united nations
educational, scientific
and cultural
organization

international atomic
energy agency

the
abdus salam
international centre for theoretical physics

SMR1327/8

Summer School on Mathematical Control Theory

(3 - 28 September 2001)

Notes on Subriemannian Geometry from 'A tour of subriemannian geometry'

Richard Montgomery
Department of Mathematics
University of California at Santa Cruz
Santa Cruz, CA 95064
U.S.A.

These are preliminary lecture notes, intended only for distribution to participants

Notes on Subriemannian Geometry from 'A tour of subriemannian geometry' - by R. Montgomery

Chapter 1

Dido Meets Heisenberg

We begin with the isoperimetric problem and Dido's problem, or more precisely, the duals of these problems. What is the shortest curve enclosing a given area? By adding an extra variable corresponding to this area we rephrase the problem as that of finding geodesics for a certain nonriemannian geometry on $\mathbb{R}^3$. This geometry is the simplest example of a subriemannian geometry. It is called the Heisenberg group, and it leads us into the basic definition of a subriemannian geometry, in section 1.4. In section 1.5 we formulate a system of subriemannian geodesic equations, which are ODEs for curves on the cotangent bundle of the underlying manifold. What might be called the main result of the chapter, Theorem 1.5.7 asserts that sufficiently short solutions to these equations project down to minimizing subriemannian geodesics. These geodesics form the "main class" of geodesics, the *normal* ones. There is another less-understood class of geodesics, the *singular* ones. The singular geodesics are not present in the Heisenberg group. By combining this fact with Theorem 1.5.7, and writing down and solving the geodesics equations for the Heisenberg group, we find that the solutions to the (dual) to Dido's problem and the isoperimetric problem are indeed what we have known for thousands of years: arcs of circles. In section 1.5 we present an overview of some of the basic results in subriemannian geodesy and geometry, along with Theorem 1.5.7. In section 1.9 we prove Theorem 1.5.7 using the Hamilton-Jacobi method. In section 1.10 we present a number of examples of subriemannian geometries, ending with my favorite, the spherical version of the Heisenberg group, which is a geometry on the three-sphere.

1.1 Dido's problem

Dido's problem is a variant of the isoperimetric problem. It was formulated in the *Aeneid*, Virgil's epic poem glorifying the beginnings of Rome.

Queen Dido had to flee across the Mediterranean in a ship with friends and servants. She had what we would nowadays call a dysfunctional family. Her brother, Pygmalion, had just murdered her husband and taken most of her

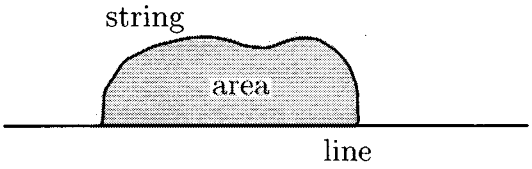

Figure 1.1: Dido's problem.

possessions. Dido landed, nearly penniless, on a part of the African coast ruled by King Jarbas. After dickering and begging, Dido persuaded Jarbas to give her as much land as she could enclose with an ox hide. Dido told her servants to cut an ox hide into a single long, narrow strip. They turned the ox hide into a single leather string.

Dido had in this way reformulated her difficult situation into the following geometric problem. Given a string of fixed length ℓ and a fixed line L (the Mediterranean coastline), place the ends of the string on L and determine the shape of the curve c for which the figure enclosed by c together with L has the maximum possible area. This is Dido's problem. It is also sometimes referred to as the *problem of Pappus*. Dido found the solution – a half-circle – and thus founded the semicircular city of Carthage.

1.2 A vector potential

Introduce the one-form

$$\alpha = \frac{1}{2}(xdy - ydx)$$

which satisfies

$$d\alpha = dx \wedge dy$$

and

$$\alpha_L = 0 \quad \text{for any ray } L \text{ through the origin.}$$

According to Stokes' theorem, the area Φ enclosed by a closed planar curve c is

$$\Phi(c) = \int_c \alpha. \tag{1.1}$$

Because of the second property, if c is a non-closed curve beginning at the origin, $\Phi(c)$ represents the area enclosed by the closed curve obtained by traversing c and then returning to the origin along the ray that connects the endpoint of c to the origin.

The length ℓ of $c = (x(t), y(t))$ is

$$\ell(c) = \int_c ds, \tag{1.2}$$

where $ds = \sqrt{dx^2 + dy^2} = \|\dot{c}\|dt$ is the usual element of arc length. In this manner Dido's problem, and the (dual) isoperimetric problem, becomes the following constrained variational problem:

Problem 1.2.1 *Minimize the length $\ell(c)$ of a closed rectifiable curve c, subject to the constraint that the signed area $\Phi(c)$ of the curve be a fixed constant.*

The introduction of α lets us extend the problem to non-closed curves. The ray used to close up c corresponds to the coastline L in Dido's problem.

1.3 Heisenberg geometry

We construct the three-dimensional geometry whose geodesics correspond to the solutions to the isoperimetric problem. Add a third direction z whose motion is linked to that of x and y according to

$$\dot{z} = \frac{1}{2}(-y\dot{x} + x\dot{y}). \tag{1.3}$$

In this way we associate a family of curves $\gamma(t) = (x(t), y(t), z(t))$ to a single planar curve $c(t) = (x(t), y(t))$, the family being parameterized by the initial value z_0 of the height z. We will call any one of these paths a *horizontal lift* of c, and more generally, any path γ in $\mathbb{R}^3$ that satisfies the differential constraint 1.3 a *horizontal path.*

Set
$$ds = \sqrt{dx^2 + dy^2}$$
and define the length of any horizontal path in $\mathbb{R}^3$ to be $\int_\gamma ds$. In other words, we have defined the length of γ to be equal to the usual length of its planar projection c.

