Abnormal

Minimizers

Richard Montgomery

July 15, 1999

Abstract

This paper constructs the first example of a singular, abnormal minimizer for the problem of Lagrange with linear velocity constraints and quadratic definite Lagrangian, or equivalently, for an optimal control system linear in the controls, with k controls, n states and a running cost function which is quadratic positive definite in the controls. In the example $k = 2$, $n = 3$ and the system is completely controllable. The example is stable: if both the control law and cost are perturbed the singular minimizer persists. Its importance is due in part to the fact that it is a counterexample to a theorem which has appeared several times in the differential geometry literature. There the problem is called the problem of finding minimizing subRiemannian gedesics and it had been claimed that all minimizers were normal Pontrjagin extremals [22]. (If the number of states equals the number of controls the problem is that of finding Riemannian geodesics.) The main difficulty is proving minimality. To do this the length (cost) of the abnormal is compared with all competing normal extremals. A detailed asymptotic analysis of the differential equations governing the normals shows that they are all longer.

1 Introduction

Consider a system

$$\dot{q} = \Sigma_{i=1}^{k} u_i X_i(q),$$

linear in the k controls u_i. Here the point q evolves on an n-manifold Q, and we assume $k < n$. If Q is connected and if the vector fields X_i are bracket-generating then according to Chow's theorem we can find piecewise continuous controls $u_i(t)$ which steer between any two points of Q. Consider the optimal control problem of finding controls $u_i(t), 0 \leq t \leq 1$ which steer between two given points in such a way as to minimize the L^2-norm $\frac{1}{2}\int_0^1 \Sigma_i u_i^2 dt$ over all controls which steer between these points. If desired, this can be reformulated as a minimum time problem. Impose the bounds

$\Sigma_i u_i^2 \leq 1$ and find the controls joining the two points in the minimum possible time.

This problem can be viewed as a generalization of the problem of finding geodesics on a Riemannian manifold. Let $D_q = Span\{X_i(q) : i = 1, 2, \ldots, k\}$. We assume that the X_i are everywhere linearly independent. Then $D = \cup_{q \in Q} D_q$ forms a *distribution*, i.e. a linear subbundle of the tangent bundle TQ of Q. By declaring the X_i to be orthonormal we define an inner product $\langle \cdot, \cdot \rangle$ on the k-planes D_q. The pair $(D, \langle \cdot, \cdot \rangle)$ is called a *SubRiemannian metric* on Q. And the X_is form a *framing* of D.

Call an absolutely continuous path *horizontal*, *integrable* or a *D-curve* if its derivative lies in D wherever it is defined. The *length* of such a path is $\int ds$ where $ds = \sqrt{\langle \dot{\gamma}, \dot{\gamma} \rangle} dt$. The *distance* between two points is the infimum of the lengths of all horizontal curves joining them. Our problem is to find a horizontal curve joining the two given points whose length realizes the distance between them. We call such a path a minimizing geodesic.

The Pontrjagin maximum principle [22] provides necessary conditions for a curve to be a minimizing geodesic. The Hamiltonian equations which govern the **normal** Pontrjagin extremals will be called the *geodesic equations*. They are similar to the geodesic equations of Riemannian geometry, the principle difference being that the fiber-quadratic form defining the Hamiltonian has rank $k < n$. We will also say that a curve in Q satisfies the geodesic equations if it is the projection of a solution to the geodesic equations.

Our main result is the construction of a subRiemannian metric on $\mathbf{R}^3$ and a horizontal curve in this space which is a minimizing geodesic but which does not satisfying the geodesic equations. This curve is necessarily the projection of an abnormal extremal. The distribution on $\mathbf{R}^3$ is bracket generating. The curve is smooth.

The assertion that every minimizer is the projection of a normal extremal, that is, that every minimizing geodesics satisfies the subRiemannian geodesic equations has occured numerous times in the literature. See for example Rayner [24], Strichartz [25], Bär [1] , or Taylor [27]. Strichartz later retracted his claim [26]. Our result came out of attempts to understand this (false) assertion. Gaveau had proposed a counterexample to this assertion in 1977 but Brockett [6] found a fundamental flaw in his reasoning. For completeness' sake we have included an appendix containing Gaveau's example and Brockett's refutation of it.

A basic property of our curve is that it is rigid in the sense that it admits no piecewise-C^1 endpoint-preserving variations through D-curves. Having no variations, such a curve is automatically an extremal (in the C^1 topology) for any functional. Since this extremality has no relation to the function being minimized (length) there is no reason that the curve should solve its Euler-Lagrange equations which are the subRiemannian geodesic equations. Our main difficulty will be to show that our curve is in fact a minimum.

One of our main tools is a partial dictionary between subRiemannian geometry and gauge theory begun in [20]. This dictionary tells us that the subRiemannian geodesic equations for our example are identical to the equations of a charged particle travelling in the plane under the influence of a magnetic field determined by the distribution.

The article is organized as follows. §2 contains basic definitions and notation. §3 describes the counterexample and some of its properties. §4 gives a heuristic proof that it is a minimizing geodesic. §5 and 6 contain the proof, the guts of the proof being §5.3 and §5.4. In §7 we show that our example is stable.

Since this article was first written [1991] several other proofs of minimality have appeared. One is due to I. Kupka [17] of the U. of Toronto and is similar in spirit to ours. Another is due to W-S. Liu and H. Sussmann [18]and is based on an inequality.

2 Preliminaries

A subRiemannian [sR] structure on a smooth manifold Q consists of a non-integrable distribution $\mathcal{D} \subset TQ$ together with a smoothly varying inner product $\langle \cdot, \cdot \rangle_q$ on the fibers $\mathcal{D}_q$ of this distribution. In this section we will describe the basic notions and notations of sR geometry. For more detailed treatments see Strichartz [25], Brockett [6], Pansu [21], Hamenstädt [13], Rayner [24], Bär [1], Hermann [15], or Vershik and Gershkovich [12] . SubRiemannian structures have also been called "Carnot-Caratheodory metrics", "singular Riemannian metrics", and "non-holonomic Riemannian metrics".

