on U. Applying (40) to an arbitrary section $Z \in \mathrm{Sec}(\mathcal{D})$ —in which case $\omega(Z)$ vanishes along γ— we conclude that the distribution $\mathcal{D}$ is invariant under Y along γ. Applying (40) to $Z = X$, we find that $\omega([X,Y])$ vanishes along γ, so $[X,Y](q) \in \mathcal{D}(q)$ for every $q \in \gamma([a,b])$. Applying (40) to $Z = [X,Y]$ we conclude that $\omega([Y,[X,Y]]$ also vanishes along γ. Since $\omega(q)$ does not annihilate $E^3(q)$ for any $q \in U$, it follows that $\omega([X,[X,Y]])(\gamma(t)) \neq 0$ for all t. Therefore $[X,[X,Y]](q) \notin \mathcal{D}(q)$ for $q \in \gamma([a,b])$, and we may assume —after shrinking U and α, if necessary— that $[X,[X,Y]](q) \notin \mathcal{D}(q)$ for all $q \in U$.

We have therefore shown that $\mathcal{D}$ has almost all the properties of Lemma 3. The only remaining problem is that X, Y need not be sections of $\mathcal{D}$. To take care of this, we let $(Z_1,\ldots,Z_{n-1})$ be a global basis of sections of $\mathcal{D}$, chosen in such a way that $Z_1 = X$ on γ and $Z_2 = Y$ on γ. (This may require shrinking U and α further.) Since $[X,[X,Y]](q) \notin \mathcal{D}(q)$ for $q \in U$, the vectors $Z_1(q),\ldots,Z_{n-1}(q),[X,[X,Y]](q)$ span T_qM for all $q \in U$. So we can write

$$X = \sum_{i=1}^{n-1}\alpha_i Z_i + \alpha[X,[X,Y]] \text{ and } Y = \sum_{i=1}^{n-1}\beta_i Z_i + \beta[X,[X,Y]]\,, \tag{41}$$

where $\alpha_i, \beta_i, \alpha, \beta$ are smooth functions. Then, since $X(q)$ and $Y(q)$ belong to $\mathcal{D}(q)$ for $q \in \gamma([a,b])$ but $[X,[X,Y]](q) \notin \mathcal{D}(q)$, we have

$\alpha(\gamma(t)) = \beta(\gamma(t)) = 0$ for all $t \in [a,b]$. Since $X = Z_1$ and $Y = Z_2$ on γ, we may assume (after shrinking U further, if necessary) that $X, Y, Z_3, \ldots, Z_{n-1}$ are linearly independent everywhere. Therefore if we let $\bar{\mathcal{D}}$ be the distribution on U spanned by $X, Y, Z_3, \ldots, Z_{n-1}$, then $\bar{\mathcal{D}}$ is a distribution of dimension $n-1$. Clearly, $\bar{\mathcal{D}}(q) = \mathcal{D}(q)$ for $q \in \gamma([a,b])$. This implies, in particular, that $[X,[X,Y]](q) \notin \bar{\mathcal{D}}(q)$ for $q \in \gamma([a,b])$.

We now show that $\bar{\mathcal{D}}$ is invariant under Y along γ. If Z is an arbitrary smooth section of $\bar{\mathcal{D}}$, then we can write

$$Z = \nu_1 X + \nu_2 Y + \sum_{k=3}^{n-1} \nu_k Z_k \,.$$

But then

$$[Y,Z] = \nu_1[Y,X] + \sum_{k=3}^{n-1} \nu_k [Y,Z_k] + (Y\nu_1)X + (Y\nu_2)Y + \sum_{k=3}^{n-1} (Y\nu_k) Z_k \,.$$

Since $\mathcal{D}$ is Y-invariant along γ, the vectors $[Y,Z_k](q)$ belong to $\mathcal{D}(q)$ for every $q \in \gamma([a,b])$. The $Z_k(q)$ also belong to $\mathcal{D}(q)$, and we know that $X(q)$, $Y(q)$ and $[Y,X](q)$ are in $\mathcal{D}(q)$ as well. So $[Y,Z](q) \in \mathcal{D}(q) = \bar{\mathcal{D}}(q)$ for $q \in \gamma([a,b])$, and $\bar{\mathcal{D}}$ is invariant under Y along γ.

We can now apply Lemma 3, and choose a cubic coordinate chart $(x_1, \ldots, x_n)$ of radius ρ with respect to which (i) holds. The existence of ε for which (ii) holds then follows from Remark 12.

We now prove the converse. The Hamiltonian of the control system $\dot{x} = uX(x) + vY(x)$ is

$$\begin{aligned} H(x,\lambda,u,v) &= H(x_1,\ldots,x_n,\lambda_1,\ldots,\lambda_n,u,v) \\ &= \lambda_1 u + \lambda_2 v + v x_1 \sum_{j=1}^{n} \lambda_j \psi_j(x) \,. \end{aligned}$$

The adjoint equations along γ say $\dot{\lambda}_1 = -\sum_{j=1}^{n} \lambda_j \psi_j(\gamma(t))$, $\dot{\lambda}_k = 0$ for $k = 2, \ldots, n$. Since ψ_3 vanishes along γ, these equations are satisfied by $\lambda_1 = \lambda_2 = 0$, $\lambda_3 = 1$, $\lambda_4 = \ldots = \lambda_n = 0$. With this choice of the λ_i, $H(\gamma(t),\lambda,u,v)$ vanishes for all (u,v). So γ is an abnormal extremal.

We now seek conditions that will guarantee that γ is not a normal extremal. For γ to be a normal extremal, there would have to be a solution $t \to \lambda(t)$ of the adjoint equations with the property that for each t the function $(u, v) \to H(\gamma(t), \lambda(t), u, v)$ is minimized on the unit disc by $u = 0$, $v = 1$, and the minimum value is -1. Since $H(\gamma(t), \lambda(t), u, v) = \lambda_1 u + \lambda_2 v$, this will happen if and only if $\lambda_1 = 0$ and $\lambda_2 = -1$. But then the first adjoint equation says that $\psi_2 = \sum_{i=4}^n \lambda_i \psi_i$. So γ is a normal extremal if and only if there are constants $\lambda_4, \ldots, \lambda_n$ such that $\psi_2 = \sum_{i=4}^n \lambda_i \psi_i$ along γ. ∎

7.3 The optimality theorem

We are now ready to prove:

Theorem 5 *If (M, E, G) is a sub-Riemannian manifold such that $\dim M \geq 3$ and E is a two-dimensional distribution, then every regular abnormal extremal is locally uniquely time optimal.*

PROOF. We will show that the interval $[a, b]$ can be covered by open intervals I_ν such that the restriction of γ to every closed subinterval of an I_ν is uniquely optimal. Once this is proved, the local optimality of γ follows by letting $\delta > 0$ be a Lebesgue number of the covering $\{I_\nu\}$. If $a \leq t_1 < t_2 \leq b$, and $t_2 - t_1 \leq \delta$, then the interval $[t_1, t_2]$ is contained in one of the I_ν, and therefore the restriction of γ to $[t_1, t_2]$ is uniquely optimal.

To prove the existence of the I_ν, it suffices to pick a $\bar{t} \in [a, b]$ and find a $\delta > 0$ such that the restriction of γ to the interval $[a, b] \cap [\bar{t} - \delta, \bar{t} + \delta]$ is uniquely time-optimal. Theorem 4 says that there exist a neighborhood V of $\gamma(\bar{t})$, a chart κ on V centered at $\gamma(\bar{t})$, and an $\varepsilon > 0$, such that (i) and (ii) of Theorem 4 hold. This means that we can apply Lemma 2, and conclude that the restriction of γ to some neighborhood of $\bar{t}$ is uniquely optimal. ∎

8 Strict abnormality

We now discuss the extra conditions that are needed for a regular abnormal extremal to be strictly abnormal, and show in particular that most regular abnormal extremals are in fact strictly abnormal. In Theorem 4 we have already characterized strict abnormality for abnormal extremals in normal form. An alternative characterization can be obtained in terms of conditions on vector fields, as we now explain.

Use ad_g to denote the map $X \to [g, X]$ on the space of smooth vector fields. Let (f, g) be an orthonormal basis of sections of E on an open set U and let $\gamma : [a, b] \to U$ be an integral curve of g. We make the following regularity assumption

(REG) *there exists an integer $k > 0$ such that, along γ, the vector field $\mathrm{ad}_g^{k+1}(f)$ is a linear combination with smooth coefficients of g and the $\mathrm{ad}_g^j(f)$ for $0 \le j \le k$.*

Write

$$\mathrm{ad}_g^{k+1}(f)(\gamma(t)) = \sum_{j=0}^{k} \alpha_j(t)\mathrm{ad}_g^j(f)(\gamma(t)) + \beta(t)g(\gamma(t)), \qquad (42)$$

where β and the α_j are smooth functions of t.

Let λ be a field of covectors along γ such that $\Gamma = (\gamma, \lambda)$ satisfies $\dot{\Gamma}(t) = \overrightarrow{H_g}(\Gamma(t))$. Then the derivative along Γ of any function of the form H_X, for X a smooth vector field on U, is the function $H_{[g,X]}$. Let $\varphi_j(t) = H_{\mathrm{ad}_g^j(f)}(\Gamma(t))$, and write $\psi(t) = H_g(\Gamma(t))$. Then the φ_j satisfy $\dot{\varphi}_j = \varphi_{j+1}$, while $\dot{\psi} = 0$. Since $\varphi_{k+1} = \alpha_0\varphi_0 + \ldots + \alpha_k\varphi_k + \beta\psi$, we see that the vector $(\varphi_0, \ldots, \varphi_k, \psi)$ satisfies a linear homogeneous system of differential equations. Therefore this vector vanishes identically if it vanishes at one point. So, if for some $t \in [a, b]$ there exists a nonzero $\bar{\lambda} \in T^*_{\gamma(t)}M$ such that $\langle \bar{\lambda}, g(\gamma(t)) \rangle = 0$ and $\langle \bar{\lambda}, \mathrm{ad}_g^j(f)(\gamma(t)) \rangle = 0$ for $j = 0, \ldots, k$, then the integral curve Γ of $\overrightarrow{H_g}$ that goes through $(\gamma(t), \bar{\lambda})$

at time t is an abnormal biextremal, and it is clear that the projection of Γ is γ.

We now investigate whether one can also find a normal biextremal Γ that projects onto γ. The arc Γ would have to be an integral curve of $\overrightarrow{H}$, where H is the Hamiltonian of (M, E, G), i.e. the function $(x, \lambda) \to -\frac{1}{2}\Big(\langle\lambda, f(x)\rangle^2 + \langle\lambda, g(x)\rangle^2\Big)$. Since $\overrightarrow{H} = -H_f\overrightarrow{H_f} - H_g\overrightarrow{H_g}$, $d\pi_M^*(\dot{\Gamma}(t)) = \dot{\gamma}(t)$, $d\pi_M^*(\overrightarrow{H_X}) = X$ for every X, and $\dot{\gamma}(t) = g(\gamma(t))$, we see that $H_f \equiv 0$ and $H_g \equiv -1$ along Γ, so Γ must be an integral curve of $\overrightarrow{H_g}$. As before, the functions φ_j satisfy $\dot{\varphi}_j = \varphi_{j+1}$, while ψ is now equal to -1. Since $\varphi_0 \equiv 0$, we see that $\varphi_j \equiv 0$ for all j. In particular, $\varphi_{k+1} \equiv 0$. Since $\varphi_{k+1} = \alpha_0\varphi_0 + \ldots + \alpha_k\varphi_k + \beta\psi$, we see that $\beta\psi \equiv 0$. Since $\psi \equiv -1$, we conclude that $\beta \equiv 0$. So γ is strictly abnormal if $\beta(t) \neq 0$ for some t.

Summarizing, we have proved:

Theorem 6 *Let (M, E, G) be a sub-Riemannian manifold such that E is two-dimensional. Let (f, g) be an orthonormal basis of sections of E on an open subset U of M, and let $\gamma : I \to U$ be an integral curve of g. Then: (i) if the regularity condition (REG) holds, and in addition for some $t \in I$ the linear span of $g(\gamma(t))$ and the vectors $\mathrm{ad}_g^j(f)(\gamma(t))$ for $0 \leq j \leq k$ is not the whole tangent space $T_{\gamma(t)}M$, then γ is an abnormal extremal; (ii) if the coefficient β of (42) does not vanish identically, then γ is strictly abnormal.* ∎

9 Some special cases

We now make a more detailed analysis of the abnormal extremals in some special cases. Specifically, we will consider regular distributions, low-dimensional manifolds, and a Lie group example. For the cases when $\dim M = 4$ or 5 we will assume that E is a regular distribution. In the three-dimensional case, if E is regular, then the only abnormal extremals are the constant trajectories, so the only interesting case is when E is nonregular at some points. We will assume that the singularities of E occur in a generic way. (The singularities of generic two-dimensional distributions in three-dimensional manifolds are well understood, cf. e.g. [29].)

9.1 2- and 3-generating distributions

The following is a direct consequence of the results of the previous sections:

Theorem 7 *Let (M, E, G) be a sub-Riemannian manifold such that E is two-dimensional. Then: (i) if E is 1-generating (in which case, of course, $\dim M = 2$), then (M, E, G) is in fact Riemannian, all extremals are normal, and the extremals are exactly the ordinary Riemannian geodesics; (ii) if E is 2-generating (in which case $\dim M \leq 3$ and E is strongly bracket-generating), then there are no nonconstant abnormal extremals, and all the minimizers are normal extremals; (iii) if E is 3-generating (in which case $\dim M \leq 5$), then $RA(E) = (E^2)^\perp \cap T^\# M$, and all locally simple abnormal extremals parametrized by arc-length are regular. In all three cases, all locally simple extremals parametrized by arc-length are smooth and locally optimal.* ∎

9.2 3-Regular distributions

Let E be a smooth $\bar{k}$-regular distribution on a manifold M. For each $p \in M$, and each integer k such that $1 \leq k \leq \bar{k}$, define $\mathcal{L}_E^k(p) = E^k(p)/E^{k-1}(p)$. We can then define bilinear maps $\mathcal{Q}_E^{k,j}(p) : \mathcal{L}_E^k \times \mathcal{L}_E^j \to \mathcal{L}_E^{k+j}$ by letting $\mathcal{Q}_E^{k,j}(p)(v, w) = z$ if $f \in \mathrm{Sec}^k(E)$, $g \in \mathrm{Sec}^j(E)$, $p \in \mathrm{Domain}(f) \cap \mathrm{Domain}(g)$, $v = f(p) \bmod E^{k-1}(p)$, $v = g(p) \bmod E^{j-1}(p)$, and $z = [f, g](p) \bmod E^{k+j-1}(p)$. (It is easy to see that z does not depend on the choice of f and g. Indeed, let $r = \dim E^{k-1}(p)$, $s = \dim E^k(p)$. Let $(Z_1, \ldots, Z_n)$ be a basis of sections of TM, chosen so that $(Z_1,\ldots,Z_r)$ is a basis of sections of E^{k-1} and $(Z_1,\ldots,Z_s)$ is a basis of sections of E^k. If $f_1,f_2\in \mathrm{Sec}^k(E)$, $f_1(p)-f_2(p)\in \mathrm{Sec}^{k-1}(E)$, and $g\in \mathrm{Sec}^j(E)$, then we can write $f_i=\sum_{\ell=1}^s \alpha_\ell^i Z_\ell$ near p, where the α_ℓ^i are smooth functions such that $\alpha_\ell^1(p)=\alpha_\ell^2(p)$ for $r<\ell\le s$. Then $[f_1-f_2,g]=\sum_{\ell=1}^s(\alpha_\ell^1-\alpha_\ell^2)[Z_\ell,g]-\sum_{\ell=1}^s g(\alpha_\ell^1-\alpha_\ell^2)Z_\ell$. Therefore $[f_1-f_2,g](p)=\sum_{\ell=1}^r(\alpha_\ell^1(p)-\alpha_\ell^2(p))[Z_\ell,g](p)-\sum_{\ell=1}^n(g\alpha_\ell^1(p)-g\alpha_\ell^2(p))Z_\ell(p)$. The brackets $[Z_\ell,g]$ are sections of E^{k+j-1} for $\ell\le r$, and the Z_ℓ are sections of E^{k+j-1} for $\ell\le s$. So $[f_1-f_2,g](p)\in E^{k+j-1}(p)$. This shows that the value of $[f,g](p)$ modulo $E^{k+j-1}(p)$ does not depend on f but only on the value of $f(p)$ modulo $E^{k-1}(p)$. A similar argument shows that $[f,g](p)$ modulo $E^{k+j-1}(p)$ only depends the the value of $g(p)$ modulo $E^{j-1}(p)$.)