Problem 1.3.1 *Minimize the length $\int_\gamma ds$ over all horizontal paths γ that join two fixed points in three-space.*

To see that this is a reformulation of the dual to Dido's problem, or the isoperimetric problem, observe that

$$z(1) - z(0) = \int_c \frac{1}{2}(xdy - ydx)$$

where $c(t) = (x(t), y(t))$ is the projection of the curve $\gamma(t) = (x(t), y(t), z(t))$ to the plane. Observe that, according to Stokes' theorem, if c joins the origin to (x_1, y_1) and if we take $z(0) = 0$, then the endpoints of γ are $(0, 0, 0)$ and $(x_1, y_1, \Phi(c))$, where $\Phi(c)$ denotes the signed area defined by the closed curve given by traversing c and then returning to the origin along a line segment.

Define the differential one-form $\Theta = dz - \frac{1}{2}(xdy - ydx)$ and write

$$\begin{aligned} \mathcal{H}_{(x,y,z)} &= \{\Theta(x,y,z) = 0\} \\ &= \left\{(v_1, v_2, v_3) : v_3 - \frac{1}{2}(xv_2 - yv_1) = 0\right\} \subset \mathbb{R}^3. \end{aligned}$$

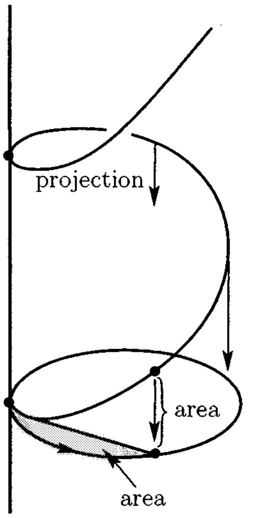

Figure 1.2: Heisenberg geometry. The height increases at a rate proportional to the area swept out.

This $\mathcal{H}$ is a field of two-planes in three-space, or what we call a *distribution*: a linear subbundle of the tangent bundle. The restriction of ds^2 to these two-planes defines a smoothly varying family of inner products $\langle \cdot, \cdot \rangle$ on the planes $\mathcal{H}$. Thus if $v, w \in \mathcal{H}_{(x,y,z)}$ then $\langle v, w \rangle = v_1 w_1 + v_2 w_2$.

Definition 1.3.2 $\mathbb{R}^3$ *endowed with the structure of this distribution* $\mathcal{H}$ *and this family of inner products* ds^2 *on* $\mathcal{H}$ *is called the Heisenberg group.*

The reason for the name Heisenberg will be explained shortly. This group is the first nontrivial example of subriemannian geometry.

1.4 The definition of a subriemannian geometry

Definition 1.4.1 *A* subriemannian geometry *on a manifold* Q *consists of a* distribution, *which is to say a vector subbundle* $\mathcal{H} \subset TQ$ *of the tangent bundle of* Q, *together with a fiber inner-product* $\langle \cdot, \cdot \rangle$ *on this subbundle.*

We will call $\mathcal{H}$ the *horizontal distribution*. An object such as a vector field or a curve on Q is called *horizontal* if it is tangent to $\mathcal{H}$. We define the *length* $\ell = \ell(\gamma)$ of a smooth horizontal curve γ as in Riemannian geometry:

$$\ell(\gamma) = \int \|\dot{\gamma}\| dt,$$

where $\|\dot{\gamma}\| = \sqrt{\langle \dot{\gamma}(t), \dot{\gamma}(t) \rangle}$ is computed using the inner product on the horizontal spaces $\mathcal{H}_{\gamma(t)}$, and the integral is over the domain of the curve. We use the length to define the subriemannian distance $d(A, B)$ between two points A and B, just as in Riemannian geometry:

$$d(A, B) = \inf \ell(\gamma)$$

where the infimum is taken over all smooth horizontal curves that connect A and B. This distance is infinite if there is no such curve joining A to B.

In order to understand and analyze this distance function, we will need to expand the domain of the length functional to the largest possible class of curves, namely the absolutely continuous horizontal curves. A curve $\gamma : I \to Q$ on a manifold is absolutely continuous if it has a derivative for almost all t, and if in any coordinate system the components of this derivative are measurable functions for which the fundamental theorem of calculus applies: the curve itself can be recovered by integration from an initial point and the derivative. (See [Royden 1968] for details in Euclidean space.) If these conditions on γ hold in one collection of coordinate systems covering γ, then they hold in any such collection, since the coordinate transition maps are themselves absolutely continuous. We say that an absolutely continuous curve is horizontal if its derivative lies in $\mathcal{H}$ wherever it exists. The length of an absolutely continuous horizontal curve is always defined, although it may be infinite. The distance $d(A, B)$ defined above remains the same if we replace the smooth curves in the infimum by the larger class of absolutely continuous horizontal curves.

Definition 1.4.2 *An absolutely continuous horizontal path that realizes the distance between two points is called a* minimizing geodesic *or simply a* geodesic.

The study of minimizing subriemannian geodesics forms one of the main currents of the book. However, the following fundamental question remains open: *Are all minimizing geodesics smooth?*

Minimizing energy instead of length. The *energy* of a horizontal curve is

$$E(\gamma) = \int_\gamma \frac{1}{2}||\dot\gamma||^2.$$

As in Riemannian geometry it is analytically more convenient to minimize E as opposed to ℓ, and this is how we will proceed. To see that minimizing E and ℓ yield the same curves we use the Cauchy-Schwartz inequality:

$$\int fg \le \sqrt{\int f^2}\sqrt{\int g^2}$$

with equality if and only if $f = cg$ for some constant c. We apply this inequality to $f = ||\dot\gamma||$ and the constant function $g = 1$. Fix the time T of the path $\gamma : [0, T] \to Q$. We obtain

$$\ell(\gamma) \le \sqrt{\int ||\dot\gamma||^2}\sqrt{T} = \sqrt{2E(\gamma)}\sqrt{T}$$

with equality if and only if γ is a constant-speed curve, meaning $\|\dot{\gamma}\| = c =$ constant. This proves the following proposition.

