A CORNUCOPIA OF FOUR-DIMENSIONAL ABNORMAL SUBRIEMANNIAN MINIMIZERS[1]

Héctor J. Sussmann[2]

Department of Mathematics, Rutgers University, New Brunswick, NJ 08903, U.S.A.
E-mail: sussmann@hamilton.rutgers.edu

"The skull seems broken as with some big weapon, but there's no weapon at all lying about, and the murderer would have found it awkward to carry it away, unless the weapon was to small to be noticed."
"Perhaps the weapon was too big to be noticed," said the priest, with an odd little giggle.
Gilder looked round at this wild remark, and rather sternly asked Brown what he meant.
"Silly way of putting it, I know," said Father Brown apologetically. "Sounds like a fairy tale. But poor Armstrong was killed with a giant's club, a great green club, too big to be seen, and which we call the earth. He was broken against this green bank we are standing on."
"How do you mean?" asked the detective quickly.
Father Brown turned his moon face up to the narrow façade of the house and blinked hopelessly up. Following his eyes, they saw that right at the top of this otherwise blind back quarter of the building, an attic window stood open.
"Don't you see," he explained, pointing a little awkwardly like a child, "he was thrown down from there?"

G.K. Chesterton, "The Three Tools Of Death," in *The Innocence of Father Brown*, The Father Brown Omnibus, Dodd, Mead & Co., New York (1983), p. 117.

ABSTRACT. We study in detail the local optimality of abnormal sub-Riemannian extremals for a completely arbitrary sub-Riemannian structure on a four-dimensional manifold, associated to a two-dimensional bracket-generating regular distribution. Using a technique introduced in earlier work with W. Liu, we show that large collections of simple (i.e. without double points) nondegenerate extremals exist, and are always uniquely locally optimal. In particular, we prove that the simple abnormal extremals parametrized by arc-length foliate the space (i.e. through every point there passes exactly one of them) and they are *all* local minimizers. Under an extra nondegeneracy assumption, these abnormal extremals are strictly abnormal (i.e. are not normal). (In the forthcoming paper [6] with W. Liu we show that in higher dimensions there are large families of "nondegenerate abnormal extremals" that are local minimizers as well. In dimension 3, for a regular distribution there are no nontrivial abnormal extremals at all, but if the distribution is not regular then, generically, there are two-dimensional surfaces that are foliated by abnormal extremals, all of which turn out to be local minimizers.) This adds up to a picture which is rather different from the one that appeared to emerge from previous work by R. Montgomery and I. Kupka, in which an example of an abnormal extremal for a nonregular distribution in $\mathbb{R}^3$ was studied and shown to be locally optimal with great effort, by means of a very long and laborious argument, and then this example was used to produce a similar one for a regular distribution in $\mathbb{R}^4$. All this may have given the impression that abnormal extremals are hard to find, and that proving them to be minimizers is an arduous task that can only be accomplished in some very exceptional cases. Our results show that abnormal extremals exist aplenty, that most of them are local minimizers, and that in some widely studied cases, such as regular distributions on $\mathbb{R}^4$, this is in fact true for *all of them.*

[1] This work was done while the author was a visitor at Institute for Mathematics and its Applications (I.M.A.), University of Minnesota, Minneapolis, Minnesota 55455, U.S.A. A preprint of this paper, entitled "A cornucopia of abnormal subriemannian minimizers, Part I: the four-dimensional case," is # 1073 of the I.M.A. Preprint series, December 1992.

[2] Supported in part by the National Science Foundation under Grant DMS92-02554.

§1. Introduction .

Ever since the early work of Brockett [2] and Strichartz [9], [10] on sub-Riemannian geometry, it has been clear that sub-Riemannian minimizers fall into two not mutually exclusive categories, namely, the "normal" and "abnormal" extremals. Normal extremals are obviously smooth, and satisfy equations that in many ways resemble those of Riemannian geodesics. (In [9], it was stated that sub-Riemannian minimizers are necessarily smooth, and this was derived from an assertion equivalent to the proposition that all minimizers are normal extremals. Subsequently, it was noticed that the proof of this assertion involved an invalid application of the Pontryagin Maximum Principle, and that a truly correct analysis based on this result from Optimal Control Theory implied the possibility that a minimizer might be "abnormal." In [10], it was pointed out that the result of [9] remained valid for a very restrictive class of sub-Riemannian manifolds, namely, those that obey the "strong bracket-generating condition.")

Until recently, it was not clear whether strictly abnormal extremals that actually are minimizers can exist. ("Strictly abnormal" means "abnormal and not normal," cf. below.) This question was answered in recent work by R. Montgomery [7], who gave one example of a sub-Riemannian structure in $\mathbb{R}^3$, associated to a two-dimensional subbundle E of the tangent bundle, for which there exists a strictly abnormal uniquely optimal extremal. (We call an admissible trajectory *optimal* if it minimizes length among all admissible trajectories with the same initial and terminal points, and *uniquely optimal* if it is the only optimal trajectory joining these two points, up to reparametrization of the time interval.) Montgomery's optimality proof is rather lengthy and involved, making it desirable to find simpler ways of establishing the result. I. Kupka provided in [3] a different proof, also quite lengthy, based on a detailed analysis of the solutions of the differential equation defining the normal extremals. The Montgomery-Kupka examples are for a two-dimensional distribution in $\mathbb{R}^3$ which of necessity cannot be regular (the definition of a "regular distribution" is given below), since regular 2-dimensional distributions in $\mathbb{R}^3$ are strongly bracket generating and hence have no abnormal extremals. But, starting from these examples, one can construct (essentially by adding an extra variable) examples of minimizing strictly abnormal extremals for a regular distribution in $\mathbb{R}^4$. However, due to the extreme complexity of the proofs, these results have failed to yield a true understanding of the real reason why the particular abnormal extremals considered there happen to be optimal, and have created the impression that abnormal extremals are very hard to find, and that proving them to be optimal may only be possible in some very exceptional situations, and may require very hard work and a large amount of luck.

In this note we shall attempt to correct that impression, by showing that abnormal extremals exist in large numbers, that most of them are optimal, and that there is a very simple technique —essentially due to W. Liu— for proving optimality for a very wide broad range of situations. To make our point clear, we will concentrate on the most dramatic case, which also happens to be the situation that is universally recognized as the simplest, namely, that of a *four-dimensional* manifold M with a sub-Riemannian metric arising

¿from a *two-dimensional regular distribution* E. We will show that the following facts are *always*[3] true:

1. There is a line subbundle L of E such that the abnormal extremals are exactly the integrals curves of L.
2. All the abnormal extremals that are parametrized by arc-length and are simple (i.e. contain no loops) are locally optimal.
3. Optimality can be proved by a simple technique involving elementary inequalities.
4. All these abnormal extremals are actually strictly abnormal, provided that a simple generic condition (stated below) is satisfied.

In other words: to find optimal abnormal extremals for the simplest regular case (i.e. a 2-dimensional distribution in $\mathbb{R}^4$), one need not think hard and wonder where to look and how to select an example. The examples are everywhere, they are *all* locally optimal, and the proof of this fact just involves some elementary inequalities.