Now suppose that (M,E,G) is a sub-Riemannian manifold such that $\dim M\ge 3$, and the distribution E is two-dimensional and 3-regular. Then the space $\mathcal{L}_E^2(p)$ is one-dimensional for each p, so $\mathcal{L}_E^2$ is a line bundle. The maps $\mathcal{Q}_E^{1,1}(p):E(p)\times E(p)\to\mathcal{L}_E^2(p)$ are skew-symmetric and nonzero. So each $\mathcal{Q}_E^{1,1}(p)$ induces a map $\tilde{\mathcal{Q}}_E^{1,1}(p)$ from the exterior power $\wedge^2E(p)$ to $\mathcal{L}_E^2(p)$, which is obviously an isomorphism. Clearly, $\wedge^2E(p)$ has a natural inner product induced by the inner product $G(p)$ on $E(p)$, so this inner product is also well defined on $\mathcal{L}_E^2(p)$. Therefore $\mathcal{L}_E^2(p)$ has two distinguished elements, namely, the vectors of norm one, corresponding to the two orientations of $E(p)$. Furthermore, once we have a distinguished element w of $\mathcal{L}_E^2(p)$, then we can use it to define a map $v\to\mathcal{Q}_E^{2,1}(p)(v,w)$ from $E(p)$ to $\mathcal{L}_E^3(p)$. If (f,g) is an orthonormal basis of sections of E near p, then the class of $[f,g](p)$ modulo $E(p)$ is the corresponding distinguished element of $\mathcal{L}_E^2(p)$. If $(\tilde f,\tilde g)$ is another orthonormal basis, then $[\tilde f,\tilde g](p)=\varepsilon[f,g](p)$ modulo $E(p)$, where $\varepsilon=\pm1$ depending on whether or not the orientations of the two bases are the same. The map from $E(p)$ to $\mathcal{L}_E^3(p)$ corresponding to (f,g) —which in fact does not depend on the choice of the orthonormal basis (f,g), but only on the orientation of the pair

$(f(p), g(p))$— is the one that sends each vector $v = af(p) + bg(p)$ to the vector $a[f,[f,g]](p) + b[g,[f,g]](p)$. If ω is an orientation of $E(p)$, we will use $\theta_E(\omega, p)$ to denote the map from $E(p)$ to $\mathcal{L}_E^3(p)$ corresponding to ω. Notice that the maps $\theta_E(\omega, p)$ are surjective, and satisfy $\theta_E(\omega_1, p) = -\theta_E(\omega_2, p)$ if ω_1, ω_2 are opposite orientations. It is clear that on any open set U on which E has a global orientation ω, this orientation can be used to define a smooth bundle map from $E\lceil U$ to $\mathcal{L}_E^3\lceil U$ which is surjective on each fiber.

We now study how the maps $\theta_E(\omega, p)$ relate to the abnormal extremals. We consider separately the two cases that arise depending on whether E^3 is of dimension 4 or 5.

Assume first that E^3 is four-dimensional, so E is a 3-regular distribution of partial type $(2, 1, 1)$. (For this it of course required that $\dim M \geq 4$.) Then at each point p the kernel of the map $\theta_E(\omega, p)$ is one-dimensional, and independent of the orientation ω. So this kernel defines a line subbundle $\tilde{L}_E$. If (f, g) is an orthonormal basis of sections of E near p, then the fiber $\tilde{L}_E(p)$ is the set of all vectors of the form $af(p) + bg(p)$ such that $a[f,[f,g]](p) + b[g,[f,g]](p) \in E^2(p)$.

Let $\xi = (p, \lambda)$ be a regular abnormal point in T^*M. Let (f, g) be a basis of sections of E in a neighborhood U of p. Then we know from Proposition 7 that $L_E(\xi)$ is spanned by the vector $\mathcal{X}_E(\xi) = a\overrightarrow{H_f}(\xi) - b\overrightarrow{H_g}(\xi)$, where $a = H_{[g,[f,g]]}(\xi), b = H_{[f,[f,g]]}(\xi)$. Since E is of type $(2, 1, 1, d_1, \ldots, d_k)$, the vector fields $[f,[f,g]]$ and $[g,[f,g]]$ are linearly dependent on U modulo E^2. Let α, β be smooth functions on U such that $\alpha[f,[f,g]] + \beta[g,[f,g]] \equiv 0$ modulo E^2 and $\alpha(p)^2 + \beta(p)^2 \neq 0$ for all $p \in U$. Then the vector (b, a) is orthogonal to $(\alpha(p), \beta(p))$, and therefore $(a, -b)$ is a multiple of $(\alpha(p), \beta(p)$. Therefore $L_E(\xi)$ is spanned by $\alpha(p)\overrightarrow{H_f}(\xi) + \beta(p)\overrightarrow{H_g}(\xi)$. The image of $L_E(\xi)$ under $d\pi_M^*(\xi)$ is the one-dimensional subspace of $E(p)$ spanned by $\alpha(p)f(p) + \beta(p)g(p)$, which is precisely the space $\tilde{L}_E(p)$ defined above, and is independent of λ. So $\tilde{L}_E(p) = d\pi_M^*(p, \lambda)(L_E(p, \lambda))$ for any $(p, \lambda) \in RA(E)$. (Notice, however, that $d\pi_M^*(\mathcal{X}_E(p,-\lambda)) = -d\pi_M^*(\mathcal{X}_E(p,\lambda))$. So the canonical orientation of L_E does not induce a natural orientation of $\tilde{L}_E$.) It is clear that M is foliated by the leaves of $\tilde{L}_E$ and every regular abnormal extremal is contained in a leaf. Conversely, if $\gamma : [a,b] \to M$ is an arc which is contained in a leaf of $\tilde{L}_E$, we claim that γ *is a projection of an abnormal biextremal contained in* $RA(E)$. Indeed, we can partition the interval $[a,b]$ into intervals $I_i = [t_{i-1}, t_i]$ such that each $\gamma(I_i)$ is contained in an open set U_i which is the domain of an orthonormal basis (f_i, g_i) of sections of E and of a coordinate chart κ_i. It is clear that we can choose the basis (f_i, g_i) in such a way that f_i is a section of $\tilde{L}_E$ and $[f_i,[f_i,g_i]]$ is a section of E^2. If $\xi = (p,\lambda)$ is a regular abnormal point such that $p \in U_i$, then we know that $L_E(\xi)$ is spanned by $\overrightarrow{H_{f_i}}(\xi)$. So the vector field $\overrightarrow{H_{f_i}}$, which is smooth on T^*U_i, is tangent to $(E^2)^\perp$ on the relatively open dense subset $RA(E)$, and therefore is tangent to $(E^2)^\perp$ everywhere. Moreover, $\overrightarrow{H_{f_i}}$ is tangent to $(E^3)^\perp$ as well. (To see this, observe that, since $[f_i,[f_i,g_i]]$ is a section of E^2, we can write

$$[f_i,[f_i,g_i]] = \alpha_1 f_i + \alpha_2 g_i + \alpha_3 [f_i,g_i] \tag{43}$$

on U_i, where the α_j are smooth functions. But then

$$\begin{aligned} &[f_i,[g_i,[f_i,g_i]]] = [g_i,[f_i,[f_i,g_i]]] \\ &= (g_i\alpha_1)f_i + (g_i\alpha_2)g_i + \Big((g_i\alpha_3) - \alpha_1\Big)[f_i,g_i] + \alpha_3[g_i,[f_i,g_i]]\,, \end{aligned} \tag{44}$$

so $[f_i,[g_i,[f_i,g_i]]]$ is a section of E^3. On the other hand, $(E^3)^\perp$ is defined by the four equations $h_1 = h_2 = h_3 = h_4 = 0$, where $h_1 = H_{f_i}$, $h_2 = H_{g_i}$, $h_3 = H_{[f_i,g_i]}$, and $h_4 = H_{[g_i,[f_i,g_i]]}$. At a point $\xi \in (E^3)^\perp$, the derivatives in the direction of $\overrightarrow{H_{f_i}}$ of the four functions h_1, h_2, h_3, h_4 are, respectively, 0, $h_3(\xi)$, $h_5(\xi)$ and $h_6(\xi)$, where $h_5 \equiv H_{[f_i,[f_i,g_i]]}$ and $h_6 \equiv H_{[f_i,[g_i,[f_i,g_i]]]}$. Now (43) and (44) imply that h_5 is a linear combination of h_1, h_2 and h_3 with smooth coefficients, so h_5 vanishes on $(E^2)^\perp$, and h_6 is a linear combination of h_1, h_2, h_3 and h_4 with smooth coefficients, so h_6 vanishes on $(E^3)^\perp$. Therefore the derivatives of h_1, h_2, h_3 and h_4 in the direction of $\overrightarrow{H_{f_i}}$ vanish on $(E^3)^\perp$, and $\overrightarrow{H_{f_i}}$ is tangent to $(E^3)^\perp$.) The arc γ satisfies, on I_i, the equation $\dot\gamma(t) = u_i(t)f_i(\gamma(t))$, where the function u_i is integrable. Consider the differential equation

$$\dot\lambda(t) = -u_i(t)\lambda(t)\frac{\partial f_i}{\partial x}(\gamma(t)), \tag{45}$$

relative to the chart κ_i. It is clear that if $\lambda(\cdot)$ is any solution of (45) on I_i such that $\lambda(t) \in E^2(p)^\perp$ but $\lambda(t) \notin E^3(\gamma(t))^\perp$ for any $t \in I_i$, then the pair $(\gamma\lceil I_i, \lambda)$ is an abnormal biextremal contained in $RA(E)$. Moreover, for any given initial condition $\lambda(t_{i-1}) = \bar\lambda$ there exists a solution of (45) on the whole interval I_i, since the right-hand side of (45) is linear in λ. If λ is any solution of (45), and we let $\Gamma = (\gamma\lceil I_i, \lambda)$, then Γ is a solution of $\dot\Gamma(t) = \overrightarrow{H_{f_i}}(\Gamma(t))$. Since $\overrightarrow{H_{f_i}}$ is tangent to $(E^2)^\perp$, which is a closed submanifold of T^*M, the curve Γ has to be entirely contained in $(E^2)^\perp$ if $\bar\lambda \in E^2(\gamma(t_{i-1}))^\perp$. Similarly, $\overrightarrow{H_{f_i}}$ is tangent to $(E^3)^\perp$, which is also a closed submanifold of T^*M, the curve Γ is either entirely contained in $(E^3)^\perp$ or disjoint from it. In particular, if $\bar\lambda \notin E^3(\gamma(t_{i-1}))^\perp$, then Γ is entirely contained in $RA(E)$. It then follows easily that γ is a projection of an abnormal biextremal contained in $RA(E)$. Indeed, we can fix a $\hat\lambda_0 \in E^2(\gamma(a))^\perp - E^3(\gamma(a))^\perp$, and solve (45) on I_1, obtaining an abnormal biextremal Γ_1 which is contained in $RA(E)$ and projects onto the restriction of γ to I_1. Then, using $\lambda(\gamma(t_1))$ as the new initial condition, we can solve (45) on I_2, obtaining an abnormal biextremal Γ_2 which is contained in $RA(E)$, projects onto the restriction of γ to I_2, and is such that $\Gamma_2(t_1) = \Gamma_1(t_1)$. This procedure can clearly be continued until we construct an abnormal biextremal $\Gamma : [a,b] \to RA(E)$ that projects onto γ.

Finally, we observe that our proof yields extra information. In fact, we have established that an abnormal biextremal that is not regular has to be entirely contained in $(E^3)^\perp$. In other words, there cannot exist abnormal biextremals that are regular on part of their interval of definition but go through $(E^3)^\perp$ at some time. Let us call a curve in M a *totally singular abnormal extremal* if it is the projection of an abnormal biextremal which is entirely contained in $(E^3)^\perp$. Then we have proved the following

Proposition 11 *Let E be a two-dimensional 3-regular distribution of partial type $(2,1,1)$ on a manifold M. Then there is a line subbundle $\tilde{L}_E$ of E such that, if U is any open subset of M on which E admits a basis (f,g) of sections, and α, β are smooth functions on U such that $\alpha[f,[f,g]] + \beta[g,[f,g]] \equiv 0$ modulo E^2 and $\alpha(p)^2 + \beta(p)^2 \neq 0$ for all $p \in U$, then the vector field $Z = \alpha f + \beta g$ is a section of $\tilde{L}_E$ on U. The regular abnormal extremals of E are exactly the arcs parametrized by arc-length that are contained in a leaf of the foliation induced by $\tilde{L}_E$. Every locally simple abnormal extremal parametrized by arc-length is either regular or totally singular.* ∎

We now consider the case when E^3 is five-dimensional, so the maps $\theta(\omega,p)$ are isomorphisms from $E(p)$ to $\mathcal{L}_E^3(p) = E^3(p)/E^2(p)$. There is a map Θ_E from $(E^2)^\perp$ onto $(E^3/E^2)^*$ that sends a pair (p,λ) to the pair $(p,\hat{\lambda})$, where $\hat{\lambda}$ is the linear functional on $E^3(p)/E^2(p)$ induced by restricting λ to $E^3(p)$ and then using the fact that λ vanishes on $E^2(p)$ to pass to the quotient. Since $\theta(\omega,p)$ is an isomorphism from $E(p)$ to $E^3(p)/E^2(p)$, we can use it to induce an inner product on $E^3(p)/E^2(p)$, which enables us to identify $E^3(p)/E^2(p)$ with its dual. Then the map $\tilde{\sigma}_{E,\omega,p} = \theta(\omega,p)^{-1}\Theta_E(p)$ is well defined, and sends $E^2(p)^\perp$ onto $E(p)$. Let $\sigma_{E,p}$ be the composite map obtained by first applying $\tilde{\sigma}_{E,\omega,p}$ and then a clockwise 90^o rotation in $E(p)$. (Here "clockwise" refers to the orientation ω. Notice that the rotation changes sign if we reverse ω, and $\tilde{\sigma}_{E,\omega,p}$ also changes sign, so $\sigma_{E,p}$ does not depend on ω.) If (f,g) is an orthonormal basis of sections of E on an open set U, ω is the orientation such that (f,g) is positively oriented, and $v \in E(p)$, $v = v_1 f(p) + v_2 g(p)$, then $\theta(\omega,p)(v) = v_1 F + v_2 G$, where F, G are, respectively, the classes modulo $E^2(p)$ of $[f,[f,g]](p)$ and $[g,[f,g]](p)$ The induced inner product on $E^3(p)/E^2(p)$ is the one that makes (F,G) an orthonormal basis. If $(p,\lambda)\in(E^2)^\perp$, then $\langle\hat\lambda,F\rangle=\langle\lambda,[f,[f,g]](p)\rangle$ and $\langle\hat\lambda,G\rangle=\langle\lambda,[g,[f,g]](p)\rangle$. So the vector in $E^3(p)/E^2(p)$ that corresponds to $\hat\lambda$ under the identification of $E^3(p)/E^2(p)$ with its dual is $\langle\hat\lambda,F\rangle F+\langle\hat\lambda,G\rangle G$. Hence $\tilde\sigma_{E,\omega,p}(\lambda)=\langle\hat\lambda,F\rangle f(p)+\langle\hat\lambda,G\rangle g(p)$, and then $\sigma_{E,p}(\lambda)=\langle\hat\lambda,G\rangle f(p)-\langle\hat\lambda,F\rangle g(p)$, i.e.

$$\sigma_{E,p}(\lambda)=\langle\lambda,[g,[f,g]](p)\rangle f(p)-\langle\lambda,[f,[f,g]](p)\rangle g(p)\,. \tag{46}$$

Now suppose that $(p,\lambda)=\xi\in RA(E)$. Then

$$d\pi_M^*(\mathcal{X}_E(p,\lambda))=\frac{\sigma_E(p,\lambda)}{\|\sigma_E(p,\lambda)\|}\,. \tag{47}$$