Proposition 1.4.3 *The horizontal curve γ minimizes the energy E among all curves joining q_0 to q_1 in time T if and only if it minimizes the length ℓ among all curves joining q_0 to q_1 and is parameterized to have constant speed $c = d(q_0, q_1)/T$.*

Example: The Heisenberg geodesics. We have seen seen that the geodesics in the Heisenberg group correspond to solutions to the isoperimetric problem, or to Dido's problem. The solutions to the isoperimetric problem are well known to be circles. For some beautiful elementary proofs of this nontrivial fact see [Howards et al. 1999]. For our version of Dido's problem, with fixed nonequal endpoints in the plane, these minimizers are arcs of circles in the plane, including line segments as degenerate cases. (Line segments through the origin have $\Phi(c) = 0$.)

It follows that the Heisenberg geodesics are precisely the horizontal lifts of arcs of circles in the plane.

Proposition 1.4.4 *The geodesics for the Heisenberg group are exactly the horizontal lifts of arcs of circles, including line segments as a degenerate case.*

This proposition is an immediate corollary of the theorem on normal geodesics (Theorem 1.5.7 below), together with the computations of section 1.7.

1.5 Geodesic equations

The cometric. A Riemannian metric is defined by a covariant two-tensor, which is to say a section of the bundle $S^2(T^*Q)$. There is no such object in subriemannian geometry. Instead, a subriemannian metric can be encoded as a contravariant symmetric two-tensor, which is a section of $S^2(TQ)$. This two-tensor has rank $k < n$, where k is the rank of the distribution, so it cannot be inverted to obtain a Riemannian metric. We call this contravariant tensor the *cometric.*

Definition 1.5.1 *A* cometric *is a section of the bundle $S^2(TQ) \subset TQ \otimes TQ$ of symmetric bilinear forms on the cotangent bundle of a manifold.*

Since TQ and T^*Q are dual, any cometric defines a fiber-bilinear form $((\cdot,\cdot)) : T^*Q \otimes T^*Q \to \mathbb{R}$, i.e. a kind of "inner product" on covectors. This form in turn defines a symmetric bundle map $\beta : T^*Q \to TQ$ by $p(\beta_q(\mu)) = ((p,\mu))_q$ for $p, \mu \in T^*_qQ$ and $q \in Q$. Thus $\beta_q(\mu) \in T_qQ$. The adjective *symmetric* means that β equals its adjoint $\beta^* : T^*Q \to T^{**}Q = TQ$.

The cometric for a subriemannian geometry β is uniquely defined by the following conditions:

1. $\operatorname{im}(\beta_q) = \mathcal{H}_q$;

2. $p(v) = \langle \beta_q(p), v \rangle$ for $v \in \mathcal{H}_q$, $p \in T_q^*Q$, where $\langle \beta_q(p), v \rangle_q$ is the subriemannian inner product on $\mathcal{H}_q$.

Conversely, any cometric of constant rank defines a subriemannian geometry whose underlying distribution has that rank.

Definition 1.5.2 *The fiber-quadratic function $H(q,p) = \frac{1}{2}(p,p)_q$, where $(\cdot,\cdot)_q$ is the cometric on the fiber T_q^*Q, is called the subriemannian* Hamiltonian, *or the* kinetic energy.

The Hamiltonian H is related to length and energy as follows. Suppose that γ is a horizontal curve. Then $\dot\gamma(t) = \beta_{\gamma(t)}(p)$ for some covector $p \in T^*_{\gamma(t)}Q$, and

$$\frac{1}{2}\|\dot\gamma\|^2 = H(q,p).$$

H uniquely determines β by polarization, and β uniquely determines the subriemannian structure. This proves the following proposition:

Proposition 1.5.3 *The subriemannian structure is uniquely determined by its Hamiltonian. Conversely, any non-negative fiber-quadratic Hamiltonian of constant fiber rank k gives rise to a subriemannian structure whose underlying distribution has rank k.*

To compute the subriemannian Hamiltonian we can start with a local frame $\{X_a\}_{a=1}^k$ of vector fields for $\mathcal{H}$. Think of the X_a as fiber-linear functions on the cotangent bundle. In so doing, we will rename them P_a. Thus

$$P_a(q,p) = p(X_a(q)), \quad q \in Q, p \in T_q^*Q.$$

Definition 1.5.4 *Let X be a vector field on the manifold Q. The fiber-linear function on the cotangent bundle $P_X : T^*Q \to \mathbb{R}$ defined by $P_X(q,p) = p(X(q))$ is called the* momentum function *for X.*

Thus the $P_a = P_{X_a}$ are the momentum functions for our horizontal frame.

If $X_a = \sum X_a^i(x)(\partial/\partial x^i)$ is the expression for X_a relative to coordinates x^i, then $P_{X_a}(x,p) = \sum X_a^i(x) p_i$, where $p_i = P_{\partial/\partial x^i}$ are the momentum functions for the coordinate vector fields. The x^i and p_i together form a coordinate system on T^*Q. They are called *canonical coordinates.*

Let $g_{ab}(q) = \langle X_a(q), X_b(q) \rangle_q$ be the matrix of inner products defined by our horizontal frame. Let $g^{ab}(q)$ be its inverse matrix. Then g^{ab} is a $k \times k$ matrix-valued function defined in some open set of Q.