A crucial difference between our point of view and that of previous authors who have studied the problem is that we make systematic use of a control-theoretic approach, and in particular work with systems of vector fields and use properties of vector fields and Lie brackets to make appropriate choices of coordinate charts, rather than carry out calculations in terms of differential forms. It is our belief that the vector field formulations are more natural and geometric, and in addition are also better for effective calculation. In our view, the present paper provides support for this assertion. We hope the reader will be persuaded that abnormal extremals, which have appeared somewhat mysterious to geometers eager to pursue the analogy with Riemannian geometry, are not at all surprising to a mathematician who operates from an Optimal Control perspective, since a routine application of the Pontryagin Maximum Principle leads to them immediately. Similarly, when one uses the language of vector fields and Lie brackets to translate the conclusions obtained from the Maximum Principle into useful information, one is led directly to the canonical forms for abnormal extremals derived below, which lend themselves to an easy optimality proof.

The techniques of this paper can also be easily applied to study nondegenerate abnormal extremals in higher dimensions, and the abnormal extremals that arise in dimension 3 for nonregular generic distributions. This will be done in [6].

[3] We emphasize that this is ALWAYS true, not just “generically” or “almost always.”

§2. Sub-Riemannian Manifolds and Abnormal Extremals: a Brief Review .

If M is a C^∞ manifold, and $p \in M$, we use T_pM, T_p^*M to denote, respectively, the tangent and cotangent spaces of M at p, and TM, T^*M to denote the tangent and cotangent bundles of M. If $\lambda \in T_p^*M$, $v \in T_pM$, we write $\lambda(v)$, $\langle \lambda, v \rangle$ or, simply, λv, to denote the value at v of the linear functional λ. A subbundle E of TM is sometimes called a *distribution* on M. *A nonholonomic subbundle* (also known as a *bracket-generating distribution*) is a subbundle E of TM such that the Lie algebra $L(E)$ of vector fields generated by the global C^∞ sections of E has the *full rank property*, i.e. satisfies $\{X(p) : X \in L(E)\} = T_pM$ for all $p \in M$.

If E is a C^∞ subbundle of TM, we use $\Gamma(E)$ to denote the set of all C^∞ sections of E defined on open subsets of M. For a positive integer k, we let $\Gamma^k(E)$ denote the set of all vector fields X such that the domain of X is an open subset of M, and X is a linear combination of iterated brackets of degree $\leq k$ of members of $\Gamma(E)$. For $p \in M$, we let $E^k(p)$ denote the set $\{X(p) : X \in \Gamma^k(E)\}$. We write $\mu_E^k(p) = \dim E^k(p)$. The subbundle E is called *regular* if for every k the integer $\mu_E^k(p)$ is independent of p.

An *E-admissible arc* is an absolutely continuous curve γ on M, defined on some compact interval $[a, b]$, such that $\dot\gamma(t) \in E(\gamma(t))$ for almost all $t \in [a, b]$. If E is nonholonomic and M is connected, then any two points in M can be joined by an E-admissible arc.

A C^∞ *Riemannian metric* on E is a C^∞ 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. A *sub-Riemannian structure* on a manifold M is a pair (E, G) where E is a nonholonomic C^∞ subbundle of TM and G is a C^∞ Riemannian metric on E. A *sub-Riemannian manifold* is a triple (M, E, G) such that M is a C^∞ manifold and (E, G) is a sub-Riemannian structure on M. We call a sub-Riemannian manifold (M, E, G) *regular* if the subbundle E is regular. 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. If $p \in M$, $v \in E(p)$, then the *length* $||v||_G$ of v is the number $G_p(v, v)^{1/2}$. The *length* $||\gamma||_G$ of an E-admissible arc $\gamma : [a, b] \to M$ is the integral $\int_a^b ||\dot\gamma(t)||_G dt$. If $p, q \in M$, then the infimum of the lengths of all the E-admissible curves γ that go from p to q is the *distance* from p to q, and is denoted by $d_G(p, q)$. If M is connected and E is nonholonomic, then $d_G(p, q) < \infty$ for all p, q, and $d_G : M \times M \to \mathbb{R}$ is a metric whose associated topology is the one of M. An E-admissible curve $\gamma : [a, b] \to M$ such that $d_G(\gamma(a), \gamma(b)) = ||\gamma||_G$ is called a *minimizer*.

An E-admissible curve 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 $\tau(t) = \int_a^t ||\dot\gamma(s)||_G ds$, so 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 $\gamma(t_2) = \gamma(t_1)$. So we can define $\tilde\gamma : [0, ||\gamma||_G] \to M$ by letting $\tilde\gamma(s) = \gamma(t)$ if $\tau(t) = s$. Then, if $s_1 < s_2$, and $s_i = \tau(t_i)$ for $i = 1, 2$, the points $\tilde\gamma(s_1)$ and $\tilde\gamma(s_2)$ can be joined by the restriction of γ to the interval $[t_1, t_2]$, whose G-length is $s_2 - s_1$. So $d_G(\tilde\gamma(s_1), \tilde\gamma(s_2)) \leq s_2 - s_1$. If $\hat G$ is a Riemannian metric on M (i.e. a metric defined on the whole tangent bundle TM) that extends G, then the $\hat G$-distance $d_{\hat G}(\tilde\gamma(s_1), \tilde\gamma(s_2))$ is *a* *fortiori* $\leq s_2 - s_1$. So $\tilde{\gamma}$ is Lipschitz as a map into $(M, d_{\hat{G}})$. Clearly, $\gamma = \tilde{\gamma} \circ \tau$. Since $\tilde{\gamma}$ is Lipschitz and τ is integrable, we have $\int_0^s ||\dot{\tilde{\gamma}}(\sigma)||_G d\sigma = \int_0^t ||\dot{\gamma}(\theta)||_G d\theta = s$, if $s = \tau(t)$. So $||\dot{\tilde{\gamma}}(s)|| = 1$ for almost all s. Therefore $\tilde{\gamma}$ is parametrized by arc length.

In particular, every minimizer γ is equivalent modulo reparametrization to an arc γ^* which is parametrized by arc length and is *time-optimal* for the control problem Σ (i.e. goes from its initial point p to its terminal point q in time not greater than that of any other trajectory of Σ that goes from p to q), where Σ is the class of all E-admissible arcs δ that satisfy $||\dot{\delta}(t)|| \leq 1$ for almost all t. (It is clear that, if γ is a minimizer, then the arc $\tilde{\gamma}$ constructed above is a solution of the minimum time problem.) Conversely, it is easy to see that, if γ is a solution of the minimum time problem, then γ is a minimizer parametrized by arc length. So the class of solutions of the minimum time control problem for Σ coincides with the class of minimizers that are parametrized by arc length.

The solutions of the minimum time problem satisfy a necessary condition for optimality given by the Pontryagin Maximum Principle (cf. [1], [4], [8]). A trajectory that satisfies this condition is called a *Pontryagin extremal.* To state the condition, we need to define, for an arbitrary E-admissible curve, what is meant by an *H-minimizing adjoint vector* along γ. We would like to say that an H-minimizing adjoint vector is an adjoint vector that is H-minimizing. Unfortunately, the concept of an adjoint vector, by itself, is not intrinsic, since it depends on choosing an orthonormal basis of sections of E. So we will first define the concept of an adjoint vector relative to a basis $\mathbf{f}$ (or $\mathbf{f}$-*adjoint vector*), and then the concept of an $\mathbf{f}$-H-minimizing adjoint vector. The latter turns out to be intrinsic, and this will give us the desired definition.