So we have proved

Proposition 12 *If (M,E,G) is a sub-Riemannian manifold such that E is 3-regular of partial type $(2,1,2)$. Then there exists a smooth surjective vector bundle map $\sigma_E:(E^2)^\perp\to E$ with the property that, if v is any unit vector in $E(p)$, and $\sigma_E(p,\lambda)=v$, then $d\pi_M^*(p,\lambda)(\mathcal{X}_E(p,\lambda))=v$. In particular, if $p\in M$ and v is any unit vector in $E(p)$, then there exists a regular abnormal extremal $\gamma:(-\varepsilon,\varepsilon)\to M$ such that $\gamma(0)=p$ and $\dot\gamma(0)=v$.* ∎

9.3 The 4- and 5-dimensional cases

The results for the four- and five-dimensional cases are essentially contained in the conclusions of the previous subsection. In dimension 4, Proposition 11 applies, and in addition the set $(E^3)^\perp\cap T^\# M$ is empty, so there are no totally singular abnormal extremals. Therefore we get the following theorem, already derived in [25]:

Theorem 8 *If (M,E,G) is a sub-Riemannian manifold such that $\dim M=4$ and E is a regular two-dimensional distribution, then* *there exists a line subbundle L of E such that the abnormal extremals are exactly the L-admissible arcs. Through every point in M there pass exactly two oriented (or one unoriented) locally simple abnormal extremals parametrized by arc-length. Moreover every locally simple abnormal extremal parameterized by arc-length is regular, so it is locally uniquely optimal.* ∎

In dimension five, the type of a regular distribution E can be $(2,1,2)$ or $(2,1,1,1)$. In the $(2,1,2)$ case Proposition 12 applies. Moreover, it is easy to see that in this case *the regular abnormal extremals are the projections on M of the trajectories of a flow on the unit sphere bundle of E.* To prove this, we work locally, and use Equations (20) and (21). Let (f,g) be an orthonormal basis of sections of E. Extend the Riemannian metric G from E to the whole tangent bundle by declaring $(f,g,[f,g],[f,[f,g]],[g,[f,g]])$ to be an orthonormal basis. For each $x \in M$, $v \in T_xM$, let $\hat{\lambda}_x(v)$ be the unique linear functional $\lambda \in T_x^*M$ such that $\langle \lambda, [g,[f,g]](x)\rangle = \langle v, f(x)\rangle$, $\langle \lambda, [f,[f,g]](x)\rangle = -\langle v, g(x)\rangle$, $\langle \lambda, [f,g](x)\rangle = \langle v, [f,g](x)\rangle$, $\langle \lambda, f(x)\rangle = \langle v, [f,[f,g]](x)\rangle$ and $\langle \lambda, g(x)\rangle = \langle v, [g,[f,g]](x)\rangle$. Notice that $\hat{\lambda}_x(v) \in E^2(x)^\perp$ if and only if $v \in E(p)$. Suppose a curve $t \to (x(t),\lambda(t)) = \Gamma(t) \to RA(E)$ is a solution of (20), (21). Reparametrize Γ so that the equations simply read

$$\begin{aligned} \dot{x} &= \langle \lambda, [g,[f,g]](x)\rangle f(x) - \langle \lambda, [f,[f,g]](x)\rangle g(x)\,, &(48)\\ \dot{\lambda} &= -\langle \lambda, [g,[f,g]](x)\rangle \lambda \frac{\partial f}{\partial x}(x) + \langle \lambda, [f,[f,g]](x)\rangle \lambda \frac{\partial g}{\partial x}(x)\,. &(49) \end{aligned}$$

Let $v(t) = \dot{x}(t)$. Since $\lambda(t) \in E^2(x(t))^\perp$, we must have $\hat{\lambda}_{x(t)}(v(t)) = \lambda(t)$. But then we can plug in $\hat{\lambda}_x(v)$ for λ in the second equation, and express the time derivative of $\hat{\lambda}_{x(t)}(v(t))$ in the form $\tilde{F}(x(t),v(t))$, where $(x,v) \to \tilde{F}(x,v)$ is smooth in both variables and quadratic in v. But

$$\frac{d}{dt}\Big(\hat{\lambda}_{x(t)}(v(t))\Big) = \frac{\partial}{\partial x}\Big(\hat{\lambda}_x(v)\Big)\Big|_{x=x(t),v=v(t)} \cdot \dot{x}(t) + \hat{\lambda}_{x(t)}(\dot{v}(t)) \quad (50)$$

$$= \frac{\partial}{\partial x}\big(\hat{\lambda}_x(v)\big)\Big|_{x=x(t),v=v(t)} \cdot v(t) + \hat{\lambda}_{x(t)}(\dot{v}(t))\,. \quad (51)$$

Since $\hat{\lambda}_x$ is invertible, we end up with an expression for $\dot{v}(t)$ of the form $F(x(t), v(t))$, where F is a smooth function of x and v, quadratic in v. Then the equations of the reparametrized abnormal biextremals are of the form $\dot{x} = v$, $\dot{v} = F(x, v)$. To go back to parametrization by arc-length, we must rewrite these equations in the form $\dot{x} = \frac{v}{\|v\|}$, $\dot{v} = \frac{F(x,v)}{\|v\|}$. If we write $w = \frac{v}{\|v\|}$, then the first equation becomes $\dot{x} = w$. A simple computation shows that

$$\dot{w} = \frac{F(x,v)}{\|v\|^2} - \frac{\langle v, F(x,v)\rangle v}{\|v\|^4} \quad (52)$$

and then $\dot{w} = F(x, w) - \langle w, F(x, w)\rangle w$, since $F(x, v)$ is quadratic in v.

So we have proved

Theorem 9 *Let (M, E, G) be a sub-Riemannian manifold such that $\dim M = 5$ and E is a regular distribution of type $(2, 1, 2)$. Then there exists a smooth flow Φ on the unit sphere bundle $\{(x, w) : x \in M\,,\, w \in E(x)\,,\, \|w\|_G = 1\}$ such that, (i) whenever $t \to (x(t), w(t))$ is a trajectory of this flow then $t \to x(t)$ is a regular abnormal extremal and $w(t) = \dot{x}(t)$, (ii) every regular abnormal extremal arises in this way. In addition, all the locally simple abnormal extremals that are parametrized by arc-length are regular, so they are locally uniquely optimal and are projections on M of trajectories of Φ.* ∎

In the type $(2, 1, 1, 1)$ case, we can apply Proposition 11. We know that the regular abnormal extremals are contained in leaves of $\tilde{L}_E$, but we don't yet have such knowledge about the totally singular abnormal extremals. We will show that the totally singular abnormal extremals are also contained in leaves of $\tilde{L}_E$, and therefore all locally simple abnormal extremals parametrized by arc-length are regular.

Let $p \in M$, and let (X, Y) be an orthonormal basis of sections of $\tilde{L}_E$ in an open neighborhood U of p, chosen so that Y is a section of $\tilde{L}_E$. We show that every locally simple abnormal extremal that is contained in U and is parameterized by arc-length is an integral curve of either Y or $-Y$. Now, let $\gamma : [a, b] \to M$ be a locally simple abnormal extremal that is parametrized by arc-length and is contained in U. We already know that γ is either regular or totally singular, and our conclusion follows trivially if γ is regular, so we will assume that γ is totally singular. Then $\dot{\gamma}(t)$ can be expressed as $\dot{\gamma}(t) = u(t)X(\gamma(t)) + v(t)Y(\gamma(t))$ for almost all $t \in [a, b]$, where u and v are measurable functions on $[a, b]$ such that $u(t)^2 + v(t)^2 = 1$. Let λ be a nonzero field of covectors along γ such that $\Gamma = (\gamma, \lambda)$ is a totally singular abnormal biextremal. Write $X_1 = X, X_2 = Y, X_3 = [X, Y], X_4 = [X, [X, Y]], X_5 = [Y, [X, Y]], X_6 = [X, [X, [X, Y]]], X_7 = [X, [Y, [X, Y]]$, and let $\varphi_k(t) = H_{X_k}(\gamma(t), \lambda(t))$. Then, since $\lambda(t)$ annihilates $E^3(\gamma(t))$ for all $t \in [a, b]$, we have $\varphi_k(t) \equiv 0$ for $k = 1, 2, 3, 4, 5$. On the other hand, since $[Y, [X, Y]]$ is a section of E^2, X_7 is a section of E^3, and then φ_7 vanishes along Γ. Now $\dot{\varphi}_4(t) = u(t)\varphi_6(t) + v(t)\varphi_7(t)$, so $0 = \dot{\varphi}_4(t) = u(t)\varphi_6(t)$. Since $\lambda(t)$ never vanishes, and the seven vector fields X_j span the tangent space at each point, we conclude that $\varphi_6(t) \neq 0$ for all t. But then $u(t) \equiv 0$, so $\dot{\gamma}(t) = v(t)Y(\gamma(t))$ for almost all $t \in [a, b]$, and then γ is contained in a leaf of the foliation induced by $\tilde{L}_E$, so γ is in fact a regular abnormal extremal as well.

So we have proved a result very similar to the derived eariler for the four-dimensional case:

Theorem 10 *Let (M, E, G) be a sub-Riemannian manifold such that $\dim M = 5$ and E is two-dimensional and regular of type $(2, 1, 1, 1)$. Then there exists a line subbundle $\tilde{L}_E$ of E such that the abnormal extremals are exactly the $\tilde{L}_E$-arcs. Every locally simple abnormal extremal parametrized by arc-length is regular. Through every point in M there pass exactly two oriented (or one unoriented) locally simple abnormal extremals parametrized by arc-length. All the locally simple abnormal extremals parametrized by arc-length are locally uniquely optimal.* ∎

9.4 The three-dimensional case

Let (M, E, G) be a sub-Riemannian manifold such that $\dim M = 3$ and E is a two-dimensional 3-generating distribution. Then it easy to see that, if the set $S_E = \{p \in M : \dim E^2(p) = 2\}$ is nonempty, then it is a smooth two-dimensional submanifold of M. Moreover, the tangent space of S_E at a $p \in S_E$ and the space $E(p)$ are transversal, so the intersection $\tilde{L}_E(p) = E(p) \cap T_pS_E$ is one-dimensional. Then $\tilde{L}_E$ is a smooth line subbundle of TS_E. Clearly, a nonzero $\lambda \in T_p^*M$ cannot annihilate $E^2(p)$ unless $p \in S_E$. So all the nonconstant abnormal extremals of E are contained in S_E. Since these extremals must have at each point a direction that is tangent to S_E and belongs to E, we see that every nonconstant abnormal extremal is contained in a leaf of the foliation $\tilde{F}_E$ of S_E induced by $\tilde{L}_E$. Conversely, it is easy to show that every locally simple curve contained in a leaf of $\tilde{F}_E$ and parametrized by arc-length is a regular abnormal extremal. Therefore the following result holds.

Theorem 11 *Let (M, E, G) be a sub-Riemannian manifold such that $\dim M = 3$ and E is two-dimensional and 3-generating. Then either $S_E = \emptyset$ or S_E is a two-dimensional smooth surface in M, and there is a line subbundle $\tilde{L}_E = (E\lceil S_E) \cap TS_E$ of TS_E such that the abnormal extremals of (M, E, G) are exactly the locally absolutely continuous curves that are contained in the leaves of $\tilde{L}_E$. All the locally simple abnormal extremals parametrized by arc-length are regular, and in particular locally uniquely optimal.* ∎

Remark 13 It is well known that the singularities of a generic two-dimensional distribution E on a three-dimensional manifold M are as follows: the points where $\dim E^2(p) = 2$ form a smooth two-dimensional surface S_E, that consists of two kinds of points: (a) points p where $\dim E^3(p) = 2$, and (b) points p where $\dim E^3(p) = 2$. The points of the second kind form a discrete subset of S_E. Theorem 11 therefore provides, for generic distributions E, a complete description of the abnormal extremals on an open dense set whose complement is at most a discrete set.

On the other hand, it is also possible to consider much more degenerate distributions E. For example, one can easily construct, for arbitrarily large ℓ, sub-Riemannian manifolds (M, E, G) of dimension 3 with the property that ℓ is the smallest number ν such that $E^{\nu}(p) = T_pM$ for all $p \in M$, on which there are E-arcs that are strictly abnormal extremals and are uniquely locally optimal. For example, let $M = \mathbb{R}^3$ with the usual coordinates x, y, z. Let E be the two-dimensional distribution spanned by $f = \frac{\partial}{\partial x}$ and $g = (1-x)\frac{\partial}{\partial y} + x^{2k}\frac{\partial}{\partial z}$, where $k > 0$ is an integer. Let G be the Riemannian metric on E under which (f, g) forms an orthonormal basis. It is easily checked that E is nonholonomic and the smallest number ℓ such that $E^{\ell}(p) = T_pM$ for all $p \in M$ is $2k + 1$. Let $\gamma : [a, b] \to M$ be a curve given by $\gamma(t) = (0, t, 0)$. Then, with an argument similar to the one used in Subsection 2.3, it can be proved that γ is a strictly abnormal extremal, and γ is uniquely optimal if $b - a < \frac{2}{2k+1}$. ∎

9.5 A Lie group example

The preceding discussion makes it easy to produce examples of locally uniquely optimal abnormal extremals for invariant sub-Riemannian metrics on Lie groups. For example, let G be any four-dimensional Lie group whose Lie algebra L has two generators f and g such that (i) f, g, $[f, g]$ and $[f, [f, g]]$ form a basis of L, (ii) $[g, [f, g]]$ belongs to the linear span of f, g, and $[f, g]$, and (iii) $[g, [f, g]]$ does not belong to the linear span of f and $[f, g]$. (For instance, one can take $G = SO(3) \times \mathbb{R}$, and let $f = K_1 \oplus 1$, $g = (K_1 + K_2) \oplus 2$, where K_1, K_2, K_3 are generators of the Lie algebra $so(3)$ of $SO(3)$ such that $[K_1, K_2] = K_3$,

$[K_2, K_3] = K_1$ and $[K_3, K_1] = K_2$, and we are identifying the Lie algebra of G with the direct sum $so(3) \oplus \mathbb{R}$. It is easily verified that f, g, $[f,g]$ and $[f,[f,g]]$ are linearly independent, and $[g,[f,g]] = 2f - g$, so all our conditions hold.) Let E be the subbundle of TG spanned by f and g, and define a sub-Riemannian structure by letting f and g be an orthonormal basis of sections. Theorem 6 then implies that the integral curves of g are strictly abnormal locally uniquely optimal trajectories.

9.6 A nonsmooth abnormal extremal

In view of the preceding discussion, it is clear that the simplest situation involving a regular bracket-generating rank-2 distribution where a nonsmooth abnormal extremal parametrized by arc-length could exist is in dimension 6. We now exhibit such an example. Let $M = \mathbb{R}^6$, and define vector fields f, g by letting $f = \frac{\partial}{x_1}$, $g = \frac{\partial}{x_2} + x_1\frac{\partial}{x_3} + x_1^2\frac{\partial}{x_4} + x_1x_2\frac{\partial}{x_5} + x_1^2x_2\frac{\partial}{x_6}$. We then let E be the distribution spanned by f and g, and define a metric on E by declaring (f,g) to be an orthonormal basis. Clearly, $[f,g] = \frac{\partial}{x_3} + 2x_1\frac{\partial}{x_4} + x_2\frac{\partial}{x_5} + 2x_1x_2\frac{\partial}{x_6}$, $[f,[f,g]] = 2\frac{\partial}{x_4} + 2x_2\frac{\partial}{x_6}$, $[g,[f,g]] = \frac{\partial}{x_5} + 2x_1\frac{\partial}{x_6}$, $[f,[f,[f,g]]] = [g,[g,[f,g]]] = 0$, and $[g,[f,[f,g]]] = [f,[g,[f,g]]] = 2\frac{\partial}{x_6}$. Therefore E is bracket-generating and regular of type $(2,1,2,1)$. Let $\gamma : \mathbb{R} \to M$ be the curve given by $\gamma(t) = (0,t,0,0,0,0)$ for $t \leq 0$ and $\gamma(t) = (t,0,0,0,0,0)$ for $t \geq 0$. Then γ is an E-curve, and a simple computation shows that γ is a simple abnormal extremal parametrized by arc-length. It is clear that γ is not smooth, and therefore γ is not regular.

It can be shown that γ is not locally optimal (cf. Appendix E.) We do not know any examples of locally optimal nonsmooth extremals.