Proposition 1.5.5 *Let P_a and g^{ab} be the functions on T^*Q that are induced by a local horizontal frame $\{X_a\}$ as just described. Then*

$$H(q,p) = \frac{1}{2}\sum g^{ab}(q) P_a(q,p) P_b(q,p). \tag{1.4}$$

We leave the proof, which is pure linear algebra, to the reader. Note, in particular, that if the X_a are an orthonormal frame for $\mathcal{H}$ relative to the subriemannian inner product, then

$$H = \frac{1}{2}\sum P_a^2.$$

Normal geodesics. Like any smooth function ("Hamiltonian") on the cotangent bundle, our function H generates a system of Hamiltonian differential equations. (See Appendix A.) In terms of canonical coordinates (x^i, p_i) these differential equations are

$$\dot{x}^i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial x^i}. \tag{1.5}$$

Definition 1.5.6 *The Hamiltonian differential equations (1.5) are called the* normal geodesic equations.

Riemannian geometry can be viewed as a special case of subriemannian geometry, one in which the distribution is the entire tangent bundle. The cometric is the usual inverse metric, written g^{ij} in coordinates. The normal geodesic equations in the Riemannian case are simply the standard geodesic equations, rewritten on the cotangent bundle. (See [Abraham and Marsden 1978; Arnol'd 1989]. For those unfamiliar with Hamiltonian formalism, we briefly review it in section 1.7 and provide more details in Appendix A.)

Theorem 1.5.7 (on normal geodesics) *Let $\zeta(t) = (\gamma(t), p(t))$ be a solution to Hamilton's differential equations on T^*Q for the subriemannian Hamiltonian H and let $\gamma(t)$ be its projection to Q. Then every sufficiently short arc of γ is a minimizing subriemannian geodesic. Moreover γ is the unique minimizing geodesic joining its endpoints.*

The theorem will be proved in section 1.9.

Definition 1.5.8 *We call the projected curves $\gamma(t)$ of Theorem 1.5.7 the* normal *subriemannian geodesics.*

Singular geodesics. Not all subriemannian geodesics are normal. There are subriemannian geometries which admit minimizing geodesics that do not solve the geodesic equations as defined by H. The first example is presented in chapter 3. This example is not pathological; perturbations cannot destroy it.

We call these peculiar geodesics *singular geodesics.* Their existence makes subriemannian geodesy very different from the study of Riemannian geodesics. In Riemannian geometry all geodesics are normal: they come from the geodesic equations.

The existence of singular geodesics can be ruled out for contact distributions. (See chapters 5 and 8.) The Heisenberg group has a contact distribution, so we need not worry about the singular geodesics. The theorems and phenomena of singular geodesics are the subject of chapters 3 and 5.

1.6 Chow's theorem and geodesic existence

We have been discussing the structure of geodesics. How do we know they exist? In other words, given points A and B, is there a subriemannian geodesic that joins them. Even more fundamentally, is there a horizontal curve that joins them? This last question is addressed by Chow's theorem, also called the Chow-Rashevskii theorem, which is the first and most basic theorem in the business.

When $\mathcal{H}$ is involutive we cannot horizontally connect arbitrary points A and B. (Recall that a distribution is called *involutive* if whenever X and Y are horizontal vector fields, their Lie bracket $[X, Y]$ is also horizontal.) The Frobenius theorem asserts that when $\mathcal{H}$ is involutive, the set of horizontal paths through a fixed point A sweeps out a smooth immersed submanifold, called the *leaf* through A, whose dimension equals k, the rank of the distribution. So if $\mathcal{H}$ is involutive and B does not lie on the leaf through A, we cannot horizontally connect A and B.

At the opposite end from the involutive distributions stand the bracket-generating distributions. Given a collection $\{X_a\}$ of vector fields, form its *Lie hull*, the collection of all vector fields $\{X_a, [X_b, X_c], [X_a, [X_b, X_c]], \ldots\}$ generated by Lie brackets of the X_a. We say that the collection $\{X_a\}$ is *bracket generating* if this Lie hull spans the whole tangent bundle.

Definition 1.6.1 *A distribution $\mathcal{H} \subset TQ$ is called* bracket generating *if any local horizontal frame $\{X_a\}$ for the distribution is bracket generating (over its domain).*

We remark that if $\{X_a\}$ is bracket generating, and if $\{Y_a\}$ is defined by $Y_a = \sum B_a^c X_c$ with B_a^c a smooth invertible matrix-valued function, then $\{Y_a\}$ is also bracket generating. Consequently, to check the bracket-generating condition near a point q we need only check it for a single horizontal frame defined in a neighborhood of q.

Theorem 1.6.2 (Chow) *If a distribution $\mathcal{H} \subset TQ$ is bracket generating then the set of points that can be connected to $A \in Q$ by a horizontal path is the component of Q containing A.*

In other words, bracket-generating plus connected implies horizontally path-connected. This theorem, and its proof, are the subject of the next chapter.

With Chow's theorem in mind, we address the problem of existence of geodesics.

Theorem 1.6.3 (Local existence) *If Q is a manifold with a bracket-generating distribution then any point A of Q is contained in a neighborhood U such that every $B \in U$ can be connected to A by a minimizing geodesic.*

In other words, on a bracket-generating subriemannian manifold any two sufficiently close points can be joined by a minimizing geodesic.

Theorem 1.6.4 (Global existence) *Suppose that Q is a connected manifold with a bracket-generating distribution and that Q is complete relative to the subriemannian distance function. Then any two points of Q can be joined by a minimizing geodesic.*

The proofs of these last two theorems appear in Appendix E.