We first define what is meant by an $\mathbf{f}$-*adjoint vector*, assuming that γ is such that the set $\gamma([a,b])$ is entirely contained in an open set Ω and that $\mathbf{f} = (f_1, \ldots, f_m)$ is a basis of smooth sections of E on Ω. Under this assumption we can express γ as a trajectory of the control system $\dot{x} = u_1 f_1(x) + \ldots + u_m f_m(x)$, that is, as a solution of the differential equation $\dot{x}(t) = u_1(t) f_1(x(t)) + \ldots + u_m(t) f_m(x(t))$ for some m-tuple $(u_1, \ldots, u_m)$ of real-valued integrable functions on $[a,b]$. (Notice that the control functions u_i are uniquely determined by γ and $\mathbf{f}$, since the f_i are linearly independent at each point.) If in addition Ω is also the domain of a coordinate chart $\kappa = (\kappa_1, \ldots, \kappa_n)$ of M, then we define an $\mathbf{f}$-*adjoint vector along* γ to be a vector-valued absolutely continuous function $\lambda^\kappa : [a,b] \to \mathbb{R}_n$ that satisfies the adjoint equations of the Pontryagin Maximum Principle:

$$\dot{\lambda}^\kappa(s) = -\sum_{i=1}^{m} u_i(s) \Big(\lambda^\kappa(s) \frac{\partial f_i}{\partial \kappa}(\gamma(s)) \Big) \ .$$

(We use $\mathbb{R}^n$, $\mathbb{R}_n$ to denote, respectively, the spaces of n-dimensional column and row vectors. We also use the familiar convention of thinking of n-tuples of coordinates of points or of components of tangent vectors as columns, so the f_i are columns of functions and the $\frac{\partial f_i}{\partial \kappa}(x)$ are square matrices. The columns of $\frac{\partial f_i}{\partial \kappa}$ are the partial derivatives $\frac{\partial f_i}{\partial \kappa_\ell}$, for $\ell = 1, \ldots, n$. Then λ^κ is a row vector, and $\lambda^\kappa \frac{\partial f_i}{\partial \kappa}$ is therefore a row vector as well.)

It is well known —and easy to prove— that, if we think of $\lambda^\kappa(t)$ as the n-tuple of components with respect to κ of a covector $\lambda(t)$ at $\gamma(t)$ then, if the adjoint equation holds for one coordinate system κ on Ω, it necessarily must hold on any other coordinate system on Ω. So, if λ is a *field of covectors along* γ (that is, λ is a section of the pullback $\gamma^*(T^*M)$ or, equivalently, λ is a mapping defined on $[a,b]$ such that, for each t, $\lambda(t)$ belongs to $T^*_{\gamma(t)}M$ of M at $\gamma(t)$), then the property that λ is a solution of the adjoint equation associated to $\mathbf{f}$ along γ is well defined, if γ is contained in an open set Ω and $\mathbf{f} = (f_1,\ldots,f_m)$ is a basis of sections on Ω. Any field of covectors along γ that has this property will be called an $\mathbf{f}$-*adjoint vector along* γ. We use $\mathrm{Adj}_{\mathbf{f}}(\gamma)$ to denote the set of all $\mathbf{f}$-adjoint vectors along γ. It is clear that $\mathrm{Adj}_{\mathbf{f}}(\gamma)$ is an n-dimensional linear space —where $n = \dim M$— and that, for each $t \in [a,b]$, the map $\lambda \to \lambda(t)$ establishes an isomorphism between $\mathrm{Adj}_{\mathbf{f}}(\gamma)$ and $T^*_{\gamma(t)}M$. A field λ of covectors along γ such that $\lambda(t) \neq 0$ for some $t \in [a,b]$ will be called *nontrivial.* If $\lambda \in \mathrm{Adj}_{\mathbf{f}}(\gamma)$ is nontrivial, then $\lambda(t) \neq 0$ for all $t \in [a,b]$.

We call a $\lambda \in \mathrm{Adj}_{\mathbf{f}}(\gamma)$ $\mathbf{f}$-*H-minimizing* if in addition the control functions u_i are such that, for almost every t, the vector $(u_1(t),\ldots,u_m(t))$ minimizes the linear function $\mathbf{u} = (u_1,\ldots,u_m) \to \mathbf{u}\Lambda_{\mathbf{f}}(s)$ (where $\Lambda_{\mathbf{f}}(s)$ is the column vector with components $\langle\lambda(s), f_1(\gamma(s))\rangle,\ldots,\langle\lambda(s), f_m(\gamma(s))\rangle$) on the unit ball $\{v : v_1^2 + \ldots + v_m^2 \leq 1\}$. Equivalently, λ is $\mathbf{f}$-H-minimizing if, for each t:

$$\sum_{i=1}^m \Big\langle \lambda(t), f_i(\gamma(t))\Big\rangle^2 > 0 \text{ implies } u_i(t) = \frac{-\langle\lambda(t), f_i(\gamma(t))\rangle}{\sqrt{\sum_{j=1}^m \Big\langle \lambda(t), f_j(\gamma(t))\Big\rangle^2}} \text{ for } i = 1,\ldots,m\,.$$

(When $\sum_{j=1}^m \Big\langle \lambda(t), f_j(\gamma(t))\Big\rangle^2 = 0$, the $u_i(t)$ can be arbitrary, provided only that they satisfy the constraint $\sum_{i=1}^m u_i(t)^2 \leq 1$.)

We can now go one step further, and drop the dependence on the basis $\mathbf{f}$, for *orthonormal bases* $\mathbf{f}$ and $\mathbf{f}$-H-minimizing adjoint vectors. To do this, it suffices to show that, if $\mathbf{f} = (f_1,\ldots,f_m)$ and $\mathbf{g} = (g_1,\ldots,g_m)$ are two orthonormal bases of sections on Ω, and λ is $\mathbf{g}$-adjoint and $\mathbf{g}$-H-minimizing, then it is also $\mathbf{f}$-adjoint and $\mathbf{f}$-H-minimizing. Write $g_i = \sum \alpha_{ij} f_j$, where the α_{ij} are smooth functions on Ω. Then the matrix $A(x) = (\alpha_{ij}(x))_{i,j=1,\ldots,m}$ is orthogonal for each x. On any subinterval I of $[a,b]$ such that $\gamma(I)$ is contained in the image of a chart κ, we can write

$$\dot\gamma(s) = \sum_i v_i(s) g_i(\gamma(s)) = \sum_j u_j(s) f_j(\gamma(s))$$

where $u_j(s) = \sum_i v_i(s)\alpha_{ij}(\gamma(s))$. Then the adjoint equation relative to $\mathbf{g}$ says

$$\dot\lambda(s) = -\sum_{i=1}^m v_i(s)\lambda(s)\frac{\partial g_i}{\partial\kappa}(\gamma(s))\ .$$

But

$$\frac{\partial g_i}{\partial\kappa}(x) = \frac{\partial(\sum_j \alpha_{ij} f_j)}{\partial\kappa}(x) = \sum_j \alpha_{ij}(x)\frac{\partial f_j}{\partial\kappa}(x) + \sum_j \frac{\partial\alpha_{ij}}{\partial\kappa}(x) f_j(x)\ .$$

The term $\sum_j \frac{\partial \alpha_{ij}}{\partial \kappa}(x) f_j(x)$ is a square matrix whose columns are the partial derivatives $\sum_j \frac{\partial \alpha_{ij}}{\partial \kappa_\ell}(x) f_j(x)$. Suppose we evaluate this matrix at $x = \gamma(s)$, then left-multiply by the row vector $\lambda(s)$, multiply by $v_i(s)$, and sum over i. The result is a row vector whose components are the inner products $\langle \mathbf{v}(s), B_\ell(\gamma(s))\Lambda_{\mathbf{f}}(s)\rangle$, where $B_\ell = \frac{\partial A}{\partial \kappa_\ell}$, $\mathbf{v} = (v_1, \ldots, v_m)^\dagger$ and, for any m-tuple $\mathbf{h} = (h_1, \ldots, h_m)$ of vector fields, we write

$$\Lambda_{\mathbf{h}}(s) \overset{\text{def}}{=} (\lambda(s)h_1(\gamma(s)), \ldots, \lambda(s)h_m(\gamma(s)))^\dagger \,.$$