Appendix A: The Gaveau–Brockett problem

We study a sub-Riemannian minimization problem first considered by Gaveau in 1977 in [9] and then by Brockett in 1981 in [5]. The manifold M for this problem is $\mathbb{R}^n \times so(n)$. Gaveau's Theorem 1, in p. 133, says, in our terminology, that if $Z \in so(n)$ has rank > 2 then there are no normal extremals joining $(0,0)$ to $(0,Z)$. If true, this would have implied that the minimum-length arcs from $(0,0)$ to $(0,Z)$ —whose existence follows by standard arguments— are strictly abnormal extremals, thereby proving, 14 years before Montgomery's 1991 example, that strictly abnormal sub-Riemannian minimizers can occur. However, Brockett explicitly computed, for arbitrary $Z \in so(n)$, the set $N(Z)$ of all normal extremals from $(0,0)$ to $(0,Z)$, and showed in particular that $N(Z) \neq \emptyset$. In addition, he determined the elements of $N(Z)$ of minimum length, and computed their length $L(Z)$. Finally, he stated that "an elementary Lagrange multiplier argument shows that there exists a skew-symmetric matrix Λ such that $\dot{u} + \Lambda u = 0$" ([5], p. 22), which is tantamount to saying that all minimizers are normal extremals. Together, these results imply that $L(Z)$ is the minimum length, and the members of $N(Z)$ of length $L(Z)$ are the minimizers. However, Brockett did not spell out his Lagrange multiplier argument. Moreover, it turns out that the full proof of Brockett's assertion shows the occurrence of a new phenomenon: if $n > 2$, then there exist abnormal extremals, and even strictly abnormal ones, but the strictly abnormal extremals cannot be minimizers. In view of the existence of conflicting claims surrounding Brockett's result, plus the fact that the proof is interesting in its own right, we will now show how to derive Brockett's conclusion, using the Maximum Principle together with a supplementary inductive argument to exclude the possibility of strictly abnormal minimizers.

Let V be a finite-dimensional linear space over $\mathbb{R}$, endowed with a Euclidean inner product $\langle\cdot,\cdot\rangle$. Let $Z(V)$ be the space of all linear, skew-symmetric maps from V to V. Let $n = \dim V$. Then $Z(V)$ has dimension $\frac{n(n-1)}{2}$. The space $Z(V)$ is also Euclidean, with the inner product defined by $\langle A, B\rangle = -Tr(AB)$ for $A, B \in Z(V)$, where $Tr(\cdot)$ is the trace. Clearly V is isomorphic to $\mathbb{R}^n$ and $Z(V)$ is isomorphic to $so(n)$. Moreover, we can identify $Z(V)$ with the second exterior power $\wedge^2 V$, by means of the canonical isomorphism $\mu_V : \wedge^2 V \to Z(V)$ defined as follows. For $v, w \in V$, let $v \wedge w \in Z(V)$ denote the linear skew-symmetric map on V defined by letting

$$(v \wedge w)(x) \stackrel{\text{def}}{=} \langle w, x\rangle v - \langle v, x\rangle w \ \text{ for } \ x \in V\,.$$

Then $(v, w) \to v \wedge w$ is a bilinear skew-symmetric map from $V \times V$ to $Z(V)$, and therefore induces a linear map $\mu_V : \wedge^2 V \to Z(V)$ that sends the exterior product $v \wedge w \in \wedge^2 V$ to the skew-symmetric map $v \wedge w \in Z(V)$. The map μ_V is easily seen to be an isomorphism.

Let $M_V = V \times Z(V)$. Then M_V is a nilpotent Lie algebra of dimension $\frac{n(n+1)}{2}$. The Lie bracket on M_V is defined by $[(x, z), (\tilde{x}, \tilde{z})] = (0, 2x \wedge \tilde{x})$. Via the exponential map, M_V can be identified with its corresponding connected simply connected Lie group, and the group multiplication on M_V is then given by $(x, z)\cdot(\tilde{x}, \tilde{z}) = (x + \tilde{x}, z + \tilde{z} + x \wedge \tilde{x})$. As a Lie algebra, M_V is generated by V, identified with $V \times \{0\}$. If we let f_u denote the left-invariant vector field on M_V that corresponds to $(u, 0)$, then f_u can be regarded as a map from M_V to M_V, given by $f_u(x, z) = (u, x \wedge u)$. It is not hard to see that $[f_u, f_v] = 2g_{u,v}$, where $g_{u,v}(x, z) = (0, u \wedge v)$. The linear subspace V of M_V gives rise to a left-invariant bracket-generating distribution E_V, which carries a left-invariant Riemannian metric G_V such that $\langle f_u, f_v\rangle_{G_V} \equiv \langle u, v\rangle$.

The *Gaveau-Brockett problem* is that of determining the structure of the trajectories that minimize length for the sub-Riemannian manifold (M_V, E_V, G_V). The E_V-admissible trajectories of (M_V, E_V, G_V) are exactly the trajectories of the control system on M_V given by

$$\dot{x} = u\,, \quad \dot{z} = x \wedge u\,, \tag{53}$$

where the control set is V. A simple application of Ascoli's Theorem shows that given any pair (p, q) of points of M_V there exists at least one E_V-arc γ such that $\partial\gamma = (p, q)$ and γ minimizes length among all the E_V-arcs $\tilde{\gamma}$ such that $\partial\tilde{\gamma} = (p, q)$. Our problem is to understand these length-minimizing arcs. (Gaveau and Brockett actually considered the problem of minimizing the "energy" $\frac{1}{2}\int_0^1 \|\dot{\gamma}(s)\|^2 ds$ among all the E_V-arcs $\gamma : [0, 1] \to M_V$ such that $\gamma(0) = p$ and $\gamma(1) = q$. It is easy to show that this problem is equivalent to ours, in the sense that an arc $\gamma : [0, 1] \to M_V$, for which $\int_0^1 \|\dot{\gamma}(s)\| ds = \tau$, solves the Gaveau-Brockett minimization problem if and only if the arc $\tilde{\gamma} : [0, \tau] \to M_V$ given by $\tilde{\gamma}(t\tau) = \gamma(t)$ for $0 \le t \le 1$ is parametrized by arc-length and minimizes length.) Since one can always reparametrize an arc by arc-length without changing the length, we will limit ourselves to considering arcs parametrized by arc-length. Finally, in view of the left-invariance of the problem, we may assume without loss of generality that $p = 0$.

Let F_V be the control system defined by the equations (53) with the control constraint $\|u\|^2 \le 1$. Then the E_V-arcs that minimize arc-length and are parametrized by arc-length are precisely the minimum-time trajectories of F_V. Let U denote the closed unit ball in V. After identifying in an obvious way the dual of M_V with M_V itself, and the cotangent bundle T^*M_V with $M_V \times M_V$, the Hamiltonian of F_V is the function H on $M_V \times M_V \times U$ given by

$$\begin{aligned} H(x, z, \lambda, \Lambda, u) &= \langle\lambda, u\rangle + \langle\Lambda, x \wedge u\rangle = \langle\lambda, u\rangle - Tr(\Lambda.(x \wedge u)) \\ &= \langle\lambda, u\rangle + 2\langle x, \Lambda u\rangle = \langle\lambda, u\rangle - 2\langle\Lambda x, u\rangle\,, \end{aligned} \tag{54}$$

for $x \in V$, $z \in Z(V)$, $\lambda \in V$, $\Lambda \in Z(V)$, $u \in U$. The adjoint equations are

$$\dot{\lambda} = -2\Lambda\, u\,, \quad \dot{\Lambda} = 0\,. \tag{55}$$

A trajectory $[0,T] \ni t \to \gamma(t) = (x(t), z(t)) \in V \times Z(V)$ corresponding to a control $u : [0,T] \to V$ is an extremal if there exist a nowhere vanishing absolutely continuous function $[0,T] \ni t \to (\lambda(t), \Lambda(t)) \in V \times Z(V)$ and a constant $\lambda_0 \geq 0$ such that (a) (λ, Λ) satisfies the adjoint equations (55), and (b) the minimization condition

$$-\lambda_0 = H(x(t), z(t), \lambda(t), \Lambda(t), u(t)) = \min_{v \in U} H(x(t), z(t), \lambda(t), \Lambda(t), v) \tag{56}$$

holds for almost all $t \in [0,T]$. If λ_0 can be chosen so that $\lambda_0 > 0$, then γ is a normal extremal. In that case, we can always assume $\lambda_0 = 1$ due to homogeneity. If λ_0 can be chosen equal to 0, then γ is an abnormal extremal. (As explained before, the two possibilities are not mutually exclusive.)

Assume that $T > 0$ and $\gamma = (x, z) : [0,T] \to M_V$ is parametrized by arc-length and is a normal extremal. Let $u = \dot{x}$, and let λ, Λ be a nontrivial solution of (55) such that the minimization condition holds with $\lambda_0 = 1$. It is clear that the minimum value on the ball U of the functional $v \to \langle \lambda(t) - 2\Lambda(t)x(t), v\rangle$ is $-\|\langle\lambda(t) - 2\Lambda(t)x(t)\|$, so $\|\langle\lambda(t) - 2\Lambda(t)x(t)\| = 1$ for a.e. t. Moreover, the minimum is attained for $v = -\lambda(t) + 2\Lambda(t)x(t)$. This yields $u(t) = -\lambda(t) + 2\Lambda x(t)$ for a.e. t, which further implies that the function $u(\cdot)$, which was only assumed *a priori* to be measurable and U-valued, is in fact absolutely continuous, so the equality $u = -\lambda + 2\Lambda x$ holds in fact for all t, and $\dot{u} = -\dot{\lambda} + 2\Lambda\dot{x} = 4\Lambda u$. So the control u satisfies the equation $\dot{u} = \bar{\Lambda}u$ for some $\bar{\Lambda} \in Z(V)$, and in addition $\|u(t)\| = 1$ for all t. Conversely, suppose $t \to (x(t), z(t)) = \gamma(t)$ is a trajectory that corresponds to a control $u = \dot{x}$ that satisfies $\dot{u} = \bar{\Lambda}u$ for some $\bar{\Lambda} \in Z(V)$ and $\|u(0)\| = 1$. Then $\|u(t)\| = 1$ for all t, so γ is parametrized by arc-length. Define (λ, Λ) by $\Lambda = \frac{1}{4}\bar{\Lambda}$, $\lambda(t) = 2\Lambda x(t) - u(t)$. Then (λ, Λ) satisfies the adjoint equations (55) and the minimization condition (63) holds with $\lambda_0 = 1$, which implies that γ is a normal extremal. So we have proved the following characterization of normal extremals:

Proposition 13 *Let $T \geq 0$. A trajectory $\gamma = (x,z) : [0,T] \to M_V$, corresponding to a control $u = \dot{x}$ is a normal extremal for (M_V, E_V, G_V) parametrized by arc-length if and only if there exists a skew-symmetric linear map $\bar{\Lambda}$ from V to V such that u satisfies $\dot{u} \equiv \bar{\Lambda} u$ and $\|u(0)\| = 1$.*

Using this characterization, we now prove:

Theorem 12 *All the minimum-length trajectories of (M_V, E_V, G_V) that are parametrized by arc-length are normal extremals.*

PROOF. We use induction on the dimension n of V. If $n = 1$, this is trivial. Assume that the conclusion is true for all $1 \leq n < k$. Let V be of dimension k. Let $\gamma = (x,z) : [0,T] \to M_V$ be a length-minimizing trajectory such that $\gamma(0 = (0,0)$. Let $u = \dot{x}$ be the corresponding control. The Maximum Principle guarantees that γ is either a normal extremal or an abnormal one. Let us assume that γ is an abnormal extremal. We will prove that in that case γ is a normal extremal as well.

Since γ is an abnormal extremal, there exists a nowhere vanishing absolutely continuous function $(\lambda, \Lambda) : [0,T] \to V \times Z(V)$ such that (λ, Λ) satisfies the adjoint equations (55) and in addition

$$\lambda - 2\Lambda\, x \equiv 0 \ \text{ on } [0,T]. \tag{57}$$

In particular (57) implies that $\lambda(0) = 0$. Differentiating (57) we have $0 = \frac{d}{dt}(\lambda - 2\Lambda x) = -4\Lambda u$, so $\Lambda u \equiv 0$. But then $\dot{\lambda} = -2\Lambda u \equiv 0$. Therefore $\lambda \equiv 0$ on $[0,T]$. The nontriviality of (λ, Λ) then implies that $\Lambda \neq 0$. Let $S = \ker \Lambda$. Then $\dim S < k$. Since $\lambda \equiv 0$, (57) implies that $x(t) \in S$ for all t, and then $u(t) \in S$ as well. Therefore $\gamma = (x,z)$

is a trajectory of (53) with initial condition $x(0) = 0$, $z(0) = 0$, corresponding to the S-valued control $u(\cdot)$. This implies that γ takes values in $S \times Z(S)$. Clearly γ also minimizes length and is parametrized by arc-length when regarded as an admissible arc for the sub-Riemannian manifold (M_S, E_S, G_S). Since S has dimension $< k$, the inductive hypothesis tells us that γ is a normal extremal of (M_S, E_S, G_S). Then Proposition 13 implies that there exists a $\Theta \in Z(S)$ such that $\dot{u} \equiv \Theta u$. Clearly, we can extend Θ to a skew-symmetric linear map $\bar{\Theta}$ on V. Since $u(t) \in S$ for all t, we have $\dot{u}(t) = \bar{\Theta}u(t)$ for $t \in [0, T]$, and $\bar{\Theta} \in Z(V)$. But then it follows from Proposition 13 that γ is a normal extremal. ∎

Remark 14 In the above proof, we have shown that all the *minimizing* abnormal extremals for the Gaveau-Brockett problem that are parametrized by arc-length are normal extremals. If $\dim V > 2$, then it is easy to produce examples of non-minimizing strictly abnormal extremals parametrized by arc-length. For example, pick any nonzero $\Lambda \in Z(V)$ with a nontrivial kernel K, let $u : [0, T] \to K$ be an arbitrary measurable function such that $\|u(t)\| = 1$ for all t, and then let $x(t) = \int_0^t u(s)ds$, $z(t) = \int_0^t (x(s) \wedge u(s))\, ds$, $\gamma(t) = (x(t), z(t))$. If we take $\lambda(t) \equiv 0$, then it is easy to see that $t \to (x(t), z(t), \lambda(t), \Lambda)$ is an abnormal biextremal, so γ is an abnormal extremal. By choosing $u(\cdot)$ to be discontinuous, we can make sure that γ is not a normal extremal. If $\dim V \geq 4$, then it is clear that $u(\cdot)$ can be taken to be smooth, and even real-analytic. (It suffices to make sure that $u(\cdot)$ is not a solution of an equation $\dot{u} = \Theta u$ for any skew-symmetric map Θ.) ∎

Brockett has shown in [5] how Proposition 13 and Theorem 12 can be used to derive, for any $Z \in Z(V)$, an explicit formula for a minimum-length trajectory γ from $(0, 0)$ to $(0, Z)$, and for the sub-Riemannian distance from $(0, 0)$ to $(0, Z)$.

Appendix B: Proof of Theorem 1

We show that $\Sigma(M, E, G)$ is the set of trajectories of a control system, and then use the "Pontryagin Maximum Principle" (cf. [2], [13], [15], [20], [26]), which gives a necessary condition for a trajectory γ of a control system to minimize a cost functional such as time. In our situation, the condition says precisely that γ is an extremal.

We will first review the definition of a control system and the statement of the Maximum Principle. For simplicity, we will only consider the case of interest, namely, *minimum time problems* with *fixed endpoints*, arising from an *autonomous* control system of class C^1, although the Maximum Principle is a much more general result, valid as well for nonautonomous systems with vector fields that are locally Lipschitz but not necessarily of class C^1, and for minimization problems with variable endpoints, and cost functionals other than time.

B.1: Control systems

Definition 18 An *autonomous control system of class* C^1 on a smooth manifold M is a parametrized family $F = \{f_u : u \in \mathcal{U}\}$ of vector fields of class C^1 on M. ∎

Remark 15 We emphasize that in the above definition $\mathcal{U}$ *is just an arbitrary set*, and there is no requirement that $\mathcal{U}$ be endowed with extra structure such as a topology or a metric. ∎

To such a system F one associates a collection $\Sigma(F)$ of *trajectories* as follows:

Definition 19 A *control* for F is a function $\eta : [a, b] \to \mathcal{U}$ defined on some compact interval $[a, b]$. An F-*trajectory* generated by a control η is an arc[‡] $\gamma : [a, b] \to M$ such that $\dot\gamma(t) = f(\gamma(t), \eta(t))$ for almost all $t \in$

[‡] Recall that arcs are absolutely continuous by definition.