1.7 Geodesic equations on the Heisenberg group

We return to the Heisenberg group. The vector fields

$$X = \frac{\partial}{\partial x} - \frac{1}{2}y\frac{\partial}{\partial z}, \quad Y = \frac{\partial}{\partial y} + \frac{1}{2}x\frac{\partial}{\partial z}$$

form an orthonormal frame for the Heisenberg geometry. This means that they frame the two-plane field $\mathcal{H}$ and that they are orthonormal with respect to the inner product $ds^2 = (dx^2 + dy^2)|_{\mathcal{H}}$ on that distribution. According to the discussion above, the subriemannian Hamiltonian is

$$H = \frac{1}{2}\left(P_X^2 + P_Y^2\right), \tag{1.6}$$

where P_X, P_Y are the momentum functions of the vector fields X, Y. Thus

$$P_X = p_x - \frac{1}{2}yp_z, \quad P_Y = p_y + \frac{1}{2}xp_z,$$

where p_x, p_y, p_z are the fiber coordinates on the cotangent bundle of $\mathbb{R}^3$ corresponding to the Cartesian coordinates x, y, z on $\mathbb{R}^3$. Again, these fiber coordinates are defined by writing a covector as $p = p_x dx + p_y dy + p_z dz$. Together, (x, y, z, p_x, p_y, p_z) are global coordinates on the cotangent bundle $T^*\mathbb{R}^3 = \mathbb{R}^3 \oplus \mathbb{R}^3$.

Hamilton's equations can be written

$$\frac{df}{dt} = \{f, H\}, \quad f \in C^\infty(T^*Q) \tag{1.7}$$

which holds for any smooth function f. The function H defines a vector field X_H, called the Hamiltonian vector field, which has a flow $\Phi_t : T^*Q \to T^*Q$. Let $f : T^*\mathbb{R}^3 = T^*Q \to \mathbb{R}$ be any smooth function on the cotangent bundle. Form the time-dependent function $f_t = \Phi_t^* f$ by pulling f back via the flow. Thus $f_t(x, y, z, p_x, p_y, p_z) = f(\Phi_t(x, y, z, p_x, p_y, p_z))$. In other words, $df/dt = X_H[f_t]$, which gives meaning to the left-hand side of Hamilton's equations.

To define the right-hand side, which is to say the vector field X_H, we will need the Poisson bracket. The Poisson bracket on the cotangent bundle T^*Q of a manifold Q is a canonical Lie algebra structure defined on the vector space $C^\infty(T^*Q)$ of smooth functions on T^*Q. (For details, see Appendix A.) The Poisson bracket is denoted $\{\cdot, \cdot\} : C^\infty \times C^\infty \to C^\infty$, where $C^\infty = C^\infty(T^*Q)$, and can be defined by the coordinate formula

$$\{f, g\} = \sum_i \frac{\partial f}{\partial x^i}\frac{\partial g}{\partial p_i} - \frac{\partial g}{\partial x^i}\frac{\partial f}{\partial p_i}.$$

This formula is valid in any canonical coordinate system, and can be shown to be coordinate independent. The Poisson bracket satisfies the Leibniz identity

$$\{f, gh\} = g\{f, h\} + h\{f, g\}$$

which means that the operation $\{\cdot, H\}$ defines a vector field X_H, called the *Hamiltonian vector field.*

By letting the functions f vary over the collection of coordinate functions x^i and p_i we get the more common form of Hamilton's equations:

$$\dot{x}^i = \frac{\partial H}{\partial p_i}, \quad \dot{p}_i = -\frac{\partial H}{\partial x^i}.$$

(These equations are in turn equivalent to our formulation 1.7.) It is more convenient to use the formulation 1.7, because the momentum function $X \mapsto P_X$ is a Lie algebra anti-homomorphism from the Lie algebra of all smooth vector fields on Q to $C^\infty(T^*Q)$ with its Poisson bracket:

$$\{P_X, P_Y\} = -P_{[X,Y]}. \tag{1.8}$$

For the Heisenberg group, with our choice of X and Y as a frame for $\mathcal{H}$, we compute

$$[X,Y] = Z := \frac{\partial}{\partial z}, \quad [X,Z] = [Y,Z] = 0.$$

Thus

$$\{P_X, P_Y\} = -p_z := P_Z, \quad \{P_X, P_Z\} = \{P_Y, P_Z\} = 0.$$

These relations can also easily be computed by hand, from our formulae for P_X, P_Y and the bracket in terms of x, y, z, p_x, p_y, p_z. By letting f vary over the collection of functions $\{x, y, z, P_X, P_Y, P_Z\}$, using the bracket relations and equation 1.8, we find that Hamilton's equations are equivalent to the system

$$\begin{aligned}
\dot{x} &= P_X \\
\dot{y} &= P_Y \\
\dot{z} &= -\frac{1}{2}yP_X + \frac{1}{2}xP_Y \\
\dot{P}_X &= -P_Z P_Y \\
\dot{P}_Y &= +P_Z P_X \\
\dot{P}_Z &= 0.
\end{aligned}$$

The last equation asserts that $P_Z = p_z$ is constant. The variable z appears nowhere in the right-hand sides of these equations. It follows that the variables x, y, P_X, P_Y evolve independently of z, and so we can view the system as defining a one-parameter family of dynamical systems on $\mathbb{R}^4$ parameterized by the constant value of P_Z. Combine x and y into a single complex variable $w = x+iy$. Note that the first two equations say that $dw/dt = P_X + iP_Y$. The fourth and fifth equations say that the time derivative of $P_X + iP_Y$ is $ip_z(P_X + iP_Y)$. All together, then, we have $\partial^2 w/\partial t^2 = ip_z w$, $p_z =$ constant. These are the famed Lorentz equations for the motion of a particle in a constant magnetic field. To convert to electromagnetic notation, we set the parameter $p_z = Be/m$, where e is the particle's charge, B is the magnetic field strength, and m is the mass

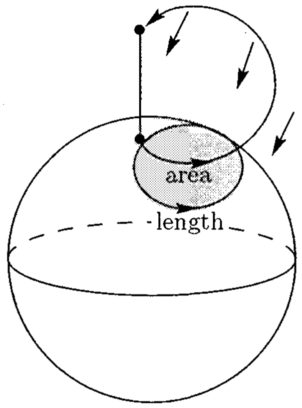

Figure 1.3: The Heisenberg sphere.

of the particle. Finally note that the third of our Hamilton's equations, the z equation, is just the differential constraint 1.3.