A path in Q is said to be horizontal if it is locally rectifiable (for example piecewise differentiable) and if its derivatives, whenever defined, lie in $\mathcal{D}$. The *length* of a horizontal path γ is the integral over t of $\sqrt{\langle \frac{d\gamma}{dt}, \frac{d\gamma}{dt} \rangle_{\gamma(t)}} dt$. The *distance* $d(q_0, q_1)$ between two points q_0, q_1 in Q is the infimum of the lengths of the horizontal paths joining them.

Definition 1 *A path $\gamma : [a, b] \to Q$ is called a minimizing geodesic if it is the shortest horizontal path joining its endpoints.*

Definition 2 *A path $\gamma : I \subset \mathbf{R} \to Q$ is called a geodesic if it is locally a minimizing geodesic. In other words, each $t_0 \in I$ is contained in a nontrivial* *closed subinterval* $J \subset I$ *such that* γ *restricted to* J *is a minimizing geodesic. If, in addition* $d(\gamma(t_2), \gamma(t_1)) = |t_2 - t_1|$ *whenever* $|t_2 - t_1|$ *is sufficiently small then we say that* γ *is a unit speed geodesic.*

When do minimizing subRiemannian geodesics exist? . The distribution is said to be bracket generating (or to satisfy Hörmander's conditon) if every point q in Q has a neighborhood on which there is a local frame field $X_i, i = 1, \ldots, rank\mathcal{D}$ for $\mathcal{D}$ such that the X_i together with all of their iterated Lie brackets $[X_i, X_j], [X_i, [X_j, X_k]], \ldots$ span the tangent space at every point of this neighborhood. (This condition is independent of the choice of local frame.) A corollary to a classical theorem of Chow [8] and Rashevski [23] asserts that if $\mathcal{D}$ is bracket generating and Q is connected then any two points of Q can be joined by a horizontal path. In particular, it follows from this and the Arzela-Ascoli theorem that if $\mathcal{D}$ is bracket generating and if the inner product on $\mathcal{D}$ is the restriction of a complete Riemannian metric then any two points of Q can be joined by a minimizing subRiemannian geodesic.

A formal application of the method of Lagrange multipliers yields differential equations for the geodesics. They can also be derived from the Pontrjagin maximum principle [22], in which case they are the equations for the normal extremals. We will call them the subRiemannian geodesic equations. To define them set $a_{ij}(q) = \langle X_i, X_j \rangle_q$ where X_i is a frame for $\mathcal{D}$ as before. Let a^{ij} be the inverse matrix. Then

$$g = \Sigma a^{ij} X_i X_j$$

is independent of the choice of frame field and defines a fiber-quadratic form on the cotangent bundle T^*Q, that is, a symmetric covariant tensor of type $(2,0)$. (We can define g intrinsically by saying that it is a vector bundle map $T^*Q \to TQ$ which satisfies $g(q)^* = g(q)$, image(g) $= \mathcal{D}$, and $p(w) =< g(q)(p), w >$ whenever $p \in T^*_q Q$, and $w \in \mathcal{D}_q$.) The function $H : T^*Q \to \mathbf{R}$ defined by

$$H(q,p) = \frac{1}{2} g(q)(p,p)$$

is the Hamiltonian which defines the geodesic equations. (In the Riemannian case $\mathcal{D}$ is the entire tangent bundle and this is the usual kinetic energy which generates geodesic flow.) If we choose coordinates q^μ on Q and the corresponding induced coordinates q^μ, p_ν on T^*Q then

$$H = \frac{1}{2} g^{\mu\nu}(q) p_\mu p_\mu$$

where $g^{\mu\nu}(q)$ is a symmetric matrix of rank equal to the rank of $\mathcal{D}$ which represents the $(2,0)$ tensor. And the subReimannian geodesic equations are

$$\dot{q}^\mu = \Sigma g^{\mu\nu}(q) p_\nu \,;\, \dot{p}_\mu = -\frac{1}{2} \Sigma \frac{\partial g^{\alpha\beta}}{\partial q^\mu} p_\alpha p_\beta \tag{1}$$

By a slight abuse of language, we will call a curve $q(t)$ in Q a solution to the subRiemannian geodesic equations if it is the projection of a solution $(q(t), p(t))$ to this system of first order ordingary differential equations on T^*Q.

Lemma 1 (Rayner [24] Corollary 2.2; Hamenstädt [13] Corollary 5.9) *Every solution to the subRiemannian geodesic equations, equation 1 above, is a geodesic.*

Our counterexample is to the converse of this lemma.

Remark 1 *The lemma does not require any assumptions on the distribution. In particular, it need not be bracket generating. Rayner's proof proceeds by showing that in a neighborhood of any non-self intersecting solution γ to the sR geodesic equation there is an extension of the distribution's inner product to a Riemannian metric such that γ is a Riemannian geodesic relative to that extension.*

3 The Counterexample

Take Q to be $\mathbf{R}^3$ for our example. Let $\mathcal{D}$ be the distribution annihilated by a smooth one-form of the type

$$\Theta = dz + A(r)d\theta$$

where (r, θ, z) are standard cylindrical coordinates on Q. This means that a vector $(\dot{r}, \dot{\theta}, \dot{z})$ is in $\mathcal{D}(r, \theta, z)$ if and only if $\dot{z} = -A(r)\dot{\theta}$. We will require that A has a single nondegenerate maximum at $r = 1$. Thus $\frac{dA}{dr}|_{r=1} = 0$ and $\frac{d^2A}{dr^2}|_{r=1} < 0$. For example,

$$A = \frac{1}{2}r^2 - \frac{1}{4}r^4 \tag{2}$$

is a good choice and defines a smooth one-form on all of $\mathbf{R}^3$. We will refer to this choice as the "model case".

Let us check the bracket generating condition for $\mathcal{D}$. $Y_1 = \frac{\partial}{\partial r}$ and $Y_2 = \frac{\partial}{\partial \theta} - A(r)\frac{\partial}{\partial z}$ form a local **nonorthonormal** frame field for $\mathcal{D}$. $[Y_1, Y_2] = -\frac{dA}{dr}\frac{\partial}{\partial z}$ so that $X_1, X_2, [X_1, X_2]$ span $\mathbf{R}^3$ everywhere except at points on the unit cylinder $r = 1$. But here $[Y_1, [Y_1, Y_2]] = -\frac{d^2A}{dr^2}\frac{\partial}{\partial z} \neq 0$ and so $Y_1, Y_2, [Y_1, [Y_1, Y_2]]$ span $\mathbf{R}^3$ and the bracket generating condition holds everywhere.