$[a,b]$, where we have written $f(x,u)\stackrel{\text{def}}{=}f_u(x)$. A control η is *admissible* if the time-varying vector field $(x,t)\to f_\eta(x,t)\stackrel{\text{def}}{=}f(x,\eta(t))$ arising from η satisfies the following *Carathéodory conditions*: (i) $f_\eta(x,t)$ is Lebesgue measurable as a function of t for each fixed x, (ii) for every chart $\kappa : V\to \mathbb{R}^n$ of M and every compact subset K of $\kappa(V)$ there exists an integrable function $\varphi:[a,b]\to\mathbb{R}$ such that

$$\|f_\eta^\kappa(x,t)\|+\left\|\frac{\partial f_\eta^\kappa}{\partial x}(x,t)\right\|\leq\varphi(t)\ \text{ for all }\ (x,t)\in K\times[a,b]\,, \tag{58}$$

where $f_\eta^\kappa:\kappa(V)\times[a,b]\to\mathbb{R}^n$ is the representation of f_η with respect to the chart κ, i.e. $f_\eta^\kappa(x,t)=d\kappa(\kappa^{-1}(x)).f_\eta(\kappa^{-1}(x),t)$. An *admissible trajectory* of F is an F-trajectory that is generated by an admissible control η. ∎

We let $\Sigma(F)$ denote the set of all admissible trajectories of F. If $\gamma\in\Sigma(F)$, and the domain of γ is the interval $[a,b]$, we write $T(\gamma)\stackrel{\text{def}}{=}b-a$. A trajectory $\gamma\in\Sigma(F)$ is *time-optimal* if $T(\gamma)\leq T(\tilde\gamma)$ for all $\gamma\in\Sigma(F)$ such that $\partial\tilde\gamma=\partial\gamma$.

Remark 16 In Control Theory, the manifold M on which the vector fields f_u are defined is known as the *state space* of the system $F=\{f_u : u\in\mathcal{U}\}$, and the set $\mathcal{U}$ is known as the "control space," or the "set of control values." The word "control" is used somewhat vaguely to mean either the $\mathcal{U}$-valued "variable" u or a $\mathcal{U}$-valued function $t\to\eta(t)$ defined on an interval. It is customary to introduce control systems by means of expressions such as "the system $\dot x=f(x,u)$, $x\in M$, $u\in\mathcal{U}$." This is to be interpreted as another way of referring to the parametrized family of vector fields $\{f(\cdot,u):u\in\mathcal{U}\}$. The formal expression $\dot x = f(x,u)$ gives rise, for each control function $[a,b]\ni t\to\eta(t)\in U$, to a non-autonomous ordinary differential equation $\dot x=f(x,\eta(t))$, whose solutions are the trajectories of the system. In addition, if $\mathcal{U}$ is described by some formula, one often uses this formula instead of the expression

"$u \in \mathcal{U}$." As an illustration of the use of this terminology, consider the example studied in Subsection 2.2. The space $M = \mathbb{R}^3$ was given a sub-Riemannian manifold structure by means of a two-dimensional distribution E, spanned by two vector fields f and g. The metric G on E was defined by declaring f and g to be an orthonormal basis. We can then construct a control system F in the sense of our definition by letting $\mathcal{U}$ be the unit ball in $\mathbb{R}^2$, and associating to each $u \in \mathcal{U}$ a vector field f_u, defined by $f_u = u_1 f + u_2 g$, if $u = (u_1, u_2) \in \mathcal{U}$. The system F is then the vector field family $F = \{f_u : u \in \mathcal{U}\}$. It is easy to verify that the set $\Sigma(F)$ is exactly $\Sigma(M, E, G)$, so we have realized $\Sigma(M, E, G)$ as the collection of trajectories of a control system. Instead of using the somewhat awkward expression "the system $F = \{f_u : u \in \mathcal{U}\}$, where $\mathcal{U}$ is the unit ball in $\mathbb{R}^2$ and $f_u = u_1 f + u_2 g$ if $u = (u_1, u_2) \in \mathcal{U}$," we can introduce F as

$$\text{the system } \dot{x} = u_1 f(x) + u_2 g(x)\,, \quad x \in \mathbb{R}^3\,, \quad u_1^2 + u_2^2 \leq 1\,, \tag{59}$$

or, using the formulas for f and g, and u, v rather than u_1, u_2, as

$$\text{the system } \dot{x} = u\,, \quad \dot{y} = v\,, \quad \dot{z} = vx - uy\,, \quad u^2 + v^2 \leq 1\,. \tag{60}$$

In expressions such as (60), the symbols x, y, z are the "state variables," and u, v are the "control variables." From now on, we will always use this Control Theory language to introduce control systems, with the understanding that the precise interpretation of expressions in this language is obtained by translating them into the vector field formulation used in our definition[§]. ∎

[§]In Control Theory, it is sometimes desirable to single out a special class of control functions, such as the set of square-integrable, or piecewise continuous, or piecewise smooth, or piecewise constant functions, assuming, in each case, that the control space has whatever extra structure is needed for such a notion to make sense. To allow for this, one has to define a control system to be a parametrized family of vector fields together with the specification of a class of control functions. In this work, this extra feature is ignored, because we are always taking the class of control functions to be the class of all controls that are admissible in the sense of Definition 19.

B.2: The Maximum Principle

Let $F = \{f_u : u \in \mathcal{U}\}$ be an autonomous control system of class C^1 on a manifold M. The *Maximum Principle* gives a necessary condition for a $\gamma \in \Sigma(F)$ to be time-optimal. To state this condition, we first introduce the *Hamiltonian lift* F^* of F. By definition, F^* is the collection of vector fields $\{\overrightarrow{H_{f_u}} : u \in U\}$ on T^*M. In other words, each f_u gives rise to a function $H_{f_u} : T^*M$ —defined by $H_{f_u}(x, \lambda) = \langle \lambda, f_u(x) \rangle$— and this function gives rise to a Hamilton vector field $f_u^* \stackrel{\text{def}}{=} \overrightarrow{H_{f_u}}$. The collection of all the f_u^* is the system F^*. If $\eta : [a, b] \to \mathcal{U}$ is an admissible control for F, then a trajectory of F^* for η is a pair (γ, ζ), where γ is a trajectory of F for η, and ζ is a field of covectors along γ (i.e. $\zeta(t) \in T^*_{\gamma(t)}M$ for each $t \in [a, b]$) such that, if we write $\Gamma(t) = (\gamma(t), \zeta(t))$, then Γ is absolutely continuous and

$$\dot{\Gamma}(t) = f^*_{\eta(t)}(\Gamma(t)) \tag{61}$$

for almost all $t \in [a, b]$. Equivalently, (γ, ζ) is a trajectory of F^* for η if and only if $\dot{\gamma}(t) = f(\gamma(t), \eta(t))$ and

$$\dot{\zeta}(t) = -\zeta(t) \frac{\partial f}{\partial x}(\gamma(t), \eta(t)) \tag{62}$$

for almost all $t \in [a, b]$. Here (62) is known as the *adjoint equation* along (γ, η). In principle, (62) makes sense in coordinates, if we write the components of points and tangent vectors as columns —so that, for instance, $f(x, u)$ is a column, and $\frac{\partial f}{\partial x}(x, u)$ is a square matrix— and those of covectors as rows. However, since (62) is equivalent to (61), the property characterized in coordinates by (62) is in fact intrinsic.

A field of covectors ζ along a trajectory $\gamma : [a, b] \to M$ of F corresponding to an admissible control $\eta : [a, b] \to \mathcal{U}$ is called an *adjoint vector* along (γ, η) if it is a solution of (62), i.e. if the curve $\Gamma : [a, b] \to T^*M$ given by $\Gamma(t) = (\gamma(t), \zeta(t))$ is absolutely continuous and satisfies (61) for almost every t. An adjoint vector ζ along (γ, η) is *minimizing* if there exists a constant $\zeta_0 \geq 0$ such that

$$-\zeta_0 = \zeta(t)f(\gamma(t),\eta(t)) = \min\{\zeta(t)f(\gamma(t),u) : u \in \mathcal{U}\} \tag{63}$$

for almost all $t \in [a,b]$. Finally, ζ is *nontrivial* if $\zeta(t) \neq 0$ for every t. (Notice that, since (62) is a linear time-varying O.D.E. for ζ, if $\zeta(t) \neq 0$ for *some* t, then ζ is nontrivial, i.e. $\zeta(t) \neq 0$ for *all* t.)

With the above definitions, the Maximum Principle says:

Theorem 13 *If F is an autonomous control system of class C^1 on a manifold M, and γ is a time-optimal trajectory corresponding to an admissible control η, then there exists a nontrivial minimizing adjoint vector along (γ,η).*

B.3: Proof of Theorem 1

We are now ready to show that Theorem 1 is a consequence of Theorem 13. To apply Theorem 13 to our situation, we must first produce, for a sub-Riemannian manifold (M,E,G), a control system F on M such that $\Sigma(F)$ is precisely the set $\Sigma(M,E,G)$ defined before. Recall that in our general formulation of the Maximum Principle the "control space" $\mathcal{U}$ was allowed to be an arbitrary set, with no additional structure whatsoever. In particular, we can take $\mathcal{U}$ to be the set of all smooth sections u of E that satisfy $\|u(x)\|_G \leq 1$ for all $x \in M$. We then define f_u to be just u, so $f_u(x)$ and $f(x,u)$ are simply alternative ways of writing $u(x)$. We let $F = \{f_u : u \in \mathcal{U}\}$. It is then easy to see that the admissible trajectories of F are precisely the arcs in $\Sigma(M,E,G)$. So our control system F has the desired property that $\Sigma(F) = \Sigma(M,E,G)$.

If we now apply the Maximum Principle to this system F, we find that, if $\gamma : [a,b] \to M$ is a time-optimal trajectory, corresponding to a control $\eta : [a,b] \to \mathcal{U}$, then there must exist a nontrivial field ζ of covectors along γ that satisfies, in local coordinates, the equation

$$\dot{\zeta}(t) = -\zeta(t)\,\frac{\partial\eta(t)}{\partial x}(\gamma(t))\,, \tag{64}$$

and a nonnegative constant ζ_0 such that the minimization condition

$$-\zeta_0 = \langle\zeta(t), \eta(t)(\gamma(t))\rangle = \min\{\langle\zeta(t), u(\gamma(t))\rangle : u \in U\} \tag{65}$$

holds for almost all t. Let us write $\Gamma(t) = (\gamma(t), \zeta(t))$.

We will show that γ is an abnormal extremal if $\zeta_0 = 0$ and a normal one if $\zeta_0 \neq 0$.

Suppose first that $\zeta_0 = 0$. Then (65) implies that $\zeta(t)v \geq 0$ for every $v \in E(\gamma(t))$ such that $||v||_G \leq 1$. Therefore $\zeta(t)$ annihilates $E(\gamma(t))$, so $\Gamma(t) \in E^\perp$, and then every $H \in \mathcal{H}_E$ vanishes along Γ. Moreover, (64), together with $\dot\gamma(t) = \eta(t)(\gamma(t))$, say that $\dot\Gamma(t) = \overrightarrow{H_{\eta(t)}}$, where $H_{\eta(t)}$ is the Hamiltonian of the vector field $\eta(t)$. Since $\eta(t)$ is a smooth section of E, we see that $\dot\Gamma(t) \in \overrightarrow{\mathcal{H}_E}(\Gamma(t))$. So Γ is an abnormal extremal.

Next consider the case when $\zeta_0 \neq 0$. We can then multiply ζ by ζ_0^{-1} and obtain a new nontrivial minimizing adjoint vector that corresponds to $\zeta_0 = 1$. Our goal is now is to prove that Γ is a normal biextremal, and this is a local property (cf. Remarks 18 and 19 below), so we may assume that M is open in $\mathbb{R}^n$ and E has a global orthonormal basis $(f_1, \ldots, f_m)$ of sections. Our trajectory γ satisfies $\dot\gamma(t) = \eta(t)(\gamma(t))$, where each $\eta(t)$ is a vector field in $\mathcal{U}$. Write $\eta(t)(x) = \sum_{k=1}^m \eta_k(x,t) f_k(x)$, so the η_k are functions on $M \times [a,b]$, smooth in x for each t, measurable in t, and such that $\sum_k \eta_k(x,t)^2 \leq 1$ for all (x,t).

The minimization condition then says that $\eta(t)(\gamma(t))$ minimizes the function $u \to \langle\zeta(t), u(\gamma(t))\rangle$ on $\mathcal{U}$, i.e. that the functional

$$\mu_t : v = (v_1, \ldots, v_m) \to \sum_k v_k \langle\zeta(t), f_k(\gamma(t))\rangle\,, \tag{66}$$

restricted to the unit ball of $\mathbb{R}^m$, attains its minimum value for the m-tuple $(\eta_1(\gamma(t),t), \ldots, \eta_m(\gamma(t),t))$. From this it clearly follows that $\eta_k(\gamma(t),t) = -\langle\zeta(t), f_k(\gamma(t))\rangle$. On the other hand, the adjoint equation says that

$$\dot\zeta(t) \;\; = \;\; -\zeta(t)\,\frac{\partial}{\partial x}\Big(\eta(t)(x)\Big)\Big|_{x=\gamma(t)} \tag{67}$$

$$
\begin{aligned}
&= -\zeta(t)\,\frac{\partial}{\partial x}\Big(\sum_k \eta_k(x,t) f_k(x)\Big)\Big|_{x=\gamma(t)} && (68)\\
&= -\sum_k \eta_k(\gamma(t),t)\zeta(t)\,\frac{\partial f_k}{\partial x}\big(\gamma(t)\big)\\
&\quad -\frac{\partial}{\partial x}\left(\sum_k \eta_k(x,t)\langle\zeta(t), f_k(\gamma(t))\rangle\right)\Bigg|_{x=\gamma(t)} && (69)\\
&= \sum_k \langle\zeta(t), f_k(\gamma(t))\rangle\zeta(t)\,\frac{\partial f_k}{\partial x}\big(\gamma(t)\big)\\
&\quad -\frac{\partial}{\partial x}\left(\sum_k \eta_k(x,t)\langle\zeta(t), f_k(\gamma(t))\rangle\right)\Bigg|_{x=\gamma(t)}. && (70)
\end{aligned}
$$

Now, we know that μ_t is minimized on the unit ball of $\mathbb{R}^m$ by taking $v = \big(\eta_1(\gamma(t),t),\ldots,\eta_m(\gamma(t),t)\big)$. On the other hand, for each $x \in M$ the m-tuple $v(x) = (\eta_1(x,t),\ldots,\eta_m(x,t))$ belongs to the unit ball. Therefore the function $x \to \sum_k \eta_k(x,t)\langle\zeta(t), f_k(\gamma(t))\rangle$ has a minimum at $x = \gamma(t)$. It follows that

$$
\frac{\partial}{\partial x}\left(\sum_k \eta_k(x,t)\langle\zeta(t), f_k(\gamma(t))\rangle\right)\Bigg|_{x=\gamma(t)} = 0\,. \tag{71}
$$

So

$$
\dot\zeta(t) = \sum_k \langle\zeta(t), f_k(\gamma(t))\rangle\zeta(t)\,\frac{\partial f_k}{\partial x}\big(\gamma(t)\big)\,. \tag{72}
$$

On the other hand, it follows from the equalities $\dot\gamma(t) = \eta(t)(\gamma(t))$, $\eta(t)(\gamma(t)) = \sum_k \eta_k(\gamma(t),t) f_k(\gamma(t))$, and $\eta_k(\gamma(t),t) = -\langle\zeta(t), f_k(\gamma(t))\rangle$, that

$$
\dot\gamma(t) = -\sum_k \langle\zeta(t), f_k(\gamma(t))\rangle f_k(\gamma(t))\,, \tag{73}
$$

It is clear that (73) and (72) are precisely the equations of a trajectory of the Hamiltonian H of (E,G). So Γ is a normal biextremal of (E,G), and therefore γ is a normal extremal of (E,G). ∎

Remark 17 The above proof can be slightly simplified in the special case when E has a global orthonormal basis $(f_1,\ldots,f_m)$ of sections.