One integration of the Lorentz equations yields the evolution of the planar velocity: $P_X + iP_Y = P(0)\exp(ip_z t)$, where the complex vector $P(0) = P_X(0) + iP_Y(0)$ describes the initial velocity vector. A second integration yields the general form of the geodesics on the Heisenberg group:

$$x(t) + iy(t) = \frac{P(0)}{ip_z}\left[\exp(ip_z t) - 1\right] + (x(0) + iy(0)) := w(t),$$

$$z(t) = z(0) + \frac{1}{2}\int_0^t \operatorname{Im}(\bar{w}dw).$$

Based on these formulae we draw a picture of the Heisenberg sphere of radius a (Figure 1.3). It is a surface of revolution in our Cartesian coordinates (x, y, z) obtained by rotating this parametric curve about the z-axis:

$$\begin{aligned} r(\lambda, a) &= \frac{a}{\lambda} 2 \sin\left(\frac{\lambda}{2}\right) \\ z(\lambda, a) &= \frac{1}{2}\left(\frac{a}{\lambda}\right)^2 [\lambda - \sin(\lambda)]. \end{aligned}$$

Conjugate points. The Heisenberg spheres are singular where they intersect the z-axis. The points along the z-axis correspond to conjugate points for the exponential map. Take a circle passing through the origin in the plane, and rotate it about the origin. In this way we get a circle's worth of circles, all of the same area A, all of the same length, and all passing through the origin. The horizontal lifts through the origin of this family of circles forms a one-parameter family of subriemannian geodesics of the same length connecting 0 to

$(0, 0, A)$. Thus the entire z-axis consists of conjugate points to the origin (where "conjugate point" has the same definition as in Riemannian geometry). Note that unlike in Riemannian geometry, the set of conjugate points to the origin passes through the origin! This is a general phenomenon in subriemannian geometry: The conjugate and cut loci of a point pass through the point.

1.8 Why call it the Heisenberg group?

The three-dimensional Heisenberg algebra is the Lie algebra with basis X, Y, Z and bracket relations $[X, Y] = Z; [X, Z] = [Y, Z] = 0$. These are the bracket relations generated by our frame $\{X, Y\}$ for the Heisenberg distribution.

Heisenberg wrote down these bracket relations in his foundational works on quantum mechanics. In Heisenberg's work X, Y, Z are self-adjoint operators on a Hilbert space, with X corresponding to measuring position, Y to measuring momentum, and Z to a multiple of the identity. (See chapter 13 for a little bit on the foundations of quantum mechanics.)

The Heisenberg Lie algebra is a *nilpotent* Lie algebra. This means there is an integer r, called the *depth* or *step* of the algebra, such that any iterated bracket involving more that r elements is zero. The step of the Heisenberg algebra is 2. Thus, for example $[[X, Y], Y] = 0$. Every nilpotent Lie algebra has a unique simply connected Lie group, called a nilpotent group. For a nilpotent Lie algebra, the exponential map is a diffeomorphism between the algebra and its simply connected group. The exponential map thus provides global coordinates for the group, and in these coordinates the group multiplication law is a polynomial of degree r. We can thus think of a nilpotent group as a vector space with a polynomial group law. (See section 1.10 for more on nilpotent groups with subriemannian structures.)

The group for the Heisenberg algebra is called the *Heisenberg group*. Its group law is the quadratic operation on $\mathbb{R}^3$ given by

$$(x_1, y_1, z_1) \cdot (x_2, y_2, z_2) = \left(x_1 + x_2, y_1 + y_2, z_1 + z_2 + \frac{1}{2}(x_1 y_2 - x_2 y_1) \right).$$

The one-parameter subgroups through the identity for this group structure are easily checked to be the standard Euclidean lines:

$$\gamma_v(t) = \exp(t(v_1, v_2, v_3)) = (tv_1, tv_2, tv_3).$$

If we let $q = (x, y, z)$ be a variable point and compute the derivative of $q \cdot \gamma_v(t)$ we find that

$$\left.\frac{d}{dt}\right|_{t=0} (q \cdot \gamma_v(t)) = v_1 X(q) + v_2 Y(q) + v_3 Z(q),$$

where X, Y, Z is the frame for our distribution. In other words, X, Y, Z form a basis for the space of left-invariant vector fields on the group.

The distribution $\mathcal{H}$, which is the span of $\{X, Y\}$, is now seen to be left-invariant with respect to Heisenberg multiplication. Thus our subriemannian structure is a left-invariant subriemannian structure on the Heisenberg group: the action of the group on itself by left multiplication is an action by subriemannian isometries. As with any group, left multiplication acts transitively on the group. So we can transform any point to any other, and this transformation takes geodesics to geodesics. In this manner we transform the problem of joining A to B by a subriemannian geodesic to the problem of joining the origin $0 = (0,0,0)$ to $A^{-1}B$.

1.9 Proof of the theorem on normal geodesics

1.9.1 Heuristics via taming

First recall how to express the *Riemannian* geodesic flow for a Riemannian metric g. Let X_μ, $\mu = 1, \ldots, n = \dim(Q)$, be a local orthonormal frame for the metric. Write $P_\mu = P_{X_\mu}$ for the associated momentum functions. The Hamiltonian $H = \frac{1}{2}\sum P_\mu^2$ generates the Riemannian geodesic flow for g.

Definition 1.9.1 *A Riemannian metric g is said to* tame *a subriemannian metric $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ if the restriction of g to the horizontal space $\mathcal{H}$ equals the subriemannian inner product $\langle\cdot,\cdot\rangle$. In this case, we also say that g and $\langle\cdot,\cdot\rangle$ are* compatible.