(Here we are using † to denote matrix transpose.) Since $A(x)$ is orthogonal, the matrices $B_\ell(x)A(x)^\dagger$ are skew-symmetric. Since $\lambda(s)g_i(\gamma(s)) = \sum_j \alpha_{ij}(\gamma(s))\lambda(s)f_j(\gamma(s))$, the vectors $\Lambda_{\mathbf{f}}(s)$ and $\Lambda_{\mathbf{g}}(s)$ are related by $\Lambda_{\mathbf{g}}(s) = A(\gamma(s))\Lambda_{\mathbf{f}}(s)$, i.e. by $\Lambda_{\mathbf{f}}(s) = A(\gamma(s))^\dagger \Lambda_{\mathbf{g}}(s)$. Then

$$\langle \mathbf{v}(s), B_\ell(\gamma(s))\Lambda_{\mathbf{f}}(s)\rangle = \langle \mathbf{v}(s), B_\ell(\gamma(s))A(\gamma(s))^\dagger \Lambda_{\mathbf{g}}(s)\rangle \,.$$

Since λ is $\mathbf{g}$-H-minimizing, the vector $\Lambda_{\mathbf{g}}(s)$ is of the form $\rho(s)\mathbf{v}(s)$ for some scalar $\rho(s)$. Therefore

$$\langle \mathbf{v}(s), B_\ell(\gamma(s))A(\gamma(s))^\dagger \Lambda_{\mathbf{g}}(s)\rangle = \rho(s)\langle \mathbf{v}(s), B_\ell(\gamma(s))A(\gamma(s))^\dagger \mathbf{v}(s)\rangle \,,$$

which is equal to zero because $B_\ell(\gamma(s))A(\gamma(s))^\dagger$ is skew-symmetric. So we have shown that

$$\langle \mathbf{v}(s), B_\ell(\gamma(s))\Lambda_{\mathbf{f}}(s)\rangle = 0 \,.$$

Using this, the equation $\frac{\partial g_i}{\partial \kappa}(x) = \sum_j \alpha_{ij}(x)\frac{\partial f_j}{\partial \kappa}(x) + \sum_j \frac{\partial \alpha_{ij}}{\partial \kappa}(x) f_j(x)$ yields, if we evaluate at $x = \gamma(s)$, left-multiply by $\lambda(s)$, multiply by $v_i(s)$, and sum over i:

$$\sum_{i=1}^m v_i(s)\lambda(s)\frac{\partial g_i}{\partial \kappa}(\gamma(s)) = \sum_{ij} v_i(s)\alpha_{ij}(\gamma(s))\lambda(s)\frac{\partial f_j}{\partial \kappa}(\gamma(s)) \ ,$$

that is

$$\sum_{i=1}^m v_i(s)\lambda(s)\frac{\partial g_i}{\partial \kappa}(\gamma(s)) = \sum_{j=1}^m u_j(s)\lambda(s)\frac{\partial f_j}{\partial \kappa}(\gamma(s)) \ ,$$

since $\sum_i v_i(s)\alpha_{ij}(\gamma(s)) = u_j(s)$. Therefore

$$\dot\lambda(s) = -\sum_{j=1}^m u_j(s)\lambda(s)\frac{\partial f_j}{\partial \kappa}(\gamma(s)) \ ,$$

so λ is $\mathbf{f}$-adjoint as well.

Since λ is $\mathbf{g}$-H-minimizing, the vector $(v_1(t), \ldots, v_m(t))$ minimizes the linear functional $\mathbf{v} \to \mathbf{v}\Lambda_{\mathbf{g}}(s)$ on the unit ball of $\mathbb{R}^m$. But then, since $A(\gamma(s))$ is orthogonal, and $u_j(s) = \sum_i v_i(s)\alpha_{ij}(\gamma(s))$, the vector $(u_1(t), \ldots, u_m(t))$ minimizes the linear functional $\mathbf{u} \to \mathbf{u}\Lambda_{\mathbf{f}}(s)$ on the unit ball. So λ is $\mathbf{f}$-H-minimizing as well.

It is not hard to see that the t-derivative of the quantity $\sum_{i=1}^m \Big\langle \lambda(t), f_i(\gamma(t))\Big\rangle^2$ is equal to $2\sum_{i,j=1}^m u_j(t)\Big\langle \lambda(t), [f_j, f_i](\gamma(t))\Big\rangle \Big\langle \lambda(t), f_i(\gamma(t))\Big\rangle$. Since the vector with components $\langle \lambda(t), f_i(\gamma(t))\rangle$ is a scalar multiple of the vector with components $u_i(t)$, and the matrix $\Big(\Big\langle \lambda(t), [f_j, f_i](\gamma(t))\Big\rangle\Big)_{ij}$ is skew-symmetric, the derivative is in fact equal to zero, so the quantity $\sum_{i=1}^m \Big\langle \lambda(t), f_i(\gamma(t))\Big\rangle^2$ is constant. The identity $\Lambda_{\mathbf{g}}(s) = A(\gamma(s))\Lambda_{\mathbf{f}}(s)$ then implies that this constant does not depend on the choice of orthonormal basis.

Summarizing, we have shown that

the condition that a field of covectors along an E-admissible trajectory γ is both **f**-*adjoint and* **f**-H-*minimizing for some orthonormal basis* **f** *of sections of E is in fact independent of the basis* **f**. *This condition makes therefore intrinsic sense, and when we want to verify it or use it we can choose the orthonormal basis arbitrarily. Moreover, the number $\sum_{i=1}^{m}\Big\langle\lambda(t), f_i(\gamma(t))\Big\rangle^2$ is in fact independent of t, and its value does not depend on the basis either.*

We call a field λ of covectors an *adjoint H-minimizing covector* along an E-admissible trajectory $\gamma : [a,b] \to M$ if, whenever $\Omega \subseteq M$ is open, $\mathbf{f} = (f_1, \ldots, f_m)$ is an orthonormal basis of sections of E defined on Ω, and I is a subinterval of $[a,b]$ such that $\gamma(I) \subseteq \Omega$, then the restriction of λ to I is an **f**-adjoint **f**-H-minimizing field of covectors. We call λ *normal* (resp. *abnormal*) if the constant $\sum_{i=1}^{m}\Big\langle\lambda(t), f_i(\gamma(t))\Big\rangle^2$ is > 0 (resp. $= 0$). An E-admissible trajectory γ for which there exists a nonzero adjoint H-minimizing covector λ will be called an *extremal*. If λ can be chosen to be normal (resp. abnormal), then γ is called a *normal* (resp. *abnormal*) extremal. Since there may exist more than one nonzero adjoint H-minimizing covector for a given trajectory, it is possible for an extremal to be both normal and abnormal. A *strictly abnormal extremal* is an abnormal extremal which is not also a normal extremal.

The Pontryagin Maximum Principle says, simply, that

Every minimizer is an extremal.

§3. Abnormal extremals in dimension 4 .

We now let (M, E, G) be a regular sub-Riemannian manifold, such that $\dim M = 4$ and $\dim E = 2$. We let Σ be the associated control problem, whose trajectories are the E-admissible curves γ such that $||\dot\gamma(t)|| \le 1$ for almost all t.