The fiber metric for our example is the restriction of the form $dx^2 + dy^2$

to $\mathcal{D}$. Here (x, y, z) are the usual Cartesian coordinates, so that $dx^2 + dy^2 = dr^2 + r^2 d\theta^2$. **In other words, the length of a horizontal curve is equal to the Euclidean length of its projection to the x, y plane.**

The horizontal curves which provide the counterexample are any horizontal curves lying on the unit cylinder $r = 1$. Thus they are helices with pitch $-A(1)$. One such curve is parameterized according to $(r, \theta, z) = (1, \theta, -A(1)\theta)$. We will call this curve $\hat{C}$. Its projection to the xy plane is the unit circle C.

Our main result is

Theorem 1 *The above defined distribution $\mathcal{D}$ is bracket generating. The helix $\hat{C}$ or any subarc thereof is a geodesic. It does not satisfy the subRiemannian geodesic equations, (1) above. In particular there are minimizing subRiemannian geodesics which do not satisfy the subRiemannian geodesic equations.*

Analytically speaking, this theorem says that for some sufficiently small positive number θ_0 the arc of the helix $\hat{C}$ from $\hat{C}(0)$ to $\hat{C}(\theta_0)$ is the shortest horizontal curve among all horizontal curves which join $\hat{C}(0)$ to $\hat{C}(\theta_0)$.

Remark 2 . *Bär [1], [2] gives a quite different kind of a counterexample. His curve γ is also a geodesic which does not satisfy the geodesic equations.*

However every sufficiently short arc of his geodesic does satisfy the geodesic equations whereas **no** *subarc of our geodesic satisfies the equations. In his example normal extremals $\xi : I \to T^*Q$ which project to subarcs of γ can be found, but they cannot not be chosen consistenly so as to be continous over all of γ. It follows from the maximum principle that his curve is not a minimizing geodesic.*

4 The Helix Does Not Satisfy the Geodesic Equations

An **orthonormal** framing of D is provided by

$$X_1 = \frac{\partial}{\partial r}$$

and

$$X_2 = \frac{1}{r}\{\frac{\partial}{\partial \theta} - A(r)\frac{\partial}{\partial z}\}.$$

It follows that the normal Pontrjagin Hamiltonian (cf. the equation immediately preceding eq. 1) is

$$H(r, \theta, z, p_r, p_\theta, p_z) = \frac{1}{2}\{p_r^2 + \frac{1}{r^2}(p_\theta - p_z A(r))^2\} \quad (3)$$

In Cartesian coordinates

$$H(x, y, z, p_x, p_y, p_z) = \frac{1}{2}\{(p_x - p_z A_x(x, y))^2 + (p_y - p_z A_y(x, y))^2\}$$

Here $A_x dx + A_y dy = A(r) d\theta$ and the Cartesian p's are related to the cylindrical p's in the usual way, that is, through the formula $p_x dx + p_y dy = p_r dr + p_\theta d\theta$.

This is exactly the Hamiltonian for a particle with charge

$$e = p_z$$

travelling in the xy plane under the influence of a magnetic field normal to the plane and with strength

$$B = \frac{1}{r}\frac{dA}{dr}.$$

It is important in this regard that H is independent of z, for it follows from this that one of the geodesic equations is

$$\dot{p}_z = 0$$

and thus the "charge" is constant along solutions.

To write these equations let

$$\pi : \mathbf{R}^3 \to \mathbf{R}^2 \; ; \; \pi(x, y, z) = (x, y)$$

be the map which projects out the vertical coordinate. If γ is a solution to the geodesic equations write

$$c = \pi \circ \gamma$$

for its planar projection. Now write c in terms of complex variables: $c = x + iy$. Then by straightforward calculation c satisfies equations

$$\ddot{c} = -ieB(c)\dot{c} \tag{4}$$

where $B(x,y) = \frac{d\Theta}{dx \wedge dy} = \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}$. In the particular case where A is a function of r alone one easily calculates that B is given by the previous formula. If γ is parameterized by arclength, or what is the same, if $H = \frac{1}{2}$ along γ, then the above differential equation can be rewritten

$$\kappa(s) = -eB(c(s)). \tag{5}$$

where κ is the curvature of the unit speed curve c. In our example $B = 0$ on the unit circle C. But C has curvature 1 and is the projection of our helix $\hat{C}$. Thus $\hat{C}$ cannot satisfy the subRiemannian geodesic equations.

Remark 3 *This connection between the equations of motion of a particle in a magnetic field and the subRiemannian geodesic equations was first pointed out in my paper [20] . See also Granchina and Vershik [29]. If the codimension of D is larger than one then the equations are related to those of a "classical quark" in a nonAbelian gauge field. The relation only holds if the subRiemannian structure admits a Lie group of symmetries acting transversely to its distribution.*

5 Rigidity: A Heuristic Proof of the Theorem

It remains to show that sufficiently small subarcs of $\hat{C}$ are minimizing geodesics. In this section we will give an incomplete but illuminating argument in support of this fact.

Consider a horizontal curve γ in $\mathbf{R}^3$ which connects the point $(x_0, y_0, 0)$ to the point (x_0, y_0, z_1). Thus the projected curve $c \subset \mathbf{R}^2$ is closed. By Stoke's theorem and the fact that $dz = -Ad\theta$ along γ

$$z_1 = -\int\int d\Theta = -\int\int B(r) r dr d\theta \tag{6}$$

where the last integral is over the (oriented) region in the plane enclosed by of c. Following the theory of magnetism we call the integral $\int\int B(r) r dr d\theta$ the *flux* enclosed by c. Recall that the (Euclidean) length of c is the same as the subRiemannian length of γ. **It follows that γ is a minimizing geodesic if and only if its projection to the xy plane minimizes length among all loops based at (x_0, y_0) and enclosing an amount of flux equal to $-z_1$.**