Indeed, in that case one can take $\mathcal{U}$ to be the closed unit ball in $\mathbb{R}^m$, define $f_u = u_1 f_1 + \ldots + u_m f_m$ for $u = (u_1, \ldots, u_m) \in \mathcal{U}$, and let $F = \{f_u : u \in \mathcal{U}\}$, so one gets a control system with a much smaller control space than the one we have considered. The proof in this case is identical to the one given above, except that the vector field $\eta(t)$ is now replaced by $\sum_k \eta_k(t) f_k$, so the second terms of (69) and (70) do not occur, and the extra step (71) is not needed to derive (72). The general form of Theorem 1 can then be derived from the one for the special case, by showing that one can always choose a basis of smooth sections of E on some neighborhood of γ. (It is clear that γ can be assumed to begin with to be a curve without double points.) ∎

Remark 18 As opposed to the Riemannian case, it is not possible to prove Theorem 1 by an argument which is purely local in time, e.g. by showing that γ satisfies the desired conclusion on sufficiently short time intervals. The reason for this is that *being an extremal is not a local property of trajectories* γ. Indeed, suppose $\gamma : [a, b] \to M$ is "locally extremal," i.e. such that the restrictions $\gamma_j : I_j \to M$ of γ to intervals I_j that cover $[a, b]$ are extremals. This means that for each j there is a field of covectors ζ_j along γ_j such that (γ_j, ζ_j) is a biextremal. However, *there is no reason why the* ζ_j *should match, making it possible to construct a global* ζ *on* $[a, b]$ *such that* (γ, ζ) *is a biextremal.* (Here again the intuitions arising from the Riemannian case are misleading. In the Riemannian case, one can easily show that the covector ζ is uniquely determined by γ, so any two ζ_j's have to coincide on the intersection of their domains, and then the ζ_j's give rise to a global ζ, so the equation of Riemannian geodesics can be proved by purely local arguments.) It follows that, in order to prove Theorem 1 from the Maximum Principle without assuming the existence of a global basis of sections, one needs a more subtle argument, such as the one given in the text or the one of Remark 17. ∎

Remark 19 Although extremality is not a local property, *biextremality is*. That is, a curve $\Gamma : I \to T^*M$ is a biextremal iff for every $t \in I$ there is an $\varepsilon > 0$ such that the restriction of Γ to $I \cap (t-\varepsilon, t+\varepsilon)$ is a biextremal. ∎

Appendix C: Local optimality of normal extremals

For completeness, we provide a direct control-theoretic proof of the local optimality of normal extremals. Let (M, E, G) be a sub-Riemannian manifold. Let $\gamma_* : (a,b) \to M$ be a normal extremal parametrized by arc-length. Let λ_* be a field of covectors along γ such that $\Gamma_* = (\gamma_*, \lambda_*)$ is a normal biextremal. (For simplicity we assume that the domain of Γ_* is an open interval. Otherwise we can always extend Γ_* to a normal biextremal defined on a slightly larger open interval.) Let H be the Hamiltonian associated to (E, G). Without loss of generality we may assume that λ_* is normalized, i.e. $H(\gamma_*(t), \lambda_*(t)) = -\frac{1}{2}$ for all $t \in (a,b)$. To show that γ_* is locally optimal, we need only to prove that every $c \in (a,b)$ has a neighborhood $J_c = (c-\varepsilon, c+\varepsilon) \subseteq (a,b)$ such that the restriction of γ_* to J_c is optimal. Fix a $c \in (a,b)$. Let $p_* = \gamma_*(c), \lambda_* = \lambda_*(c)$. Pick a smooth hypersurface S of dimension $n-1$ in M passing through p_* such that λ_* vanishes on $T_{p_*}S$. Let $\bar{\lambda}$ be a smooth 1-form on an open neighborhood Ω of p_* such that $\bar{\lambda}(p_*) = \lambda_*$, $\bar{\lambda}(p)$ annihilates T_pS and $H(p, \bar{\lambda}(p)) = -\frac{1}{2}$ for all $p \in S \cap \Omega$. For later use we assume that Ω is small enough so that there exists an orthonormal basis $(X_1, \ldots, X_m)$ of sections of E on Ω, where m is the dimension of E, and also that Ω is the domain of a coordinate chart. Then

$$H(x, \lambda) = -\frac{1}{2} \sum_{k=1}^{m} \langle \lambda, X_k(x) \rangle^2 . \qquad (74)$$

For each $p \in \Omega \cap S$, let $\Gamma_p = (\gamma_p, \lambda_p)$ be the solution of $\dot{\Gamma}(t) = \vec{H}(\Gamma(t)), \Gamma(c) = (p, \bar{\lambda}(p))$. Then clearly $\Gamma_* = \Gamma_{p_*}$. Since $\dot{\gamma}_*(c) \notin T_{p_*}S$ (because $\dot{\gamma}_*(c) = -\sum_{j=1}^{m}\langle\lambda_*(c), X_j(c)\rangle X_j(c)$, so that $\langle\lambda_*(c), \dot{\gamma}_*(c)\rangle = -\sum_{j=1}^{m}\langle\lambda_*(c), X_j(c)\rangle^2 = -1)$, we can apply the Implicit Function Theorem to conclude that, if $\varepsilon > 0$ is small enough and W is a sufficiently small open neighborhood of p_* in S, then the map

$$\mu : (c-\varepsilon, c+\varepsilon) \times W \ni (t,p) \to \gamma_p(t) \in M$$

maps $(c-\varepsilon, c+\varepsilon) \times W$ diffeomorphically onto an open neighborhood U of p_* in M. We may assume that U is contained in Ω. Define a smooth function $V : U \to \mathbb{R}$ and a 1-form ω on U by letting $V(x) = t$ and $\omega(x) = \lambda_p(t)$ if $x = \gamma_p(t)$.

We will show that $dV \equiv -\omega$ on U. For this purpose, let X be the vector field on U given by $X(x) = d\pi_M^*(\Gamma_p(t))\vec{H}(\Gamma_p(t))$ if $x = \gamma_p(t)$. Then γ_p satisfies $\dot{\gamma}_p(t) = X(\gamma_p(t))$. Let $\overrightarrow{H_X}$ be the Hamiltonian lift of X. We claim that $\overrightarrow{H_X}(\Gamma_p(t)) = \vec{H}(\Gamma_p(t))$ for $p \in W$, $t \in (c-\varepsilon, c+\varepsilon)$. Indeed, let $\alpha_k(x) = \langle\omega(x), X_k(x)\rangle$ for $k = 1\ldots, m$, if $x \in U$. Then $X = -\sum_{k=1}^{m}\alpha_k X_k$ on U. Since H is constant along integral curves of $\vec{H}$, we have $H(\Gamma_p(t)) = -\frac{1}{2}$ for $(t,p) \in (c-\varepsilon, c+\varepsilon) \times W$. So $\sum_{k=1}^{m}\alpha_k(x)^2 = 1$ for $x \in U$. It follows from (74) that $\vec{H}(x,\lambda) = -\sum_{k=1}^{m}\langle\lambda, X_k(x)\rangle\overrightarrow{H_{X_k}}(x,\lambda)$. In particular $\vec{H}(\gamma_p(t), \lambda_p(t)) = -\sum_{k=1}^{m}\alpha_k(\gamma_p(t))\overrightarrow{H_{X_k}}(\gamma_p(t), \lambda_p(t))$. On the other hand, it is clear that $\overrightarrow{H_X}(x,\lambda) = -\sum_{k=1}^{m}\alpha_k(x)\overrightarrow{H_{X_k}}(x,\lambda) - \sum_{k=1}^{m}\langle\lambda, X_k(x)\rangle\overrightarrow{\alpha_k}(x,\lambda)$. Since $\sum_{k=1}^{m}\alpha_k(x)^2 = 1$, it follows that $\sum_{k=1}^{m}\alpha_k(x)\overrightarrow{\alpha_k}(x,\lambda) \equiv 0$. Therefore

$$\begin{aligned}
\overrightarrow{H_X}(\gamma_p, \lambda_p) &= -\sum_{k=1}^{m}\alpha_k(\gamma_p)\overrightarrow{H_{X_k}}(\gamma_p, \lambda_p) - \sum_{k=1}^{m}\langle\lambda_p, X_k(\gamma_p)\rangle\overrightarrow{\alpha_k}(\gamma_p, \lambda_p) \\
&= -\sum_{k=1}^{m}\alpha_k(\gamma_p)\overrightarrow{H_{X_k}}(\gamma_p, \lambda_p) - \sum_{k=1}^{m}\alpha_k(\gamma_p)\overrightarrow{\alpha_k}(\gamma_p, \lambda_p) \\
&= \vec{H}(\gamma_p, \lambda_p),
\end{aligned}$$

where γ_p, λ_p are evaluated at t. So Γ_p is also a solution of $\dot{\Gamma} =$

$\overrightarrow{H_X}(\Gamma), \Gamma(c) = (p, \bar{\lambda}(p))$ for $p \in W$. Let $(p,t) \to \theta_t(p)$ be the flow of X on U, so that $\dot{\theta}_t(p) = X(\theta_t(p))$ and $\theta_c(p) = p$ for $p \in U$. Then it is clear that $\gamma_p(t) = \theta_t(p)$ for $p \in W$. For each $t \in (c-\varepsilon, c+\varepsilon)$, let $W_t = \{x \in U : V(x) = t\}$. Then W_t is a smooth hypersurface, and $W_t = \theta_t(W)$. The tangent space T_xW_t of W_t at $x = \gamma_p(t)$ is equal to $d\theta_t(p)T_pW$. We are now ready to show that $dV \equiv -\omega$ on U. Pick $x \in U$, $w \in T_xM$. Let t, p be such that $x = \gamma_p(t)$, $p \in W$, so $x \in W_t$. Let $v \in T_pM$ be such that $d\theta_t(p)v = w$. Define $y(s) = d\theta_s(p)v$, so $y(s) \in T_{\gamma_p(s)}W_s$ and $y(t) = w$. Then y satisfies the variational equation $\dot{y} = \frac{\partial X}{\partial x}(\gamma_p(s))y$, and λ_p satisfies the adjoint variational equation $\dot{\lambda}_p = -\frac{\partial H_X}{\partial x}(\gamma_p(s)) = -\lambda_p \frac{\partial X}{\partial x}(\gamma_p(s))$. This implies that the function $s \to \langle \lambda_p(s), y(s)\rangle$ is constant on $(c-\varepsilon, c+\varepsilon)$. If we choose $w \in T_xW_t$, then $v \in T_pW$, so $\langle \lambda_p(c), v\rangle = 0$, and therefore $\langle \omega(x), w\rangle = \langle \lambda_p(t), y(t)\rangle = \langle \lambda_p(c), y(c)\rangle = 0 = -\langle dV(x), w\rangle$, since V is constant on W_t. If we choose $w = \dot{\gamma}_p(t)$, then $v = \dot{\gamma}_p(c)$, so $\langle \omega(x), w\rangle = \langle \lambda_p(t), y(t)\rangle = \langle \lambda_p(c), y(c)\rangle = -1 = -\langle dV(x), w\rangle$, since the derivative of V along γ_p is obviously equal to 1. Since T_xM is the direct sum of T_xW_t and the span of $\dot{\gamma}_p(t)$, we conclude that $\omega(x) = -dV(x)$.

Since $\omega \equiv -dV$ and $H(x, \omega(x)) \equiv -\frac{1}{2}$, we have shown that V satisfies the Hamilton-Jacobi equation $H(x, dV(x)) = -\frac{1}{2}$ or, equivalently, $\|dV(x)\|_G = 1$. Now suppose that $\gamma : [0, \alpha] \to U$ is an E-arc parametrized by arc-length that goes from $\gamma_*(t_1)$ to $\gamma_*(t_2)$ for some t_1, t_2 in the interval $(c-\varepsilon, c+\varepsilon)$. Then

$$\begin{aligned} t_2 - t_1 &= V(\gamma_*(t_2)) - V(\gamma_*(t_1)) = \int_0^\alpha \frac{dV(\gamma(s))}{ds} ds \\ &= \int_0^\alpha \langle dV(\gamma(s)), \dot{\gamma}(s)\rangle ds \le \int_0^\alpha \|\dot{\gamma}(s)\|_G ds = \alpha\,, \end{aligned}$$

so the restriction of γ to any closed subinterval of $(c-\varepsilon, c+\varepsilon)$ minimizes length in the class of U-valued E-curves. Now let δ be such that (i) $0 < \delta < \varepsilon$, and (ii) the sub-Riemannian distance from $\gamma([c-\delta, c+\delta])$ to the complement of U in M is larger than 2δ. Then the restriction of γ_* to the interval $[c-\delta, c+\delta]$ minimizes length in the class of all M-valued E-curves. ∎

Appendix D: Rigid sub-Riemannian arcs and local optimality

Rigid curves of rank-two distributions were studied by R. Bryant and L. Hsu in [7]. It is not hard to show that rigid curves necessarily are abnormal extremals. The purpose of this appendix is to discuss the relationship between optimality and rigidity. We will establish, by means of examples, the following facts:

1 *A rigid curve need not be locally optimal.* This disproves, at least for general —not necessarily regular— sub-Riemannian structures, a conjecture made by R. Montgomery in [19]. Our example involves a real-analytic sub-Riemannian structure in $\mathbb{R}^3$, associated to a two-dimensional, non-regular, bracket-generating distribution, and the arc which is rigid but not locally optimal is real analytic, so we show in fact that rigidity does not imply local optimality even for real analytic curves.

2 *An abnormal minimizer need not be rigid.* This will be done by exhibiting, on the space $\mathbb{R}^4$ endowed with a sub-Riemannian structure arising from a three-dimensional distribution, an abnormal sub-Riemannian minimizer that is not rigid.

3 *Every abnormal extremal that can be put in the normal form of Theorem 4 is locally rigid.* This yields in particular an alternative proof of the main theorem of Bryant and Hsu [7] —which says, in our terminology, that every regular abnormal extremal is locally rigid— for a slightly larger class of distributions (cf. Remark 2).

Our examples show that, although all rigid sub-Riemannian curves are abnormal extremals, and the abnormal extremals studied here — i.e. regular abnormal extremals for two-dimensional distributions— are both minimizing and rigid, there is no general relation between the concepts of rigidity and optimality.

We first review the definition of rigidity. Let M be an n-dimensional connected manifold and E be an m-dimensional distribution of M. For any two points p, q in M, and any interval $[a, b]$, let $\Omega_E^{a,b}(p, q)$ be the space of C^1 E-arcs $\gamma : [a, b] \to M$ that satisfy $\gamma(a) = p, \gamma(b) = q$, endowed with the C^1-topology. Let $\Omega_E(p, q) = \cup_{-\infty<a\leq b<+\infty}\Omega_E^{a,b}(p, q)$. A curve $\gamma : [a, b] \to M$ is *rigid* if there exists a C^1-neighborhood $\mathcal{U}$ of γ in $\Omega_E^{a,b}(\gamma(a), \gamma(b))$ such that every $\tilde{\gamma} \in \mathcal{U}$ is a reparametrization of γ. We say that γ is *locally rigid* if every point of $[a, b]$ lies in the interior —relative to $[a, b]$— of a subinterval $J \in [a, b]$ such that γ restricted to J is rigid.

Our first example will show that rigid curves in sub-Riemannian manifolds need not be locally optimal. We consider the 3-dimensional control system Σ given by:

$$\dot{x} = v \ , \ \ \dot{y} = u \ , \ \ \dot{z} = \varphi(x, y)^2 v \ . \tag{75}$$

where u, v are controls subject to the constraint $u^2 + v^2 \leq 1$, and $\varphi(x, y) = x^3(y - x^2)$. Clearly, Σ is of the form $\dot{X} = uf(X) + vg(X)$, where f and g are the vector fields on $\mathbb{R}^3$ given by

$$f = \frac{\partial}{\partial y} \ , \ g = \frac{\partial}{\partial x} + \varphi(x, y)^2 \frac{\partial}{\partial z} \, .$$

Moreover, $h \stackrel{\text{def}}{=} \text{ad}_g^6 \text{ad}_f^2(g) = 1440\frac{\partial}{\partial z}$. So the vector fields f, g and h are linearly independent at every point. Therefore, if we define $E(X)$ to be the linear span of $f(X)$ and $g(X)$, then E is a two-dimensional bracket-generating distribution on $\mathbb{R}^3$. On the other hand, E is not regular, since the dimension of the linear span of $f(X)$, $g(X)$ and $[f, g](X)$ is not constant. We define a metric on E by declaring $f(X)$ and $g(X)$ to be an orthonormal basis of $E(X)$ for each X.