We necessarily have $\mu_E^1 = 2$, $\mu_E^2 = 3$, $\mu_E^3 = 4$. Moreover, if f and g are two smooth sections of E with domain $\Omega \subseteq M$ that are linearly independent at each point of Ω, then the three vectors $f(p)$, $g(p)$, $[f,g](p)$ span the space $E^2(p)$ —and are therefore linearly independent— for every $p \in \Omega$, whereas the five vectors $f(p)$, $g(p)$, $[f,g](p)$, $[f,[f,g]](p)$, $[g,[f,g]](p)$ span the tangent space T_pM for every $p \in \Omega$.

We now determine the abnormal extremals for (M, E, G). Call a smooth section $g \in \Gamma(E)$, with domain $\Omega_g \subseteq M$, an *abnormal infinitesimal generator* (AIG) if

1. $||g(p)||_G = 1$ for all $p \in \Omega_g$,

and

2. if $f \in \Gamma(E)$ is any smooth section of E with domain Ω_f, then the vectors $f(p)$, $g(p)$, $[f,g](p)$ and $[g,[f,g]](p)$ are linearly dependent for every $p \in \Omega_f \cap \Omega_g$.

It is easy to see that, if Ω is any nonempty open subset of M, and g_1, g_2 are smooth sections of E with domain Ω that satisfy Condition **2**, then g_1, g_2 must be linearly dependent at each point of Ω. (Indeed, if there is a $p \in \Omega$ such that $g_1(p)$ and $g_2(p)$ are independent, then the three vectors $g_1(p)$, $g_2(p)$, $[g_1, g_2](p)$ are linearly independent and each of the two 4-tuples $(g_1(p), g_2(p), [g_1, g_2](p), [g_2, [g_1, g_2]](p))$, $(g_1(p), g_2(p), [g_1, g_2](p), [g_1, [g_2, g_1]](p))$ is linearly dependent, as can be seen by applying Condition **2** to both g_1 and g_2. Since the first three vectors of these 4-tuples are independent, it follows that both $[g_2, [g_1, g_2]](p)$ and $[g_1, [g_2, g_1]](p)$ are linear combinations of $g_1(p)$, $g_2(p)$, and $[g_1, g_2](p)$. But this contradicts the fact that $g_1(p)$, $g_2(p)$, $[g_1, g_2](p)$, $[g_2, [g_1, g_2]](p)$, and $[g_1, [g_2, g_1]](p)$ span T_pM.)

On the other hand, given any point $p \in M$, we can construct an AIG on a neighborhood of p as follows. Start with any two linearly independent sections f, h, defined on a neighborhood Ω of p. Let ω be a smooth nowhere vanishing 1-form such that the annihilator $\{v \in T_qM : \omega(q)(v) = 0\}$ is exactly $E^2(q)$ for each $q \in \Omega$. Let $g = \alpha f + \beta h$ be a linear combination of f and h with C^∞ coefficients. Then, modulo members of $\Gamma^2(E)$, the brackets $[g, [f, g]]$ and $[g, [h, g]]$ are given by $\alpha\beta[f, [f, h]] + \beta^2[h, [f, h]]$ and $-\alpha^2[f, [f, h]] - \alpha\beta[h, [f, h]]$. So, if we choose $\alpha = \langle \omega, [h, [f, h]] \rangle$, $\beta = -\langle \omega, [f, [f, h]] \rangle$, we see that $\langle \omega, [g, [f, g]] \rangle \equiv \langle \omega, [g, [h, g]] \rangle \equiv 0$, so that $[g, [f, g]]$ and $[g, [h, g]]$ are both in $\Gamma^2(E)$. From this it follows easily that g satisfies Condition **2**. Moreover, g can never vanish, because $g(q) = 0$ would imply $\langle \omega, [h, [f, h]] \rangle(q) = \langle \omega, [f, [f, h]] \rangle(q) = 0$, from which it would follow that both $[h, [f, h]](q)$ and $[f, [f, h]](q)$ belong to $E^2(q)$, so $E^3(q) = E^2(q)$, which is a contradiction.

Finally, it is easy to see that, if a vector field $g \in \Gamma(E)$ satisfies Condition **2**, and φ is a smooth real-valued function on the domain of g, then φg satisfies **2** as well. Therefore we can modify g by multiplying it by the smooth function $||g||_G^{-1}$, and we obtain a vector field that satisfies both conditions **1** and **2**, i.e. an AIG.

So we have shown that every point p has a neighborhood Ω such that there is a $g \in \Gamma(E)$ with domain Ω which is an AIG. For a given Ω, g is obviously unique up to sign. Therefore there is a well defined line subbundle L of E, characterized by the fact that the fiber $L(p)$ at each point $p \in M$ is the linear span of $g(p)$, g being any AIG whose domain contains p. The AIG's are then exactly the smooth sections of L that have length 1 at each point. A global AIG, defined on all of M, may or may not exist, depending on whether the bundle L is orientable.

Now, let $\gamma : [a, b] \to M$ be an E-admissible curve parametrized by arc-length. Assume that γ is an abnormal extremal. Let λ be a nonzero field of covectors along γ which is an abnormal adjoint vector for γ. Let $t \in [a, b]$ and choose, near $p = \gamma(t)$, a basis of sections consisting of an AIG g and a smooth section $f \in \Gamma(E)$ that has length 1 and is orthogonal to g at each point. We can then express the curve $s \to \gamma(s)$, for s near t, as a solution of the system $\dot{x}(s) = u(s)f(x(s)) + v(s)g(x(s))$, where u and v are measurable functions such that $u(s)^2 + v(s)^2 = 1$. The adjoint equation for λ then implies that for every smooth vector field X the derivative of the function $s \to \langle \lambda(s), X(\gamma(s)) \rangle$ is the function $s \to u(s)\langle \lambda(s), [f, X](\gamma(s)) \rangle + v(s)\langle \lambda(s), [g, X](\gamma(s)) \rangle$. The abnormality condition says that the functions $s \to \langle\lambda(s), f(\gamma(s))\rangle$ and $s \to \langle\lambda(s), g(\gamma(s))\rangle$ vanish identically. Differentiating these two functions we see that $s \to u(s)\langle\lambda(s), [f,g](\gamma(s))\rangle$ and $s \to v(s)\langle\lambda(s), [f,g](\gamma(s))\rangle$ vanish identically. Since $u^2 + v^2 \neq 0$, it follows that $s \to \langle\lambda(s), [f,g](\gamma(s))\rangle \equiv 0$. One more differentiation implies the identity $u(s)\langle\lambda(s), [f,[f,g]](\gamma(s))\rangle + v(s)\langle\lambda(s), [g,[f,g]](\gamma(s))\rangle \equiv 0$. Since $\lambda(s)$ annihilates the vectors f, g and $[f,g]$, evaluated at $\gamma(s)$, and $\lambda(s) \neq 0$, it follows that the annihilator of $\lambda(s)$ is precisely the linear span of these three vectors, i.e. the subspace $E^2(\gamma(s))$. Therefore the vector $V(s) = u(s)[f,[f,g]](\gamma(s)) + v(s)[g,[f,g]](\gamma(s))$ belongs to $E^2(\gamma(s))$. On the other hand, since g is an AIG, it follows that $[g,[f,g]](\gamma(s)) \in E^2(\gamma(s))$. Since, at each point, the vectors $[f,[f,g]]$ and $[g,[f,g]]$ span E^3 modulo E^2, we can conclude that $[f,[f,g]](\gamma(s)) \notin E^2(\gamma(s))$. But then $V(s)$ can only belong to $E^2(\gamma(s))$ if $u(s) = 0$, in which case $v(s) = \pm 1$. Therefore γ is in fact L-admissible, that is, $\dot\gamma(t) \in L(\gamma(t))$ for each t.