Now consider our helix $\hat{C}$. The flux enclosed by going once around $C = \pi \circ \hat{C}$ is $2\pi A(1) = \int\int_D B dx dy$ where D is the unit disc. Imagine perturbing $\hat{C}$ to form the horizontal curve γ and consequently perturbing its projection C to form c. If we push part of C into the interior of the unit disc we have subtracted flux since B is positive in the interior. On the other hand if we push part of C to the exterior of the disc then we add negative flux, that is, we also subtract flux. **See figure 1.** No matter how we perturb C we decrease the flux and hence increase the height difference z_1 and thus violate the endpoint conditions. We conclude that **there are no allowable variations of** $\hat{C}$ and consequently it is a local minimum for our constrained variational problem. This heuristic argument can easily be turned into a rigorous proof of the following

Assertion 1 *There are no piecewise C^1 endpoint preserving variations of $\hat{C}$ through horizontal curves except those whose variation field is tangential to $\hat{C}$.*

This shows that $\hat{C}$ is an isolated point in the space of all horizontal piecewise C^1 unparameterized curves which join $\hat{C}(0)$ to $\hat{C}(2\pi)$ with the (piecewise) C^1 topology on this path space. Being an isolated point, it is automatically a local minimum **for any function on this space** and in particular for the length functional.

Remark 4 *Extremals of this type have been a source of major difficulties to practitioners of "classical" (one independent variable) calculus of variations. Such extremals are called "rigid" by Young [31], "abnormal" by Bliss*

*[4], and of "maximal class" by Caratheodory [7]. If the Pontrjagin maximum principle [22] is applied to the problem of finding subRiemannian geodesics one finds that they are precisely the abnormal extremals, and that these are also the singular extremals in this case. We especially recommend the treatment in the beginning of volume 2 of Young's book regarding these extremals. Also of interest is the treatment of Morse and Myers [**?**].*

$\hat{C}$ is not isolated in the C^0 or even the Sobolev H^1 topology on the space of all horizontal curves with its endpoints. One can find arbitrarily C^0- or H^1- close curves to $\hat{C}$ which have $\hat{C}$'s endpoints by making tiny kinks in chords to the circle C. One might then wonder whether it is possible to make sufficiently small kinks in C so as shorten its length while keeping its flux the same. The theorem says that this is not possible, at least for small enough arcs of $\hat{C}$. Our work lies in showing this.

6 Proof

6.1 The proof from two propositions

Theorem 1 follows immediately from the following two propositions.

Proposition 1 *Any minimizer not lying in the cylinder must satisfy the sR geodesic equations.*

Proposition 2 *There is a positive number θ_* such that for all $\theta \leq \theta_*$ every solution γ to the subRiemannian geodesic equations with endpoints $\hat{C}(0)$ and $\hat{C}(\theta)$ has $length(\gamma) > \theta$.*

Proof of Theorem 1. Combine the two propostions, recalling that θ is the subRiemannian arclength along $\hat{C}$.

6.2 Proof of Proposition 1

The salient fact is that the unit cylinder is precisely where the distribution D fails to be a contact distribution.

Definition 3 *A rank two distribution on a three-manifold is said to be contact at a point q if for some (and hence any) frame field X_1, X_2 of D defined near q the vectors $X_1(q), X_2(q)$ together with their Lie bracket $[X_1, X_2](q)$ form a basis for the tangent space at q.*

With this in mind, one can find various proofs of Proposition 1 in the literature. See for example Zhong [10], Hamenstädt [13], Strichartz [25] or Hermann [15]. We will give another proof based on the maximum principle.

Proof of propositon 1. According to that principle every minimizer is a Pontrjagin extremal and all such extremals can be divided into the abnormal and the normal ones. The normal ones satisfy, by definition, the sR geodesic equations. The propositon now follows from the following lemma.

Lemma 2 . *Every nonconstant abnormal Pontrjagin extremals projects to a D-curve which lies on the locus of points where D fails to be contact.*

Proof of Lemma. Let X_1, X_2 be the frame for D used earlier and let X_3 be their Lie bracket. Let P_i be the corresponding "power functions" as above $P_i(q,p) = p(X_i(q))$. The maximum principle tells us to introduce the multipliers $(p(t), \lambda_0) \in (\mathbf{R}^3)^* \oplus \mathbf{R}$, $(p(t), \lambda_0) \neq (0,0)$ For fixed (q, p, λ_0) the controls $u(t)$ leading to an extremal must maximize

$$H(q, u, p, \lambda_0) = u_1 P_1 + u_2 P_2 - (\lambda_0)\frac{1}{2}(u_1^2 + u_2^2).$$

For an abnormal extremal $\lambda_0 = 0$ so the only possibility is that $P_1 = P_2 = 0$, $p \neq 0$ along the extremal. If $u(t)$ is the control inducing the abnormal extremal then the extremal is an integral curve of the Hamiltonian flow of the time-dependent Hamiltonian $H_*(q,p,t) = u_1(t)P_1 + u_2(t)P_2$. Since $\{P_1, P_2\} = -P_3$ we find that $\dot{P}_1 = -u_2 P_3\, , \dot{P}_2 = u_1 P_3$. Thus we require that

$P_1 = P_2 = P_3 = 0$ along an abnormal. (The u_i cannot both be zero since the extremal is not a single point.) But these are the components of the covector p on the vectors X_1, X_2, X_3. Consequently, if these vectors form a basis at a point q we must have $p = 0$ which is not allowed. Therefore the curve $q(t)$ must lie on the locus of points where X_1, X_2 and X_3 become linearly dependent, which is to say, where D fails to be contact.

Remark 5 *From the point of view of differential topology, proposition 1 is a statement about the regularity of the endpoint map. This is the map which assigns to each D-path γ beginning at a point q_0 its endpoint $end(\gamma) = q(1)$. (Alternatively, end is the input-output map.) The proposition states that if D is contact at one point of the curve γ then end is a submersion at γ. Hence, by the implicit function theorem the subset $end^{-1}(q_1)$ of D- paths which join q_0 to q_1 forms a smooth submanifold of the space of all horizontal paths based at q_0. It is then legitimate to use the standard method of Lagrange multipliers applied to the functional $(length) + \lambda(end)$ to find the equations characterizing extremals. These are the subRiemannian geodesic equations.*

7 Proof of Proposition 2

The proof is by contradiction. Suppose it is false. Then there must exist a sequence of lengths s_i decreasing to zero and a sequence $\gamma_i : [0, t_i] \to \mathbf{R}^3$ of **minimal** solutions to the sR geodesic equations which have the same endpoints as $\hat{C}[0, s_i]$, are parameterized by arclength, and are shorter than s_i. Thus $\gamma_i(0) = \hat{C}(0)$, and $\gamma_i(t_i) = \hat{C}(s_i)$. Also $\|\frac{dc_i}{dt}\| = 1$ where we write $c_i = \pi \circ \gamma_i$ and finally

$$sin(s_i) \leq t_i \leq s_i.$$

The first inequality expresses the fact that the Euclidean distance between the endpoints of $C[0, s_i]$ is $sin(s_i)$.