We now let $\hat{\gamma} : \mathbb{R} \to \mathbb{R}^2$ be the curve uniquely characterized by the following conditions: (1) the locus of $\hat{\gamma}$ is the parabola $y = x^2$, (2) $\hat{\gamma}$ is parametrized by arc-length, (3) $\hat{\gamma}(0) = (0,0)$, and (4) $\hat{\gamma}$ goes from left to right. We write $\hat{\gamma}(t) = (x_\gamma(t), y_\gamma(t))$, and define a curve $\gamma : \mathbb{R} \to \mathbb{R}^3$ by just letting $\gamma(t) = (x_\gamma(t), y_\gamma(t), 0)$. The function x_γ is the solution of the initial value problem

$$\dot{x}_\gamma(t) = \frac{1}{\sqrt{1 + 4x_\gamma(t)^2}} \ , \quad x_\gamma(0) = 0 \ ,$$

and then y_γ is given by $y_\gamma(t) = x_\gamma(t)^2$. If we define $u_\gamma(t) = \dot{y}_\gamma(t)$, and $v_\gamma(t) = \dot{x}_\gamma(t)$, we have $u_\gamma(t)^2 + v_\gamma(t)^2 = 1$. Since the function $\varphi(x,y)$ vanishes for $x = x_\gamma(t)$, $y = y_\gamma(t)$, we see that the choice $z_\gamma(t) \equiv 0$ satisfies the equation $\dot{z}_\gamma(t) = \varphi(x_\gamma(t), y_\gamma(t))^2 v_\gamma$. Therefore γ is a trajectory of Σ. It is clear that γ is real analytic.

For $L > 0$, let γ_L be the portion of γ that goes from $p_L = (-L, L^2, 0)$ to $q_L = (L, L^2, 0)$. Then the domain of γ_L is the closed interval $I(L) = [-\tau(L), \tau(L)]$, where $\tau(L) = \int_0^L \sqrt{1+4t^2}dt$, so that $\tau(L) = L + \frac{2}{3}L^3 + O(L^5)$ as $L \to 0$. We now prove:

Proposition 14 *For every $L > 0$ the arc γ_L is rigid, but not time-optimal.*

PROOF. Notice first that the function $\dot{x}_\gamma$ is bounded below on the interval $I(L)$ by a positive constant c_L. Let $\delta : I(L) \to \mathbb{R}^3$ be an E-arc that goes from p_L to q_L and is such that $||\delta - \gamma||_{C^1} \leq \frac{c_L}{2}$. If $\delta(t) = (x(t), y(t), z(t))$, we then have $|\dot{x}(t) - \dot{x}_\gamma(t)| \leq \frac{c_L}{2}$, so the control $v = \dot{x}$ satisfies $v(t) \geq \frac{c_L}{2}$. On the other hand, $0 = z(\tau(L)) - z(-\tau(L)) = \int_{-\tau(L)}^{\tau(L)} \varphi(x(t), y(t))^2 v(t)dt$. Since $v(t) > 0$, the function φ must vanish identically along δ. But this means that the curve $\hat{\delta} : t \to (x(t), y(t))$ is entirely contained in the zero set of φ, which is the union of the straight line $x = 0$ and the parabola $y = x^2$. Since $\hat{\delta}$ goes from left to right (because $v(t) > 0$) and goes from $(-L, L^2)$ to (L, L^2), the only possibility is that, up to a time reparametrization, $\hat{\delta}$ coincides with the projection of γ on the (x, y) plane. But then δ and γ are equivalent modulo time reparametrization. So γ is rigid, as stated.

We now prove that γ_L is not time-optimal. We will construct, for each $L > 0$ and each ε such that $0 < \varepsilon \leq L$, a trajectory $\gamma_{L,\varepsilon}$ that also goes from p_L to q_L, and we will show that $\gamma_{L,\varepsilon}$ is shorter than γ_L if ε is small enough. We construct $\gamma_{L,\varepsilon}$ by first constructing a curve $\hat{\gamma}_{L,\varepsilon} : t \to (x_{\varepsilon,L}(t), y_{\varepsilon,L}(t))$ in $\mathbb{R}^2$, defined on an interval $I_\varepsilon(L) = [-\tau_\varepsilon(L), \tau_\varepsilon(L) + 4\hat{\varepsilon}]$, and parametrized by arc-length, that goes from $(-L, L^2)$ to (L, L^2). Here the number $\tau_\varepsilon(L)$ will be defined below, and $\hat{\varepsilon}$ is a positive number that will be chosen as a function of ε and L. We then define the third component $z_{\varepsilon,L} : I_\varepsilon(L) \to \mathbb{R}$ by letting $z(t) = \int_{-\tau(L)}^{t} \varphi(x_{\varepsilon,L}(s), y_{\varepsilon,L}(s))^2 v_{\varepsilon,L}(s)ds$, where $v_{\varepsilon,L} = \dot{x}_{\varepsilon,L}$. The resulting curve $\gamma_{L,\varepsilon} : t \to (x_{\varepsilon,L}(t), y_{\varepsilon,L}(t), z_{\varepsilon,L}(t))$ is then a trajectory of the control system Σ that starts at p_L and ends at a point $q_{\varepsilon,L}$. If $z_{\varepsilon,L}(\tau_\varepsilon(L)) = 0$, then $q_{\varepsilon,L}$ is in fact q_L.

We construct $\hat{\gamma}_{L,\varepsilon}$ in two stages. Starting from the plane curve $\hat{\gamma} : t \to (x_\gamma(t), y_\gamma(t))$, defined on $I(L)$, we first replace the portion that goes from $(-\varepsilon, \varepsilon^2)$ to $(\varepsilon, \varepsilon^2)$ by the straight-line segment σ_ε that joins these two points. This produces a slightly shorter curve, which we parameterize by arc-length, choosing the initial condition so that the curve goes through $(0, \varepsilon^2)$ at time 0. Then the new curve is defined on a time interval $[-\tau_\varepsilon(L), \tau_\varepsilon(L)]$. We then insert at the end a square $S_{\varepsilon,L}$, oriented clockwise, whose sides are parallel to the coordinate axes and have length $\hat{\varepsilon}$. Moreover, we choose $S_{\varepsilon,L}$ so that its first segment is horizontal and going from left to right. This implies that $S_{\varepsilon,L}$ lies below the curve $y = x^2$. Then the total length Λ_{new} of the new curve is $\Lambda_{new} = 2\tau_\varepsilon(L) + 4\hat{\varepsilon}$. The length of the segment σ_ε is equal to 2ε, whereas the length of the piece replaced by σ_ε is $2\varepsilon + \frac{4}{3}\varepsilon^3 + O(\varepsilon^5)$. So $\tau_\varepsilon(L) = \tau(L) - \frac{2}{3}\varepsilon^3(1 + o(1))$ as $\varepsilon \to 0$. Therefore Λ_{new} is equal to the length of γ_L plus Δ, where $\Delta = -\frac{4}{3}\varepsilon^3(1 + o(1)) + 4\hat{\varepsilon}$.

On the segment σ_ε, the inequalities $|x| \leq \varepsilon$ and $|y - x^2| \leq \varepsilon^2$ hold. Therefore φ^2 is bounded by ε^{10}. Since $0 < v_{\varepsilon,L}(t) \leq 1$, we conclude that the integral of $\varphi(x_{\varepsilon,L}, y_{\varepsilon,L})^2 v_{\varepsilon,L}$ over σ_ε, which is equal to $z_{\varepsilon,L}(\tau_\varepsilon(L))$, is bounded by $2\varepsilon^{11}$.

We will choose $\hat{\varepsilon}$ as a function of ε and L in such a way that $\hat{\varepsilon} \to 0$ as $\varepsilon \to 0$. The final value $z_{\varepsilon,L}(\tau_\varepsilon(L)+4\hat{\varepsilon})$ is equal to $z_{\varepsilon,L}(\tau_\varepsilon(L))+\int_{S_{\varepsilon,L}} \varphi^2 dx$. By Green's Theorem, the integral is equal —since $S_{\varepsilon,L}$ is oriented clockwise— to $2\int\int_{R_{\varepsilon,L}} \varphi\frac{\partial\varphi}{\partial y} dx\, dy$, where $R_{\varepsilon,L}$ is the region enclosed by $S_{\varepsilon,L}$. Since $\varphi(L, L^2) = 0$ and $\frac{\partial\varphi}{\partial y}(L, L^2) = L^3 > 0$, the double integral is negative, since $S_{\varepsilon,L}$ lies below the curve $y = x^2$, so that φ is negative on $R_{\varepsilon,L}$. Moreover, the Taylor expansion for φ near (L, L^2) is

$$\varphi(x, y) = -2L^4(x - L) + L^3(y - L^2) + \text{higher order terms}\,,$$

which implies that the double integral is equal to $-k\hat{\varepsilon}^3(1 + o(1))$ as $\hat{\varepsilon} \to 0$, where k is a positive (L-dependent) constant. More precisely, the double integral is equal to $-k\hat{\varepsilon}^3(1 + \psi(\hat{\varepsilon}))$, where $\psi(\hat{\varepsilon}) \to 0$ as $\hat{\varepsilon} \to 0$.

We now choose $\hat{\varepsilon}$ so that $k\hat{\varepsilon}^3(1+\psi(\hat{\varepsilon})) = z_{\varepsilon,L}(\tau_\varepsilon(L))$. (The existence of $\hat{\varepsilon}$ is obvious. Indeed, the function $\theta \to k\theta(1 + \psi(\theta^{1/3}))$ is of class C^1 near $\theta = 0$, and its derivative at $\theta = 0$ is k. Hence, for every μ close to 0 there exists a unique $\theta(\mu)$, also close to 0, such that $\theta(\mu)(1 + \psi(\theta(\mu)^{1/3})) = \mu$. If we let $\hat{\varepsilon} = \hat{\theta}(\mu)^{1/3}$, with $\mu = z_{\varepsilon,L}(\tau_\varepsilon(L))$, then $\hat{\varepsilon}$ satisfies the desired equation.)

Since $z_{\varepsilon,L}(\tau_\varepsilon(L))$ is bounded by $2\varepsilon^{11}$, we conclude that $\hat{\varepsilon}$ is bounded by $C\varepsilon^{\frac{11}{3}}$, where C is an L-dependent positive constant. With our choice of $\hat{\varepsilon}$, the integral of $\varphi^2 dx$ over $S_{\varepsilon,L}$ exactly cancels that over σ_ε. So the resulting trajectory $\gamma_{L,\varepsilon}$ goes from p_L to q_L.

We now show that if ε is small enough then $\gamma_{L,\varepsilon}$ is shorter than γ_L. Recall that the length of $\gamma_{L,\varepsilon}$ is equal to the length of γ_L plus Δ, where $\Delta = -\frac{4}{3}\varepsilon^3(1 + o(1)) + 4\hat{\varepsilon}$. Since $\hat{\varepsilon}$ is bounded by $C\varepsilon^{\frac{11}{3}}$, and $\frac{11}{3} > 3$, we have $\Delta < 0$ if ε is small enough, and our desired conclusion follows. ∎

Our next example shows that an abnormal minimizer need not be rigid in general. The example is essentially a modification of the one discussed in Section 2.3. Let (M, E, G) be the sub-Riemannian manifold of Section 2.3.

Let $\hat{M} = \mathbb{R}^4$ with coordinates $(x_1, \ldots, x_4)$ and take $\hat{E}$ to be the kernel of the 1-form $\hat{\omega} = x_1^2 dx_2 - (1 - x_1)dx_3$. Since $\hat{\omega}$ never vanishes, $\hat{E}$ is a smooth 3-dimensional distribution. The three vector fields

$$\hat{f} = \frac{\partial}{\partial x_1}\,, \quad \hat{g} = (1 - x_1)\frac{\partial}{\partial x_2} + x_1^2\frac{\partial}{\partial x_3}\,, \quad h = \frac{\partial}{\partial x_4} \tag{76}$$

form a global basis of sections of $\hat{E}$. Clearly $\hat{E}$ is bracket-generating. Let

$$\hat{G} = dx_1^2 + \psi(x_1)(dx_2^2 + dx_3^2) + dx_4^2\,, \quad \text{where} \quad \psi(x_1) = \frac{1}{(1 - x_1)^2 + x_1^4}\,,$$

so $\hat{f}$, $\hat{g}$ and h form an orthonormal basis of sections of $\hat{E}$.

Let $\gamma(t) = (0, t, 0, 0)$ for $t \in \mathbb{R}$. Then γ is $\hat{E}$-admissible. From Proposition 1 we can conclude that the restriction $\gamma\lceil[a, b]$ is a minimizer if $b - a \leq \frac{2}{3}$. (To see this, let $\tilde{\gamma} : [\alpha, \beta] \to \mathbb{R}^4$ be another $\hat{E}$-arc that goes from $p = (0, a, 0, 0)$ to $q = (0, b, 0, 0)$. Let δ, $\tilde{\delta}$ be the projections of γ, $\tilde{\gamma}$ to the first three coordinates. Then δ and $\tilde{\delta}$ are E-arcs going from $(0, a, 0)$ to $(0, b, 0)$. Proposition 1 implies that δ is a minimizer for (M, E, G). So $\|\delta\|_G \leq \|\tilde{\delta}\|_G$. Clearly, $\|\tilde{\delta}\|_G \leq \|\tilde{\gamma}\|_{\hat{G}}$ and $\|\delta\|_G = \|\gamma\|_{\hat{G}}$. So $\|\gamma\|_{\hat{G}} \leq \|\tilde{\gamma}\|_{\hat{G}}$.) It is easy to see that γ is an abnormal minimizer. We show that γ is not rigid. For each integer $j > 0$, take a smooth real valued function θ^j on $[a, b]$ such that $\theta^j(a) = \theta^j(b) = 0$, and $\|\dot{\theta}^j\|_{\sup} \leq 1/j$. Then the $\hat{E}$-arc γ^j given by $\gamma_j(t) = (0, t, 0, \theta^j(t))$ for $a \leq t \leq b$ goes from p to q. Clearly, $\gamma^j \to \gamma$ and $\dot{\gamma}^j \to \dot{\gamma}$ uniformly on $[a, b]$, which shows that γ is not rigid. ∎

Finally we discuss the rigid curves of rank-two distributions. We prove:

Theorem 14 *Let E be a two-dimensional distribution on a manifold M. Then all the regular abnormal extremals of E are locally rigid.*

PROOF. In view of Theorem 4, it suffices to prove

Lemma 5 *Let $n \geq 3$, and let U be an open subset of $\mathbb{R}^n$. Let φ, $\psi_1, \psi_2, \psi_4, \ldots, \psi_n$, $\eta_1, \eta_3, \ldots, \eta_n$ be smooth functions on U, and let E be the distribution on U spanned by*

$$X = \varphi(x)\frac{\partial}{\partial x_1}\,, \quad Y = \frac{\partial}{\partial x_2} + x_1 \sum_{k=1}^{n} \psi_k(x)\frac{\partial}{\partial x_k}\,, \tag{77}$$

where $\psi_3 = x_1\eta_1 + x_3\eta_3 + \ldots + x_n\eta_n$. Let $x^* : [a,b] \to U$ be an E-arc given by $x^*(t) = (0, \bar{x}_2 + t - a, 0, \ldots, 0)$. Assume that $\eta_1(x^*(t)) \neq 0$ for $a \leq t \leq b$. Then x^* is locally rigid.

PROOF. Let $K = \{x^*(t) : a \leq t \leq b\}$, so K is a compact subset of U. Let $V \subseteq U$ be open, and such that $K \subseteq V$ and the closure $\bar{V}$ of V is compact and contained in U. We may assume that V is small enough so that $\eta_1(x) \neq 0$ and $1 + x_1\psi_2(x) > 0$ on $\bar{V}$. Without loss of generality we assume that $\eta_1(x) > 0$ on $\bar{V}$. Let

$$\begin{aligned}
C_0 &= \inf\left\{\frac{\eta_1(x)}{1 + x_1\psi_2(x)} : x \in V\right\}, \\
C_1 &= \sup\left\{\frac{|\eta_i(x)|}{1 + x_1\psi_2(x)} : x \in V, i = 3, \ldots, n\right\}, \\
C_2 &= \sup\left\{\frac{|\psi_i(x)|}{1 + x_1\psi_2(x)} : x \in V, i = 3, \ldots, n\right\}.
\end{aligned}$$

Let $\delta = \min\{\|x - y\| : x \in K, y \in \mathbb{R}^n - V\}$. Then $\delta > 0$. Let $\tilde{\tau} = \min\left\{\frac{\delta}{2}, \frac{2C_0}{9nC_1C_2}\right\}$.