Conversely, if $\gamma : [a,b] \to M$ is any L-admissible curve which is parametrized by arc length, then we claim that γ is an abnormal extremal. To see this, we must find a nowhere vanishing covector λ along γ which is an abnormal adjoint vector for γ. Pick a nonzero covector $\bar\lambda \in T^*_{\gamma(a)}M$ that annihilates $E^2(\gamma(a))$. We claim that there exists an abnormal adjoint vector λ along γ such that $\lambda(a) = \bar\lambda$. To see this, we pick a maximal subinterval I of $[a,b]$ such that $a \in I$ and there is an H-minimizing abnormal adjoint vector λ_I for the restriction of γ to I such that $\lambda_I(a) = \bar\lambda$. We claim that $I = [a,b]$. If this is not true, let $\tau = \sup I$, so $a \le \tau \le b$. Find an interval J that contains τ in its interior relative to $[a,b]$, and is such that $\gamma(J)$ is contained in an open set Ω on which there exist both a coordinate chart and an orthonormal basis $\mathbf{f} = (f,g)$ of sections of E, such that g is an AIG. Pick a $\theta \in I \cap J$. Solve the adjoint equation with initial condition $\lambda(\theta) = \lambda_I(\theta)$. A solution λ_J exists on J because the adjoint equation is linear with respect to λ, and this solution must agree with λ_I on $I \cap J$, by uniqueness. Our curve γ satisfies, on J, an equation $\dot\gamma(t) = v(t)g(\gamma(t))$, where $|v(t)| = 1$ a.e., because γ is L-admissible. If we let $f_1 = f$, $f_2 = g$, $f_3 = [f,g]$, $\varphi_i = \langle\lambda_J, f_i\rangle$, then the functions φ_i satisfy, for some functions ψ_i, the system of differential equations $\dot\varphi_1 = -v\varphi_3$, $\dot\varphi_2 = 0$, $\dot\varphi_3 = \sum_{i=1}^3 \psi_i\varphi_i$, where the third equation follows from the fact that $[g, f_3]$ is a linear combination of f_1, f_2 and f_3, because g is an AIG. Since λ_J is abnormal on $I \cap J$, the functions φ_1 and φ_2 vanish there. Then the equation $\dot\varphi_1 = -v\varphi_3$, implies that φ_3 also vanishes on $I \cap J$. By uniqueness, the φ_i vanish on J. So λ_J is abnormal on J. The field λ of covectors that agrees with λ_I on I and with λ_J on J is therefore an abnormal adjoint vector on $I \cup J$. The maximality of I then implies that $J \subseteq I$. Then $I = [a,b]$, and our conclusion follows.

It is clear that an L-admissible curve parametrized by arc-length satisfies, locally, an equation $\dot\gamma(t) = v(t)g(\gamma(t))$, where g is an AIG and the measurable function v takes values ± 1.

Let us call a curve *simple* if it has no double points, i.e. if it is a one-to-one map or, equivalently, if it contains no loops. We call γ *locally simple* if there is a $\delta > 0$ such that the restriction of γ to every interval of length $\le \delta$ is simple.

It is clear that a curve which is not simple cannot be time-optimal, since by removing a loop one gets a shorter curve with the same initial and terminal points. The equation $\dot\gamma(t) = v(t)g(\gamma(t))$, implies that γ is contained in an integral curve of g. More precisely, if we fix a t_0 in the domain of γ, and let $s \to \xi(s)$ be the integral curve of g which goes through $\gamma(t_0)$ when $s = 0$, then we have $\gamma(t) = \xi(V(t))$, where $V(t) = \int_{t_0}^{t} v(s)ds$. The function V is absolutely continuous and satisfies $|\dot V(s)| = 1$ for almost all s. If V is not one-to-one, then clearly γ is not simple. For γ to be simple, then V has to be one-to-one, and hence monotonic, which implies that $\dot V$ is either ≥ 0 a.e. or ≤ 0 a.e., and if we combine this observation with the fact that $|\dot V(s)| = 1$ a.e., we see that either $v \equiv 1$ or $v \equiv -1$. From this it follows easily that *the locally simple abnormal extremals parametrized by arc-length are precisely the curves that satisfy, locally, an equation of the form $\dot\gamma(t) = g(\gamma(t))$ or $\dot\gamma(t) = -g(\gamma(t))$, where g is an AIG.*

So we have proved:

THEOREM 1. *If (M, E, G) is a sub-Riemannian manifold such that $\dim M = 4$ and E is two-dimensional and regular, then there exists a line subbundle L of E such that the abnormal extremals parametrized by arc length are exactly the L-admissible curves parametrized by arc length. An abnormal extremal cannot be locally optimal unless it is simple. The locally simple abnormal extremals parametrized by arc length are precisely the curves that satisfy, locally, an equation of the form $\dot\gamma(t) = g(\gamma(t))$ or $\dot\gamma(t) = -g(\gamma(t))$, where g is an AIG. In particular, through every point there pass exactly two oriented (or one unoriented) locally simple abnormal extremals parametrized by arc-length.* ∎

§4. Optimality .

We now prove

THEOREM 2. *If (M, E, G) is a sub-Riemannian manifold such that $\dim M = 4$ and E is two-dimensional and regular, then every locally simple abnormal extremal is locally uniquely optimal.*

To prove Theorem 2, we pick a locally simple abnormal extremal $\gamma : [a, b] \to M$. 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 optimal. Let $p = \gamma(\bar t)$. Let Ω be an open subset of M that contains p and is such that on Ω there is an orthonormal basis (f, g) of sections of E such that g is an AIG.

Consider the map Φ defined by

$$\Phi(x_1, x_2, x_3, x_4) = pe^{x_3[f,[f,g]]}e^{x_4[f,g]}e^{x_2 g}e^{x_1 f} \quad ,$$

where we are using exponential notation for the flow of a vector field, and have the exponentials act on points on the right, so that $t \to xe^{tX}$ is the integral curve of the vector field X that goes through x at time $t = 0$. Since f, g, $[f, g]$ and $[f, [f, g]]$ are independent at p, the map Φ is well defined on a neighborhood of the origin in $\mathbb{R}^4$, and maps diffeomorphically some cube $C^4(\rho) = \{(x_1, x_2, x_3, x_4) : |x_i| < \rho \text{ for } i = 1, 2, 3, 4\}$ onto a neighborhood U of p in M. The inverse map Φ^{-1} defines a chart, with respect to which we are going to identify U with $C^4(\rho)$, so that p becomes the point $(0, 0, 0, 0)$. Clearly, f then just equals $\frac{\partial}{\partial x_1}$. Moreover, g is equal to $\frac{\partial}{\partial x_2}$ whenever $x_1 = 0$. So $g = \frac{\partial}{\partial x_2} + x_1 \sum_{i=1}^4 \psi_i \frac{\partial}{\partial x_i}$, where the ψ_i are smooth functions. Then

$$[f, g] = \sum_{i=1}^4 \psi_i \frac{\partial}{\partial x_i} + x_1 \sum_{i=1}^4 \frac{\partial \psi_i}{\partial x_1} \frac{\partial}{\partial x_i} \ .$$

In particular, $[f, g] = \sum_{i=1}^4 \psi_i \frac{\partial}{\partial x_i}$ whenever $x_1 = 0$. On the other hand, $[f, g] = \frac{\partial}{\partial x_4}$ whenever $x_1 = x_2 = 0$. So the functions ψ_1, ψ_2, ψ_3 and $\psi_4 - 1$ vanish when $x_1 = x_2 = 0$.