Let

$$m_{*i} = max|1 - r_i(t)|$$

where

$$\gamma_i(t) = (r_i(t), \theta_i(t), z_i(t))$$

is the expression for γ_i in cylindrical coordinates. Then we also must have

$$m_{*i} \leq s_i \tag{7}$$

for otherwise the length of γ_i would be greater than s_i.

The proof will be based on an analysis of the geodesic equations as $i \to \infty$. The salient fact will be that the 'charges' e_i must go to infinity as i does. To perform the analysis we will need the equations in both their geometric form , $\kappa(s) = -eB(c(s))$, and their Hamiltonian form,

$$\dot{\theta} = \frac{1}{r^2}(p_\theta - eA(r)) \tag{8}$$

$$\begin{aligned} \dot{r} &= p_r \\ \dot{z} &= -A(r)\dot{\theta} \\ \dot{p}_\theta &= 0 \\ \dot{p}_z &= 0 \end{aligned}$$

The differential equation for p_r which we will use is found by solving for $\dot{r}$ in the formula for the Hamiltonian

$$\frac{1}{2} = H = \frac{1}{2}\{\dot{r}^2 + \frac{1}{r^2}(p_\theta - eA(r))^2\}. \tag{9}$$

The following facts regarding solutions will be needed.

Fact [1]. p_θ and $p_z = e$ are constant along solutions.

Fact [2]. Suppose that the projection $c = \pi \circ \gamma$ of a solution is tangent to a circle $r = const.$ at some point $c_* = c(t_*)$. Then c is symmetric with respect to reflection about the radial line through this point. Analytically, this reads $c(t_* + s) = R(c(t_* - s))$ where R is the reflection about the line joining the origin to c_*.

These two facts follow directly from the equations and their invariance with respect to orthogonal transformations. The following facts are direct consequences of facts [1] , [2] and the equations.

Fact [3]. Let $\beta(n)$ denote the angle a projected solution c makes with the unit circle at its nth crossing. Then $\beta(n+1) = -\beta(n)$. In particular, since $\dot{\theta}(n) = cos(\beta(n))$ at each crossing, we have $\dot{\theta}(n) = \dot{\theta}(n+1)$.

Definition 4 *An arc of solution γ will mean that part lying between two consecutive intersections with the unit cylinder $r = 1$, or the projection to the xy plane of such an arc. Thus if there are $n+1$ such intersections then there are n arcs.*

An arc will be called "interior" or "exterior" depeding on whether or not it is interior or exterior to the unit disc. By the 'corresponding arc' of C we will mean the shorter arc of C lying between the endpoints of such an arc. By the "height" or "vertical coordinate" of an arc we will mean the difference between the z-coordinates of γ corresponding to the two endpoints of the arc. Thus the height is $-\int_{arc} Ad\theta$.

Fact[4]. Extremals of $B(c(s))$ along any projected solution c coincide with points of tangency to some circle. Since the curve is symmetric upon reflection about the line through such a point there is exactly one such extremal on any arc and it is a strict maximum for $|B|$ and for $|1-r|$ along the arc.

Fact [5]. For $e \neq 0$ and $r \neq 1$ the sign of the curvature κ is equal to $-sgn(e)sgn(1-r)$.

We will begin our analysis by eliminating certain solution curves. Solutions with $e = 0$ can be eliminated because they project to straight line segments. If such a curve satisies the xy endpoint conditions its projection must be a chord of the circle. So, by the argument of §4 it cannot satisfy the z-endpoint condition. Solutions with $\dot{\theta}_0 = 0$ can be eliminated because they must remain exterior to the cylinder (see fact [3]). It follows from the convexity of the unit disc that their projections are always longer than the corresponding segment of the circle, and so they cannot be minimizing.

Solutions with $\dot{\theta}_0 > 0$ and $e > 0$ or with $\dot{\theta}_0 < 0$ and $e > 0$ can be eliminated because they can never satisfy the z-component of the endpoint conditions. To see this observe that the projected arcs of such a curve can never have self-intersections. For example, if $0 < \beta_0 < \pi/2$ and $e > 0$ then, by Fact 5 the curvature in the arc's interior is strictly negative. Coupled with Fact 2 this shows that the arc can have no self-intersection. (Eg: Viewed as a graph over its tangent at $t = 0$ it forms a concave function. See figure (i).) Upon crossing the cylinder, the sign of the curvature of the arc switches and the same argument applies again. (As a graph it becomes strictly convex.) Therefore the Green's theorem argument of §4 (the 'heuristic proof') is valid and these curves can never satisfy the vertical (z) endpoint condition.

We are left with analyzing sequences of solutions $\gamma_i(t) = (x_i(t), y_i(t), z_i(t))$ with initial conditions $\dot{\theta}_{i,0} > 0, e_i < 0$, or $\dot{\theta}_{i,0} < 0, e_i > 0$. Set

$$h_i = z_i(t_i) - (-A(1)s_i),$$

so that our last endpoint condition is

$$h_i = 0.$$

We have

$$h_i = -\int_{c_i} A d\theta + A(1)s_i.$$

Now

$$A(r) = A(1) - \eta(r)^2$$

where

$$\eta = \eta(r) = b(1-r) + O((1-r)^2)\,;\, b > 0$$

is a smooth invertible function of $1-r$ near $r = 1$. This follows from the fact that A has a nondegenerate maximum at $r = 1$. (For the model case (eq. (2)), $A(r) = \frac{1}{2}r^2 - \frac{1}{4}r^4$ we find $A = \frac{1}{4} - \eta(r)^2$ with $\eta(r) = \frac{(1-r^2)}{2} =$

$(1-r)-\frac{1}{2}(1-r)^2$. Thus $b=1$. We have chosen $b>0$ so that $\eta>0$ when $r<1$. Then $b=+\sqrt{-\frac{1}{2}\frac{d^2A(1)}{dr^2}}$.) It follows that

$$h_i=\int_0^{t_i}\eta^2\dot{\theta}dt \tag{10}$$

Since η^2 is positive off of the unit cylinder, it follows that somewhere along the solution the sign of $\dot{\theta}$ must become opposite to its initial sign, $\dot{\theta}_0$.