It is easy to find many Riemannian metrics taming a given subriemannian one. Let us choose one, say $g = g_1$. Let $\mathcal{V}$ denote the orthogonal complement to $\mathcal{H}$ with respect to this metric. Then we can write

$$g = g_{\mathcal{H}} \oplus g_{\mathcal{V}}$$

where $g_{\mathcal{H}} = \langle, \cdot, \cdot\rangle$ is the given subriemannian metric and $g_{\mathcal{V}}$ is a fiber inner product on $\mathcal{V}$. Now form the family of Riemannian metrics

$$g_\lambda = g_{\mathcal{H}} \oplus \lambda^2 g_{\mathcal{V}}, \quad \lambda \to \infty$$

taming our subriemannian metric. By letting λ tend to infinity we are severely punishing motion in the vertical ($\mathcal{V}$) direction.

Heuristic proof of Theorem 1.5.7. Let X_a, X_i be an adapted orthonormal frame for $g = g_1$, so that the X_a are orthonormal on $\mathcal{H}$ and the X_i are an orthonormal frame for $\mathcal{V}$. Then $X_a, (1/\lambda)X_i$ is an orthonormal frame for g_λ. It follows that

$$H_\lambda = \frac{1}{2}\left(\sum P_a^2 + \frac{1}{\lambda^2}\sum P_i^2\right)$$

is the Hamiltonian generating the geodesics for g_λ. As $\lambda \to \infty$, H_λ tends to the Hamiltonian H governing the normal geodesics, so the γ_λ converge to normal geodesics.

The fault in this alleged proof is the last sentence. The limit $\lambda \to \infty$ is singular, and the γ_λ may not converge to solutions γ for H. They may instead converge to something else: the singular geodesics discussed in section 1.5. This phenomenon does indeed occur (see chapter 3).

1.9.2 Calibrations

We will give a complete proof that the normal geodesics are indeed locally minimizing geodesics. The proof is by the Hamilton-Jacobi method, put in modern dress.

Definition 1.9.2 *A* calibration Θ *on a subriemannian manifold is a closed one-form on Q with the property that $|\Theta(v)| \leq ds(v) := \sqrt{\langle v, v\rangle}$ for all horizontal vectors v. A horizontal curve γ is said to be* calibrated *by the calibration Θ if equality holds: $\Theta(\dot\gamma) = ds(\dot\gamma)$.*

The term *calibration* is usually associated to higher-dimensional gadgets. Thus, in the theory of minimal d-dimensional surfaces it is a d-form satisfying an analogous inequality involving d-area instead of arc length. For example, on a Kahler manifold, the powers of the Kahler form define a calibration which can be used to show that any complex submanifold of a Kahler manifold is locally area-minimizing. See for example [Harvey 1990, esp. ch. 7].

Lemma 1.9.3 *If γ is a calibrated horizontal curve, then every sufficiently short subarc of γ is a minimizing geodesic.*

Proof: Let γ_2 be any horizontal curve with the same endpoints as γ that is homologous to γ. Then $\int_{\gamma_2} \Theta = \int_\gamma \Theta$ by Stokes' formula. But $\int_{\gamma_2} ds \geq \int_{\gamma_2} \Theta$ and $\int_\gamma ds = \int_\gamma \Theta$ by the definition of calibrations. Therefore $\ell(\gamma_2) \geq \ell(\gamma)$. This shows that γ minimizes length among all homologous curves sharing its endpoints.

Fix a point P on γ. Fix any Riemannian metric taming the subriemannian structure, and a radius r small enough that the Riemannian ball of radius r about P is homeomorphic to the standard Euclidean ball. We claim that any segment $c \subset \gamma$ starting at P whose length is less than or equal to r is a globally minimizing subriemannian geodesic. For suppose γ_2 is another horizontal arc that shares its endpoints with such an arc. If γ_2 leaves B then its length is greater than r. If not, then it is homotopic to c, within B. So, by the result of the previous paragraph, it is at least as long as γ.

QED

1.9.3 Hamilton-Jacobi theory

Hamilton-Jacobi theory provides a method for constructing calibrations. The Hamilton-Jacobi equation we need is

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

viewed as a partial differential equation for S.

Lemma 1.9.4 *If S is a solution to the Hamilton-Jacobi equation above, then $\Theta = dS$ is a calibration.*

Proof: Clearly dS is closed. We check that it satisfies the calibration inequality. Let X_a be a local orthonormal frame for the horizontal distribution. Expand $v \in \mathcal{H}$ in terms of the frame: $v = \sum v_a X_a$. Then $ds(v)^2 = \sum v_a^2$ and $dS(v) = \sum v_a dS(X_a)$, so by the Cauchy-Schwartz inequality we have

$$|dS(v)| \leq ds(v)\sqrt{\sum dS(X_a)^2}.$$

But $H(q, dS) = \frac{1}{2}\sum dS(X_a)^2$ so that

$$\sqrt{\sum dS(X_a)^2} = 1,$$

since S satisfies our Hamilton-Jacobi equation.
QED

Example. If Q is a Riemannian manifold then our Hamilton-Jacobi equation is $\|\nabla S\|^2 = 1$. For any fixed $q_0 \in Q$ the function $S(q) = d_R(q_0, q)$ provides a smooth solution to this equation away from $q = q_0$ and the cut and conjugate loci of q_0.

Lemma 1.9.5 *In a neighborhood of any sufficiently short solution arc γ to Hamilton's equation for the subriemannian Hamiltonian H we can construct a local solution $q \mapsto S(q)$ to the Hamilton-Jacobi equation which calibrates this arc γ.*

Proof. The construction is the method of (bi-)characteristics. We may suppose that this normal geodesic is parameterized by arc length, which is the same as supposing that it is the projection of a solution for which $H = \frac{1}{2}$. We obtain S by the standard method of characteristics for constructing local solutions to the Hamilton-Jacobi equation.