An easy calculation shows that the x_3-component of $[g, [f, g]]$ is equal, for some smooth function η, to the function $\frac{\partial \psi_3}{\partial x_2} - \psi_1 \psi_3 + x_1 \eta$. When $x_1 = 0$ this component is therefore just equal to $\frac{\partial \psi_3}{\partial x_2} - \psi_1 \psi_3$. On the other hand, the third component of $[f, g]$ is equal to ψ_3 when $x_1 = 0$, whereas the third components of f and g vanish when $x_1 = 0$. Since $[g, [f, g]]$ is in the linear span of f, g and $[f, g]$, we conclude that on the set defined by $x_1 = 0$ the function $\frac{\partial \psi_3}{\partial x_2}$ is equal to ψ_3 times a smooth function. Since ψ_3 vanishes when $x_1 = x_2 = 0$, it follows that $\psi_3 = 0$ whenever $x_1 = 0$. So ψ_3 is in fact equal to x_1 times a smooth function η. Then g has the form

$$g = \frac{\partial}{\partial x_2} + x_1 \sum_{i \in \{1,2,4\}} \psi_i \frac{\partial}{\partial x_i} + x_1^2 \eta \frac{\partial}{\partial x_3} \ .$$

This means that the trajectories of the restriction to U of the control system Σ are given by the equations:

$$\begin{aligned} \dot{x}_1 &= u + v x_1 \psi_1 \ , \\ \dot{x}_2 &= v(1 + x_1 \psi_2) \ , \\ \dot{x}_3 &= v x_1^2 \eta \ , \\ \dot{x}_4 &= v x_1 \psi_4 \ , \end{aligned}$$

where u, v are controls that are required to satisfy $u^2 + v^2 \leq 1$. Our locally simple abnormal extremal γ satisfies the above equations with $u(t) \equiv 0$, and either $v(t) \equiv 1$ or $v(t) \equiv -1$.

We are now in a situation nearly identical to that of [5], and an argument similar to the one given there will enable us to prove optimality.

§5. An optimality lemma .

In this section we prove a general optimality lemma that transforms the main technical idea of [5] into a widely applicable method. We state once and for all the general version of the lemma that will be used in [6], rather than the slightly weaker result that would suffice for the four-dimensional case considered here.

LEMMA. *Let $n \geq 3$, and let Ω 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 Ω, and let Σ be the control system*

$$\begin{aligned}
\dot{x}_1 &= u\varphi(x) + vx_1\psi_1(x) \ , \\
\dot{x}_2 &= (1 + x_1\psi_2(x))v \ , \\
\dot{x}_3 &= vx_1\psi_3(x) \ , \\
\dot{x}_4 &= vx_1\psi_4(x) \ , \\
&\ \vdots \\
\dot{x}_n &= vx_1\psi_n(x) \ ,
\end{aligned}$$

where

$$\psi_3 = x_1\eta_1 + x_3\eta_3 + \ldots + x_n\eta_n \ ,$$

and the controls u, v are subject to the constraint $u^2 + v^2 \leq 1$. Let $x^* : [a, b] \to \Omega$ 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.

PROOF. Let $K = \{x^*(t) : a \leq t \leq b\}$, so K is a compact subset of Ω. Let $U \subseteq \Omega$ be open, and such that $K \subseteq U$ and the closure $\bar{U}$ of U is compact and contained in Ω. 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 U. Let

$$\delta = \min\{||x - y|| \ : \ x \in K, y \in \mathbb{R}^n - U\} \ .$$

Then $\delta > 0$. Let $\hat{\tau}_1$ be such that $0 < \hat{\tau}_1 < \frac{\delta}{C_1}$. Then, if $x(\cdot) : [t_1, t_2] \to \Omega$ 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 U. We let

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

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

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

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 optimal. Assume that $\xi : [s_1, s_2] \to \Omega$ 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 $u_\xi, v_\xi : [s_1, s_2] \to \mathbb{R}$ such that $u_\xi(s)^2 + v_\xi(s)^2 \leq 1$ for a.e. s. We will show that $\sigma \geq \tau$, with equality holding iff $\xi(s) = \gamma(s - s_1 + t_1)$ 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 U, 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 $\alpha = \sigma - h(s_2)$, $\beta = \sup\{|\xi_1(s)| : s_1 \leq s \leq s_2\}$. We then have

$$\begin{aligned}\tau &= x_2^*(t_2) - x_2^*(t_1)\\ &= \xi_2(s_2) - \xi_2(s_1)\\ &= \int_{s_1}^{s_2} \Big(1 + \xi_1(s)\psi_2(\xi(s))\Big) v_\xi(s)ds\\ &\leq \int_{s_1}^{s_2} v_\xi(s)ds + \beta\sigma C_2\\ &= h(s_2) + \beta\sigma C_2\\ &= \sigma - \alpha + \beta\sigma C_2\ .\end{aligned}$$

On the other hand,

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

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

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

So

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

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

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

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

Also,
$$\left|\int_{s_1}^{s_2} \xi_1(s)^2v_\xi(s)\frac{\eta_1(\xi(s_1))-\eta_1(\xi(s))}{\eta_1(\xi(s_1))}ds\right| \le \frac{1}{4}\int_{s_1}^{s_2}\xi_1(s)^2ds \ ,$$
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
$$\int_{s_1}^{s_2}\xi_1(s)^2ds \le \beta^2\alpha + \left(\frac{1}{4} + (n-2)C_2C_3\sigma\right)\int_{s_1}^{s_2}\xi_1(s)^2ds \ .$$

Since $\sigma \le \hat{\tau}_3 \le \Big(4(n-2)C_2C_3\Big)^{-1}$, we have the bound
$$\int_{s_1}^{s_2}\xi_1(s)^2ds \le \beta^2\alpha + \frac{1}{2}\int_{s_1}^{s_2}\xi_1(s)^2ds \ ,$$

from which it follows that
$$\int_{s_1}^{s_2}\xi_1(s)^2ds \le 2\beta^2\alpha \ .$$
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
$$\int_{s_1}^{s_2}\xi_1(s)^2ds \ge \int_{I_1}\xi_1(s)^2ds + \int_{I_2}\xi_1(s)^2ds \ge \frac{2\beta^3}{3C_1} \ .$$
Combining the upper and lower bounds for $\int_{s_1}^{s_2}\xi_1(s)^2ds$ we get
$$\frac{2\beta^3}{3C_1} \le 2\beta^2\alpha \ .$$

Therefore
$$\beta \le 3C_1\alpha \ .$$

But then
$$\begin{aligned}\tau &\le \sigma - \alpha + \beta\sigma C_2 \\ &\le \sigma - \alpha + 3C_1C_2\sigma\alpha \\ &= \sigma + \Big(3C_1C_2\sigma - 1\Big)\alpha \ .\end{aligned}$$
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 $v_\xi = 1$ a.e. But then $u_\xi = 0$ a.e., and $\xi(s) = \gamma(s - s_1 + t_1)$. So γ is uniquely optimal. ▮

§6. End of the proof .