We now fix attention on the particular arc for which this sign switch occurs. Later, we will show that the sign switch must occur on every subarc of the solution. Most importantly, we will show that the length of this solution arc, (and all of these other subarcs) is longer than the corresponding subarc of the circle.

Without loss of generality we will assume that $\dot{\theta}_0>0$, the argument for the contrary case being identical. The curvature and symmetry conditions guarantee that the sign switch occurs at the midpoint of the arc. The qualitative shape of the arc is indicated in figure (ii). We will use a subscript $*$ to denote any variable evaluated at this midpoint. Thus

$$\dot{r}_*=0.$$

Now $H=\frac{1}{2}=\frac{1}{2}\{\dot{r}^2+r^2\dot{\theta}^2\}$ for any unit speed solution. Therefore

$$1=r_*^2\dot{\theta}_*^2$$

As we saw at the beginning of the proof of proposition 2, (eq. 7) $m_* = |1 - r_*| \to 0$ so that

$$r_* \to 1 \ , \ \eta_* \to 0$$

and

$$\dot{\theta}_* \to -1$$

as $i \to \infty$.

Equation (8) at $r = 1$ tells us that the the initial angular velocity is

$$\dot{\theta}_{0,i} = p_{\theta,i} - e_i A(1)$$

Since p_θ is constant, we can rewrite that equation (8) as

$$\dot{\theta} = \frac{1}{r^2}\{\dot{\theta}_{0,i} + e_i \eta(r)^2\} \tag{11}$$

Evaluating this equation at the midpoint, and taking the limit as $i \to \infty$ and recalling that $\eta_* \to 0$ we see that in order for the limit just calculated for $\dot{\theta}_*$ to be valid we must have

$$e_i \to -\infty$$

as $i \to \infty$. More precisely

$$e_i \eta_* \to -2 \tag{12}$$

Define the small parameter ϵ by

$$|e| = \frac{1}{\epsilon^2}.$$

Define the rescaled variable Y through the relation

$$\eta = \epsilon Y$$

The above limit (12) becomes

$$Y_{*,i} \to \sqrt{2} \tag{13}$$

(The +-sign occurs since the arc is interior to the unit disc.) And $Ad\theta = A(1)d\theta - \epsilon^2 Y^2 d\theta$.

RemarkIt follows from inequality (7) above that

$$\epsilon_i = O(s_i)$$

since

$$\eta_* = bm_* + O(\eta_*^2)$$

(A refinement of that inequality 7 shows that $m_* = o(s_i)$ and so in fact $\epsilon = o(s_i)$.)

Now

$$\dot{\eta}^2 = R_i(Y) \tag{14}$$

where the function $R = R_i$ is obtained by solving equation (9) for $\dot{r}^2$ and then using our change of variables $r \to \eta \to Y$. We calculate

$$R_i(Y) = R(Y, \epsilon_i, \dot{\theta}_{i,0}) = b^2[1 - (1 - Y^2)^2] + O(\epsilon_i) + O(1 - \dot{\theta}_{i,0})$$

Now η_* is the unique zero of R_i on its arc, since the midpoint of the arc is the only place where $\dot{\eta} = 0$. (See fact 4 above.) **Since R_i and the r.h.s. of the equation for $\dot{\theta}$ (??) do not depend on time, it follows that on every arc of our solution $\dot{\theta}$ is negative and close to -1 at that arc's midpoint, and that the value of η there is η_*.**

We complete the argument by now showing that for i large enough each subarc of the solution is longer than the corresponding arc of the circle. [If the arc is exterior to the circle this follows immediately from convexity. But this observation does not seem to lead to simplification of the proof. For we do not know that a single arc satisfies the endpoint condition $h_i = 0$ and so several may need to be concatenated to satisfy the vertical endpoint condition.] For this purpose, let us change the meaning of the original notation so that now our curves γ_i consist each of single arcs, with lengths t_i and subtending angles s_i: $\theta(t_i) = s_i$. But we do not insist on the vertical endpoint condition. By reflectional symmetry it suffices to show that $t_{*,i} - \theta_{*.i}$ is eventually positive. Now this quantity is the integral of $dt - d\theta$ over half an arc. Thus we will now show that

$$\int_0^{t_{*,i}} (1 - \dot{\theta}_i(t))dt > 0 \tag{15}$$

provided i is sufficiently large. To do this we change variables to Y. Now $dt = \frac{dt}{d\eta} d\eta = \frac{1}{\dot{\eta}}(\epsilon)dY$ It follows that

$$dt = \epsilon \frac{dY}{\sqrt{R_i(Y)}}.$$

Hamilton's equation for $\dot{\theta}$ in terms of the the scaled variable Y is

$$\dot{\theta} = \frac{1}{r^2}(\dot{\theta}_0 - Y^2)$$

Thus

$$\begin{aligned} 1 - \dot{\theta} &= 1 - \tfrac{1}{r^2}(\dot{\theta}_o - Y^2) \\ &= (1 - Y^2) - \tfrac{1}{r^2}[1 - Y^2 + (\dot{\theta}_0 - 1)] + Y^2 \\ &= (1 - \tfrac{1}{r^2})(1 - Y^2) + Y^2 + \tfrac{1 - \dot{\theta}_0}{r^2} \end{aligned}$$

Now

$$1 - \frac{1}{r^2} = \frac{r^2 - 1}{r^2} = -\frac{2\epsilon}{b} Y g(Y, \epsilon_i)$$

where $g(Y, \epsilon)$ is a smooth function converging uniformly to 1 on the interval of integration, $[0, Y_*]$, as $\epsilon \to 0$. This follows from the expansion $\eta = b(1 - r) + O(\eta^2)$. [For the special model we have $\eta = \frac{1-r^2}{2} = (1 - r) - \frac{1}{2}(1 - r)^2$ exactly.] Thus our integrand is

$$(1 - \dot{\theta})dt = \epsilon_i[Y^2 + \epsilon_i Y k_i(Y)] \frac{dY}{\sqrt{R_i(Y)}} + \frac{1 - \dot{\theta}_0}{r^2} dt$$