Now let $a \leq t_1 \leq t_2 \leq b$ be such that $\tau = t_2 - t_1 < \tilde{\tau}$. Let γ be the restriction of x^* to the interval $[t_1, t_2]$. We show that γ is rigid. Assume that $\xi : [t_1, t_2] \to U$ is an E-arc that goes from $\gamma(t_1)$ to $\gamma(t_2)$ and satisfies $\|\dot{\xi}(t) - \dot{\gamma}(t)\| < \frac{1}{2}$ for all $t \in [t_1, t_2]$. Then $\|\dot{\xi}(t)\| < 2$, and therefore ξ is contained in V.

Write $\xi(t) = (\xi_1(t), \ldots, \xi_n(t))$. Then we have $\frac{1}{2} \leq \dot{\xi}_2(t) \leq \frac{3}{2}$ for all $t \in [t_1, t_2]$. Now since ξ is E-admissible, we have

$$\dot{\xi}_k(t) = \frac{\xi_1(t)\psi_k(\xi(t))\dot{\xi}_2(t)}{1 + \xi_1(t)\psi_2(\xi(t))}$$

for $k \geq 3$. Therefore $|\xi_k(t)| \leq \frac{3C_2}{2}\int_{t_1}^{t_2}|\xi_1(s)|ds$ for $k = 3, \ldots, n$. Since $\psi_3(\xi) = \xi_1\eta_1(\xi) + \sum_{k=3}^n \xi_k\eta_k(\xi)$, we have

$$\begin{aligned}
0 &= \xi_3(t_2) - \xi_3(t_1) \\
&= \int_{t_1}^{t_2} \frac{\xi_1^2(s)\eta_1(\xi(s))\dot{\xi}_2(s)}{1+\xi_1(s)\psi_2(\xi(s))}ds + \sum_{k=3}^{n}\int_{t_1}^{t_2}\frac{\xi_1(s)\xi_k(s)\eta_k(\xi(s))\dot{\xi}_2(s)}{1+\xi_1(s)\psi_2(\xi(s))}ds \\
&\geq \frac{C_0}{2}\int_{t_1}^{t_2}\xi_1^2(s)ds - \frac{9nC_1C_2}{4}\Big(\int_{t_1}^{t_2}|\xi_1(s)|ds\Big)^2 \\
&\geq \frac{2C_0 - 9n\tau C_1C_2}{4}\int_{t_1}^{t_2}\xi_1^2(s)ds\,.
\end{aligned}$$

Therefore $\xi_1 \equiv 0$, which implies that ξ is a reparametrization of γ. ∎

Appendix E: A nonoptimality proof

We prove that the nonsmooth abnormal extremal γ of Section 9.6 is not locally optimal. It suffices to show that for every $L > 0$ the restriction γ_L of γ to $[-L, L]$ is not optimal. In view of the symmetry of E and the metric with respect to the dilations $D_r : (x_1, x_2, x_3, x_4, x_5, x_6) \to (rx_1, rx_2, r^2x_3, r^3x_4, r^3x_5, r^4x_6)$, it suffices to prove that γ_1 is not optimal. We let $\xi : [-1, 1] \to \mathbb{R}^2$ denote the projection of γ_1 to the first two coordinates. We will construct, for $\varepsilon > 0$, ε small, a polygonal curve ξ_ε in $\mathbb{R}^2$ going from $(0, -1)$ to $(1, 0)$ such that (i) ξ_ε is shorter than ξ, and (ii) ξ_ε is the projection of an E-curve γ_ε in $\mathbb{R}^6$ that goes from $(0, -1, 0, 0, 0, 0)$ to $(1, 0, 0, 0, 0, 0)$. Actually, ξ_ε will also depend on a 5-tuple $\delta = (\delta_0, \ldots, \delta_4)$ of positive numbers, which will later be chosen to depend on ε, but will be $o(\varepsilon)$ (in fact, $O(\varepsilon^{4/3})$) as $\varepsilon \to 0$. To define the ξ_ε, let a_1, a_2, a_3, a_4 be four fixed numbers to be chosen later, such that $-1 < a_1 < 0$, $-1 < a_2 < 0$, $0 < a_3 < 1$, $0 < a_4 < 1$. Let $a = \min(-a_1, -a_2, a_3, a_4)$. For $p_1, \ldots, p_n \in \mathbb{R}^2$, use $p_1 \to p_2 \to \ldots \to p_n$ to denote the polygonal curve obtained by concatenating the straight-line segments from p_1 to p_2, p_2 to p_3, $\ldots$, p_{n-1} to p_n. For $\varepsilon < \frac{1}{3}$, define ξ_ε^1 by modifying ξ as follows: the portion of ξ that goes from $(0,-\varepsilon)$ to $(3\varepsilon,0)$ is replaced with the polygonal $(0,-\varepsilon)\rightarrow(\varepsilon,0)\rightarrow(\varepsilon,\frac{\varepsilon}{4})\rightarrow(3\varepsilon,\frac{\varepsilon}{4})\rightarrow(3\varepsilon,0)$. Then observe that, if $3\varepsilon<a$, and we let $p_0=(\frac{\varepsilon}{2},\frac{-\varepsilon}{2})$, $p_i=(0,a_i)$ for $i=1,2$, $p_i=(a_i,0)$ for $i=3,4$, then ξ^1_ε goes through the five points $p_0,\ldots,p_4$. We then define ξ_ε as the modification of ξ_ε obtained by inserting five squares $S_0,\ldots,S_4$ at the points $p_0,\ldots,p_4$. Here S_i is the square $p_i\rightarrow q_i\rightarrow r_i\rightarrow s_i\rightarrow p_i$ where we let, using $e_1=(1,0)$, $e_2=(0,1)$:

$$\begin{aligned}
q_0 &= p_0+\delta_0e_1\,,\ r_0=q_0-\delta_0e_2\,,\ s_0=r_0-\delta_1e_1\,,\\
q_1 &= p_1+\delta_1e_2\,,\ r_1=q_1-\delta_1e_1\,,\ s_1=r_1-\delta_1e_2\,,\\
q_2 &= p_2-\delta_2e_1\,,\ r_2=q_2+\delta_2e_2\,,\ s_2=r_2+\delta_2e_1\,,\\
q_3 &= p_3+\delta_3e_1\,,\ r_3=q_3+\delta_3e_2\,,\ s_3=r_3-\delta_3e_1\,,\\
q_4 &= p_4+\delta_4e_2\,,\ r_4=q_4+\delta_4e_1\,,\ s_4=r_4-\delta_4e_2\,.
\end{aligned}$$

Then —writing $|\delta|=\delta_0+\ldots+\delta_4$— ξ_ε has length $2+4|\delta|+(\sqrt{2}+\frac{1}{2}-2)\varepsilon$. Since the δ_i will be $o(\varepsilon)$, this will guarantee that ξ_ε is shorter than ξ for small ε.

Now let γ_ε be the unique E-curve that projects down to ξ_ε and starts at $(0,-1,0,0,0,0)$. Let $(1,0,\nu_1,\nu_2,\nu_3,\nu_4)$ be the coordinates of the terminal point of γ_ε. Our desired conclusion will follow if we show that the functions δ_i can be chosen so that $\nu_1=\ldots=\nu_4=0$. Expressing the ν_i as functions of all six variables $\varepsilon,\delta_0,\ldots,\delta_4$, an elementary computation shows that

$$\begin{aligned}
\nu_1 &= \delta_1^2-\delta_2^2+\delta_3^2-\delta_4^2-\delta_0^2\,, &(78)\\
\nu_2 &= 2a_3\delta_3^2-2a_4\delta_4^2+\mu_2(\varepsilon,\delta) &(79)\\
\nu_3 &= 2a_1\delta_1^2-2a_2\delta_2^2+\mu_3(\varepsilon,\delta) &(80)\\
\nu_4 &= -a_1\delta_1^3+a_2\delta_2^3+a_3\delta_3^3-a_4\delta_4^3+\frac{\varepsilon^4}{6}+\mu_4(\varepsilon,\delta)\,, &(81)
\end{aligned}$$

where the μ_i are real-analytic functions that satisfy

$$\mu_i(\varepsilon,\delta)=O(\varepsilon^3+\varepsilon\delta_0^2+|\delta|^3)\ \text{ for }\ i=2,3\,, \qquad (82)$$

$$\mu_4(\varepsilon,\delta) = O(|\delta|^4 + \varepsilon^2\delta_0^2) \tag{83}$$

as $\varepsilon, \delta \to 0$. Write $\varepsilon = \rho^3$, $\delta_i = \rho^4\Delta_i$. Then the conditions $\nu_1 = \ldots = \nu_4 = 0$ are equivalent to

$$\begin{aligned}
\Delta_1^2 - \Delta_2^2 + \Delta_3^2 - \Delta_4^2 &= \Delta_0^2\,, &(84)\\
2a_3\Delta_3^2 - 2a_4\Delta_4^2 + \rho\theta_2(\rho,\delta) &= 0\,, &(85)\\
2a_1\Delta_1^2 - 2a_2\Delta_2^2 + \rho\theta_3(\rho,\delta) &= 0\,, &(86)\\
-a_1\Delta_1^3 + a_2\Delta_2^3 + a_3\Delta_3^3 - a_4\Delta_4^3 + \frac{1}{6} + \rho\theta_4(\rho,\delta) &= 0\,, &(87)
\end{aligned}$$

where $\theta_i(\rho,\delta) = \rho^{-9}\mu_i(\rho^3,\delta)$ for $i = 2,3$, $\theta_4(\rho,\delta) = \rho^{-13}\mu(\rho^3,\delta)$. Then the θ_i are still analytic near the origin, in view of (82) and (83). Fix the value of Δ_4, say by taking $\Delta_4 = 1$. For $(\Delta_1,\Delta_2,\Delta_3)$ in the set $C = \{(x_1,x_2,x_3) : x_1^2 + x_3^2 > 1 + x_2^2\}$, define Δ_0 as an analytic function of $(\Delta_1,\Delta_2,\Delta_3)$ with positive values, via (84). Then the left-hand sides of (85), (86), (87) are analytic functions of $\Delta_1,\Delta_2,\Delta_3$, and ρ. Pick a solution $(\bar\Delta_1,\bar\Delta_2,\bar\Delta_3) \in C$ of (85), (86), (87) for $\rho = 0$, such that $\bar\Delta_1 \neq \bar\Delta_2$. (For example, we may choose $a_1 = -\frac{1}{3}$, $a_2 = -\frac{1}{12}$, $a_3 = \frac{1}{108}$, $a_4 = \frac{1}{12}$, $\bar\Delta_1 = 1$, $\bar\Delta_2 = 2$, $\bar\Delta_3 = 3$.) The Jacobian determinant with respect to $(\Delta_1,\Delta_2,\Delta_3)$ of the left-hand sides of (85), (86), (87) is equal, at $(\bar\Delta_1,\bar\Delta_2,\bar\Delta_3)$, to $48a_1a_2a_3\bar\Delta_1\bar\Delta_2\bar\Delta_3(\bar\Delta_1 - \bar\Delta_2)$, which is $\neq 0$. Then the Implicit Function Theorem implies that the system (85), (86), (87) — with $\Delta_4 \equiv 1$, $\Delta_0 > 0$ defined by (84), and $\delta = \rho^4\Delta$— has, for small ρ, a solution $(\Delta_1(\rho),\Delta_2(\rho),\Delta_3(\rho))$ that depends analytically on ρ and is equal to $(\bar\Delta_1,\bar\Delta_2,\bar\Delta_3)$ for $\rho = 0$. Letting $\delta_i(\varepsilon) = \varepsilon^{4/3}\Delta_i(\varepsilon^{1/3})$, all our requirements are satisfied. ∎

References

[1] C. Bär, *Carnot-Caratheodory Metriken,* Diplomarbeit, Bonn, 1988.

[2] L.D. Berkovitz, *Optimal Control Theory*, Springer-Verlag, New York, 1974.

[3] J.M. Bismut, *Large Deviations and the Malliavin Calculus*, Birkhäuser, 1984.

[4] V.G. Boltyanskii, "Sufficient conditions for optimality and the justification of the Dynamic Programming Principle," *SIAM J. Control* **4** (1966), pp. 326-361.

[5] R.W. Brockett, "Control Theory and Singular Riemannian Geometry," in *New Directions in Applied Mathematics* (P.J. Hilton and G.S. Young, eds.), Springer-Verlag (1981), pp. 11-27.

[6] R.W. Brockett, "Nonlinear Control Theory and Differential Geometry," in *Proc. 1983 Int. Congress of Mathematicians*, Vol. 2, North-Holland (1984), pp. 1357-1368.

[7] R. Bryant, L. Hsu, "Rigidity of Integral Curves of Rank 2 Distributions," *Inventiones Mathematicae* **114** (1993), pp. 435-461.

[8] W.L. Chow, "Uber Systeme Von Linearen Partiellen Differentialgleichungen Erster Ordnung", *Math. Ann.* **117** (1940-1941), pp. 98-105.

[9] B. Gaveau, "Principe de moindre action, propagation de la chaleur et estimées sous elliptiques sur certains groupes nilpotents," *Acta Math.* **139** (1977), pp. 94-153.

[10] Z. Ge, "Horizontal Path Spaces and Carnot-Carathéodory Metrics," *Pacific Journal of Mathematics* **161**, No. 2 (1993), pp. 255-286.

[11] Z. Ge, "On the Global Geometry of Sub-Riemannian 3-Manifolds, I," preprint.

[12] U. Hamenstädt, "Some Regularity Theorems for Carnot-Carathéodory Metrics," *J. of Diff. Geometry* **32** (1990), pp. 819-850.

[13] B. Kaskosz and S. Lojasiewicz Jr., "A Maximum Principle for generalized control systems," *Nonlinear Anal. TMA* **9** (1985), pp. 109-130.

[14] I. Kupka, "Abnormal extremals," preprint, 1992.

[15] E.B. Lee and L. Markus, *Foundations of Optimal Control Theory*, Wiley, New York, 1968.

[16] W.S. Liu and H.J. Sussmann, "Abnormal Sub-Riemannian Minimizers," in *Differential Equations, Dynamical Systems and Control Science*, K.D. Elworthy, W.N. Everitt, and E.B. Lee Eds., Lecture Notes in Pure and Applied Mathematics, Vol. 152. M. Dekker, New York, 1993. Pages 705 - 716. (IMA technical report #1059, November, 1992.)

[17] R. Montgomery, "Abnormal Minimizers," to appear in *SIAM J. Control and Opt.*

[18] R. Montgomery, "Characteristics, Singularities, and Abnormals for Systems Linear in Controls," 1993 preprint.

[19] R. Montgomery, "Survey of Singular Geodesics," 1992 preprint.

[20] L.S. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidze and E.F. Mischenko, *The Mathematical Theory of Optimal Processes*, Wiley, New York, 1962.

[21] E.D. Sontag, "Universal nonsingular controls," *Systems and Control Letters* **19** (1992), pp. 221-224.

[22] R. Strichartz, "Sub-Riemannian Geometry," *J. Diff. Geom.* **24** (1986), pp. 221-263.

[23] R. Strichartz, "Corrections to 'Sub-Riemannian Geometry'," *J. Diff. Geom.* **30** (1989), pp. 595-596.

[24] H. J. Sussmann, "Orbits of families of vector fields and integrability of distributions," *Trans. Amer. Math. Soc.* **180** (1973), pp. 171-188.

[25] H. J. Sussmann, "A Cornucopia of Abnormal Sub-Riemannian Minimizers. Part I: The Four-Dimensional Case," IMA technical report #1073, December 1992.

[26] H. J. Sussmann, "An Introduction to the Coordinate-Free Maximum Principle," 1993 preprint.

[27] T. J. Taylor, "Some Aspects of Differential Geometry Associated with Hypoelliptic Second Order Operators," *Pacific Journal of Mathematics* **136**, No. 2 (1989) pp. 355-378.

[28] T. J. Taylor, "Off Diagonal Asymptotics of Hypoelliptic Diffusion Equations and Singular Riemannian Geometric," *Pacific Journal of Mathematics* **136**, No. 2 (1989) pp. 379-399.

[29] M. Zhitomirskii, *Typical Singularities of Differential 1-Forms and Pfaffian Equations*, American Mathematical Society Translations of Mathematical Monographs, Vol. 113, 1992.

Wensheng Liu
Department of Mathematics
Rutgers University
New Brunswick, NJ 08903
E-mail: wliu@hilbert.rutgers.edu

Héctor J. Sussmann
Department of Mathematics
Rutgers University
New Brunswick, NJ 08903
E-mail: sussmann@hamilton.rutgers.edu