We now return to the locally simple abnormal extremal $\gamma : [a,b] \to M$ of §4, which goes through $p = (0,0,0,0)$ at time $\bar{t}$, and recall that we are trying to find a δ such that the restriction of γ to the interval $[a,b] \cap [\bar{t}-\delta, \bar{t}+\delta]$ is uniquely optimal. We observe that

$$\eta(0,0,0,0) \neq 0 \ .$$

(Indeed, $\eta(0,0,0,0)$ is the third component of $[f,[f,g]](0,0,0,0)$. Since the linear span of f, g and $[f,g]$ at $(0,0,0,0)$ is precisely the set of vectors whose third component vanishes, and $[f,[f,g]]$ does not belong to that linear span, our statement follows.)

It follows that the equations given at the end of §4 are exactly of the kind occurring in the lemma, and all the hypotheses of the lemma are satisfied, provided that we first restrict γ to a subinterval $[\tilde{a},\tilde{b}]$ of $[a,b]$ that contains $\bar{t}$ in its interior relative to $[a,b]$ and is small enough that $\eta(\gamma(t)) \neq 0$ for all $t \in [\tilde{a},\tilde{b}]$. Then we can apply the lemma and conclude that our trajectory is locally optimal for the system Σ restricted to the domain U. The following remark then concludes the proof.

REMARK. Local time-optimality is a "local" property in the following sense. Consider a control system Σ on a manifold M, subject only to the condition that the velocity is uniformly bounded on compact sets, that is, that for every compact subset K of M there is a constant C such that $||\dot{x}(t)|| \leq C$ for a.e. t, for every trajectory $x(\cdot)$ of Σ which is entirely contained in K. (Here $||\ldots||$ is the norm with respect to some Riemannian metric on M.) Let Ω be an open subset of M. Let Σ_Ω be the restriction of Σ to Ω. Let $\gamma : [a,b] \to \Omega$ be locally optimal for Σ_Ω. Then γ is locally optimal for Σ. To see this, pick a compact subset K of Ω whose interior contains the set $K_0 = \{\gamma(t) \, : \, a \leq t \leq b\}$. Because of the bound on the velocities, there is a $\delta_1 > 0$ such that every trajectory $\tilde{\gamma}$ of Σ that goes through a point of K_0 and is defined on an interval of length $\leq \delta_1$ is entirely contained in K. Since γ is locally optimal for Σ_Ω, there is a $\delta_2 > 0$ such that, if $a \leq t_1 \leq t_2 \leq b$, $t_2 - t_1 \leq \delta_2$, and $\tilde{\gamma} : [\tilde{t}_1, \tilde{t}_2] \to \Omega$ is another trajectory of Σ such that $\tilde{\gamma}(\tilde{t}_1) = \gamma(t_1)$ and $\tilde{\gamma}(\tilde{t}_2) = \gamma(t_2)$, then it follows that $\tilde{t}_2 - \tilde{t}_1 \geq t_2 - t_1$. If we take $\delta = \min(\delta_1, \delta_2)$, then it is clear that, if $a \leq t_1 \leq t_2 \leq b$, $t_2 - t_1 \leq \delta$, and $\tilde{\gamma} : [\tilde{t}_1, \tilde{t}_2] \to M$ is another trajectory of Σ such that $\tilde{\gamma}(\tilde{t}_1) = \gamma(t_1)$ and $\tilde{\gamma}(\tilde{t}_2) = \gamma(t_2)$, then $\tilde{t}_2 - \tilde{t}_1 \geq t_2 - t_1$. (Indeed, if $\tilde{t}_2 - \tilde{t}_1 < t_2 - t_1$, then it would follow that $\tilde{t}_2 - \tilde{t}_1 < \delta_1$, and therefore $\tilde{\gamma}$ is in fact contained in Ω. But then $\tilde{t}_2 - \tilde{t}_1 \geq t_2 - t_1$ because $\delta \leq \delta_2$.) ∎

§7. Strict abnormality .

We must now take care of the possibility that our abnormal extremals might be normal as well, i.e. might not be strictly abnormal. Suppose g is an AIG, f is a smooth section of E of unit length and orthogonal to g, and γ is an integral curve of g or of $-g$, i.e. a locally simple abnormal extremal. Then a normal H-minimizing covector λ along γ must satisfy $\langle\lambda, f\rangle = 0$, $\langle\lambda, g\rangle = -c$, where c is a strictly positive constant. Differentiation yields $\langle\lambda, [f,g]\rangle = 0$, and then again $\langle\lambda, [g,[f,g]]\rangle = 0$. Since g is an AIG, we have $[g,f,g]] = \mu f + \nu g + \rho[f,g]$ for some smooth functions μ, ν, ρ. But then $\langle\lambda, [g,[f,g]]\rangle = -\nu c$. If $\nu \neq 0$, this is a contradiction.

Let us call a point $p \in M$ *nice* if the vectors $g(p)$, $[f,g](p)$ and $[g,[f,g]](p)$ are linearly independent, where g is an AIG defined near p, and $f \in \Gamma(E)$ is of unit length and orthogonal to g. Then, if an abnormal extremal goes through a nice point and is not strict, we must necessarily have $\nu = 0$ at that point, contradicting the fact that the point is nice. So we have shown:

THEOREM 3. *If (M, E, G) is a sub-Riemannian manifold such that $\dim M = 4$ and E is two-dimensional and regular, then every locally simple abnormal extremal that goes through a point p where g, $[f,g]$ and $[g,[f,g]]$ are linearly independent is strictly abnormal.* ∎

§8. Conclusion: An Easy Recipe for Producing Millions of Strictly Abnormal Minimizers on General Four-Dimensional Sub-Riemannian Manifolds, Including Lie Groups with an Invariant Metric .

Just take any pair of vector fields f, g on a four-dimensional manifold M, such that f, g, $[f,g]$ and $[f,[f,g]]$ are linearly independent everywhere but $[g,[f,g]]$ is a linear combination $\mu f + \nu g + \rho[f,g]$ with smooth coefficients. Make sure that the function ν never vanishes. Define a sub-Riemannian structure on M by letting E be the span of f and g, and declaring f, g to be an orthonormal basis. Then the locally simple abnormal extremals parametrized by arc-length are precisely the integral curves of g. And all these locally simple abnormal extremals are strictly abnormal and locally uniquely optimal.

The regular four-dimensional examples of Montgomery [7], Kupka [3], and Liu and Sussmann [5] can all be obtained in this way.

One can also use the above construction to produce examples of locally uniquely optimal abnormal extremals for invariant sub-Riemannian metrics on Lie groups. It suffices to 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 $[f,g]$,
(iii) $[g,[f,g]]$ does not belong to the linear span of f and $[f,g]$.

(For example, 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.)

Then we can 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. The integral curves of g are then strictly abnormal locally uniquely optimal trajectories.

REFERENCES

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

[2] 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).

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

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

[5] W.S. Liu and H.J. Sussmann, "Abnormal Sub-Riemannian Minimizers," Institute for Mathematics and its Applications, University of Minnesota, IMA preprint series # 1059, October 1992. To appear in *Differential Equations, Dynamical Systems, and Control Science: A Festschrift in honor of Lawrence Markus*, K. D. Elworthy, W. N. Everitt and E. B. Lee Eds., M. Dekker, Inc., New York.

[6] W.S. Liu and H.J. Sussmann, "Shortest paths in Sub-Riemannian manifolds," in preparation.

[7] R. Montgomery, "Geodesics which do not Satisfy the Geodesic Equations," 1991 preprint.

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

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

[10] R. Strichartz, "Corrections to 'Sub-Riemannian Geometry'," *J. Diff. Geom.* 30, no. 2, 595-596, (1989).