Here $k_i(Y) = -2g(Y, \epsilon_i)/b$ which is a smooth bounded function uniformly converging to a constant on the interval of integration. **The last term in the integrand, the one involving dt is always nonnegative**, since $1 > \dot{\theta}_{0,i}$ for unit speed curves. (In fact this term is strictly positive for interior arcs.) It follows that

$$limsup \frac{t_{*,i} - \theta_{*.i}}{\epsilon_i} \geq limsup \int_0^{Y_*} [Y^2 + \epsilon_i Y k_i(Y)] \frac{dY}{\sqrt{R_i(Y)}} \tag{16}$$

Assuming the validity of interchanging limits and integration we obtain

$$limsup \frac{t_{*,i} - s_{*,i}}{\epsilon_i} \geq \int_0^{\sqrt{2}} \frac{Y^2}{b\sqrt{1-(1-Y^2)^2}} dY = \frac{\sqrt{2}}{b} > 0.$$

This completes the proof, modulo showing that we may interchange the limit with the integration.

7.1 Interchange of limit and integration

The only difficulty is that the limiting denominator, $\sqrt{R(Y,0,0)}$, has a simple zero at $Y = 0$ in the case of a self-intersecting arc. This pole cancels with a factor of Y in the limiting numerator, yielding an integrable function. But we cannot blindly apply the dominated convergence theorem. (To see the difficulties consider, for example, the function $Y^2/|Y - \epsilon|$ which has a simple pole at ϵ and whose limiting denominator has a simple zero at zero.

Its integral over the unit interval is $+\infty$ for $\epsilon > 0$ but the integral of its limit is $\frac{1}{2}$.)

Recall that the tangent vector to any smooth planar curve c turns by a total amount $\int_c \kappa(s)ds$ upon traversing the curve. For our self-intersecting arcs c_i this amount of turning is $2\pi + \theta_i - 2\beta_i$ where θ_i is the angle subtended by the arc, and β_i is the angle the arc initially makes with the circle. (Thus if the curve consists of one arc, we have $\theta_i = s_i$.)

*******figure 4 = figure on turning tangent*****

Since our arcs solve the geodesic equations they have curvature

$$\kappa(t) = \epsilon B(t) = \frac{1}{\epsilon^2} 2\eta(s) b(1 + O(\epsilon)) = \frac{1}{\epsilon} 2Y(s) b K_i(Y),$$

where the function $K_i = K_i(Y) = 1 + O(\epsilon_i)$ is a smooth function on the interval $[0, Y_{*,i}]$ and tends uniformly to 1. (For the model case $b = 1$ and $K_i = 1$, exactly.) Changing variables from t to Y and using the reflectional symmetry of the arc we obtain:

$$\int_0^{Y_{*,i}} \frac{Y b K_i(Y)}{\sqrt{R_i(Y)}} dY = \frac{\pi}{2} + \frac{\theta_i}{4} - \frac{\beta_i}{2}$$

Let g_i denote the integrand:

$$g_i = \frac{Y(bK_i(Y))}{\sqrt{R_i(Y)}} = \frac{Yb(1 + O(\epsilon_i))}{b\sqrt{1 - ((1 - Y^2)^2 + O(\epsilon_i)}}$$

Then

$$g_i \to g = \frac{1}{\sqrt{2 - Y^2}}$$

where the convergence is uniform on compact subsets of the open interval $(0, \sqrt{2})$. One calculates directly that

$$\int g(Y)dY = \frac{\pi}{2}$$

so that $\int g_i \to \int g$ as well.

Referring back to the integrand of the desired limit, equation 16, set

$$f_i = Y(1 + \epsilon_i k_i(Y))$$

Then the integrand for the desired limit is $f_i g_i$. Also

$$f_i \to f = Y$$

where the convergence is now uniform on the entire closed interval $[0, \sqrt{2}]$

The following lemma now validates interchanging the limit with the integral sign.

Lemma 3 *Let b_i be a sequence of numbers converging to a number b. Let g_i be a sequence of nonnegative continuous functions defined on the open intervals (a, b_i) which converge uniformly on every compact subset, $[a+\delta, b-\delta]$, $\delta > 0$ of the limit interval (a, b) to the function g. Suppose that $\int_a^{b_i} g_i \to$*

$\int_a^b g$, this integral being finite. Let f_i be a sequence of continuous functions on a neighborhood of the **closed** *interval $[a,b]$ which converge uniformly on that interval to a function f. Then $\int_a^{b_i} f_i g_i \to \int_a^b fg$.*

The proof of this lemma is a straightforward exercise in real analysis. We will not give the proof except in so far as to say that the only difficulty is that the g_i can have poles at the endpoints. The value of f at these endpoints together with the uniform convergence of the f_i and the boundedness of $\int g$ serves to control the sequence of integrals near the endpoints.

This completes the proof of Proposition 2 and hence the main Theorem.

8 Stability

The germ of a distribution at a point is called *stable* if any other distribution which is sufficiently (Whitney-) close to it is locally diffeomorphic to it. For example, the Darboux theorem states that any contact distribution is stable.

The distribution in our example is stable. To see this we formalize some properties of that distribution. Suppose that $\mathcal{D}$ is a distribution on a 3-manifold which is defined in a neighborhood of a point q by the vanishing of a nonzero one-form Θ. Define the function B by writing

$$\Theta \wedge d\Theta = B d^3 x$$

where d^3x is a locally defined volume form. Suppose that B has a transverse zero at q: $B(q) = 0$ and $dB(q) \neq 0$. These are precisely the properties of our example which we will need.

Definition 5 *A distribution with the above properties will be called a simply degenerate contact structure. The surface $\{B = 0\}$ is called the singular surface.*

By a theorem of Martinet [19] if Θ is any one-form with the above properties then there exists a smooth nonzero function g and coordinates x, y, z defined in a neighborhood of q such that

$$g\Theta = dz - y^2 dx.$$

See also Zhitomirskii [32]. Since Θ and $g\Theta$ define the same distribution this proves

Lemma 4 *Distributions of simply degenerate contact type are stable.*

Theorem 2 (Stable Counterexample.) *Suppose that Q is a 3-dimensional sR manifold whose underlying distribution is a simply degenerate contact structure . Then every horizontal curve on the singular surface is a geodesic.*

For an open dense set of fiber inner products on the distribution these geodesics do not satisfy the geodesic equations. In particular, the abnormal *geodesics of our Theorem 2 are stable under C^2 perturbations of the sR structure.*

We will not prove this theorem. It follows the same lines as our proof of Theorem 2 but the algebra and analysis is considerably messier. There is now a simple proof due to Liu and Sussmann.

Remark 6 *The fact that the result is purely differential-topological suggests that there is a differential-topological proof. We have not found one yet.*

9 Acknowledgements

I thank Alan Weinstein for bringing Rayner's paper [24] to my attention and Stanislaw Janeczko for referring me to the papers of Martinet [19] and Zhitomirskii [32]. I also thank Ge Zhong and Danny Goroff for extended discussions while at MSRI. Part of this work was done while in MSRI in Berkeley with the financial support of an NSF Postdoctoral Grant, number DMS-8807219. I would like to thank both of these institutions and UC Santa Cruz for support.

References

[1] C. Bär, *Geodesics for Carnot-Caratheodory Metrics*, preprint, (1988).

[2] C. Bär, *Carnot-Caratheodory-Metriken*, Diplomarbeit. Bonn, (1988).

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

[4] G.A. Bliss, **Lectures on Calculus of Variations**, Univ. of Chicago Press, 1946.

[5] R.W. Brockett, *Nonlinear Control Theory and Differential Geometry*, Proc. of the Int. Congress of Mathematicians, Warszawa,(1983).

[6] R.W. Brockett, *Control Theory and Singular Riemannian Geometry*, in New Directions in Applied Mathematics, P.J. Hilton and G.S.,Young, editors, Springer, (1981).

[7] C. Carathéodory, **Calculus of Variations and Partial Differential Equations of the First Order, vol. 2**, Holden-Day, S.F., CA, (1967).

[8] W.L. Chow, *Uber Systeme van Linearen partiellen Differentialgleichungen erster Ordnung*, Math. Ann., vol. 117, 98-105, (1939).

[9] B. Gaveau, *Principle de Moindre Action propagation de la chaleur et estimees sons elliptiques sur certains groupes nilpotents*, Acta Math., **139**, 95-153, (1977).

[10] Zhong Ge , *On a Variational Problem and the spaces of Horizontal Paths*, Pacific J. Math., v. 149, 61-93, (1991).

[11] V. Ya Gershkovich and A.M. Vershik and *Estimation of the Functional Dimension of the Orbit Space of Germs of Distributions in General Position*, Math. Notes. of the Acad. of Sci. of USSR, **44**, no. 5-6, 806-810 (1988).

[12] V. Ya Gershkovich and A.M. Vershik, *Non-holonomic Manifolds and Nilpotent Analysis*, J. Geom. and Phys. vol. 5, no. 3, (1988).

[13] U. Hamenstädt, *Some Regularity Theorems for Carnot-Caratheodory Metrics* , J.D.G. , v. 32, 819-850 (1990).

[14] G. H. Hardy, Y. E. Littlewood, and G. Polya,**Inequalities**, (1934) .

[15] R. Hermannn, *Some Differential Geometric Aspects of the Lagrange Variational Problem* , Indiana Math. J., 634-673 (1962).

[16] R. Hermannn, *Geodesics of Singular Riemannian metrics*, Bull. AMS, vol. 79, no.4, 780-782, (1973).

[17] I. Kupka , preprint, U. of Toronto (1992).

[18] W-S. Liu and H. Sussmann *Abnormal SubRiemannian Minimizers*, IMA preprint, (1993).

[19] J. Martinet, *Sur les Singularites Des Formes Différentielles*, Ann. Inst. Fourier , vol. 20, no. 1, 95-178, (1970).

[20] R. Montgomery, *The Isoholonomic Problem and Some Applications*, Comm. Math. Phys., vol. 128, 565-592, (1990).

[21] P. Pansu, *Metriques de Carnot-Caratheodory et quasiisometries des espaces symetriques de rang un*, Ann. of Math., **129**, 1-60, (1989).

[22] L. S. Pontrjagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, **The Mathematical Theory of Optimal Processes**, Wiley, Interscience, 1962.

[23] P. K. Rashevski, *About connecting two points of complete nonholonomic space by admissible curve*, (in Russian), Uch. Zapiski ped. inst. Libknexta, no. 2, 83-94, (1938).

[24] C.B. Rayner, *The Exponential Map for the Lagrange Problem on Differentiable Manifolds*, Phil. Trans. of the Royal Soc. of London, ser. A, Math. and Phys. Sci. no. 1127, v. 262, 299-344, (1967).

[25] R. Strichartz, *Sub-Riemannian Geometry*, J. Diff. Geom. 24, 221-263, (1983).

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

[27] T.J.S. Taylor, *Some Aspects of Differential Geometry Associated with Hypoelliptic Second Order Operators*, Pac. J. Math. v. 136, no. 2 , 355-378, (1989).

[28] A. Weinstein, *Fat Bundles and Symplectic Manifolds*, Adv. in Math, v. 37 , 239-250, (1980).

[29] A.M. Vershik and O. Granchina ,*Reduction of Non-holonomic Variational Problems to Isoperimetric Ones and Connections in Principal Bundles*, preprint (in Russian) (1990).

[30] A.M. Vershik and V. Ya Gershkovich, *Non-holonomic Riemannian Manifolds*, in Dynamical Systems v. 7, part of the Mathematical Encyclopaedia series, vol.16 (in Russian), MIR pub., 1988.

[31] , L.C. Young, **Lectures on the Calculus of Variations and Optimal Control Theory**, Chelsea, 1980.

[32] , M. Ya. Zhitomirskii, *Singularties and Normal Forms of Odd-Dimensional Pfaff Equations*, Funktsional'nyia Analiz I Ego Prilozxheniya, vol. 23, no. 1, 70-71, (1989).