Step 1. Let $z_0 = (q_0, p_0) \in T^*Q$ be the initial condition for the solution to Hamilton's equation that projects to our normal geodesic $\gamma(t)$. We suppose that γ is nonconstant, so that the Hamiltonian vector field $X_H(z_0)$ is nonzero. We will work microlocally, i.e. in the cotangent bundle, near z_0. Choose a local $(n-1)$-dimensional manifold $\Sigma \subset T^*Q$ which passes through z_0 and which enjoys the following properties:

1. Σ is isotropic, meaning that $\omega|_\Sigma = 0$, where ω is the canonical symplectic form on $T^*\Sigma$.
2. $H|_\Sigma = \frac{1}{2}$.
3. $X_H(z_0) \oplus T_{z_0}\Sigma$ is projected linearly isomorphically onto $T_{q_0}Q$ by the differential of the projection $T^*Q \to Q$.

To see that this is possible is an exercise in local symplectic linear algebra (Appendix A). Indeed, because $dH(z_0) \neq 0$ we can choose canonical coordinates $\{q_i, p^i\}$ centered at z_0 so that $H = -p_1 + \frac{1}{2}$ and so that the tangent space to the fiber $T^*_{q_0}Q$, which is the kernel of the differential of the projection, is spanned by the $\partial/\partial p_i$. Then $X_H(z_0) = \partial/\partial q_1$. Take $\Sigma = \{q_1 = 0, p_1 = 0, p_2 = 0, \ldots, p_n = 0\}$.

Step 2. Apply the flow ϕ_t of X_H to Σ, for $-\epsilon < t < \epsilon$, sweeping out an n-manifold $\tilde{\Sigma}$. By elementary symplectic geometry this n-manifold is Lagrangian (Appendix A). For ϵ small it remains transverse to the fibers of $T^*Q \to Q$, so it is the graph of a local one-form $Q \to T^*Q$.

Step 3. A basic result in symplectic geometry asserts that, in a cotangent bundle, *any* Lagrangian submanifold that is transverse to the fibers is locally the graph of a *closed* one-form. Since we are speaking locally, we may take this one-form to be exact. Thus $\tilde{\Sigma}$ is the graph of a one-form $\Theta = dS$ for some function S defined near q_0. (In the local coordinates of step 1, we have $S = q_1/2$, which we check by noting that $\tilde{\Sigma}$ is the graph of the one-form $dq_1/2$.)

Step 4. We check that S satisfies the Hamilton-Jacobi equation. Observe that any $(q,p) \in \tilde{\Sigma}$ can be represented two different ways, either as $p = dS(q)$ or as $(q,p) = \phi_t(z)$ for some $z \in \Sigma$. It follows that $H(q, dS(q)) = H(\phi_t(z)) = H(z) = \frac{1}{2}$ where the second equality follows from the fact that the flow of X_H preserves the values of the energy H.

We have succeeded in constructing a local solution S. It remains to show that dS calibrates our normal geodesic γ. First, observe that if $(q(t), p(t))$ is *any* solution to the subriemannian geodesic equations, then we have $\dot{q} = \sum v_a X_a$ with the components v_a given by $v_a = p(X_a) := P_a(q(t), p(t))$.

By construction the curve $(\gamma(t), dS(\gamma(t)))$ is a solution to the subriemannian Hamilton's equations. Applying the formula above with $p = dS$ we find that the velocity components of $\dot{\gamma}$ are $v_a = dS(X_a)$. This implies that equality holds in the Cauchy-Schwartz argument in the proof of Lemma 1.9.4, so that $dS(\dot{\gamma}) = ds(\dot{\gamma})$, and dS does indeed calibrate γ.

QED

The previous two lemmas combine to prove all of Theorem 1.5.7 except for its final uniqueness statement.

1.9.4 Uniqueness of the minimum

The final uniqueness statement of Theorem 1.5.7 follows immediately from this lemma:

Lemma 1.9.6 *Let γ be a normal geodesic of unit speed which is short enough that it admits a calibration dS as in Lemma 1.9.5. Then any minimizing geodesic of unit speed that shares γ's endpoints is equal to γ.*

Proof. Let γ_2 be another horizontal curve sharing γ's endpoints. According to the calibration argument (the proof of Lemma 1.9.4), the length of γ_2 equals that of γ if and only if $|dS|_{\gamma_2}| = ds|_{\gamma_2}$ almost everywhere (a.e.). Let $v_a(t)$ be the components of γ_2's velocity relative to the orthonormal frame X_a. Recall that the calibration inequality relied on the Cauchy-Schwartz inequality applied to $\sum v_a dS(X_a)$. We have equality if and only if there is a time-dependent constant $c(t)$ such that $v_a(t) = c(t)dS(\gamma_2(t))(X_a(\gamma_2(t)))$ a.e. After reparameterizing γ_2 we may assume that $c(t) = 1$ a.e. This is equivalent to assuming that γ_2 is parameterized by arc length a.e. So $\dot{\gamma}_2 = \sum v_a X_a(q)$ with

$$v_a = dS(\gamma_2) \tag{1.9}$$

along γ_2.

Set $p(t) = dS(\gamma_2(t))$ and form the curve $z(t) = (\gamma_2(t), p(t))$. Uniqueness follows immediately from the uniqueness of solutions to ODEs, together with the following two claims:

Claim 1. We have $z(0) = \zeta(0)$ where ζ is the solution to Hamilton's equations that projects to the normal geodesic $\gamma(t)$ in T^*Q.

Claim 2. The curve $z(t)$ satisfies Hamilton's equations for the subriemannian Hamiltonian.

Proof of Claim 1. By construction of S (see step 3), we have that $\zeta(0) = dS(\gamma(0))$. But $\gamma(0) = \gamma_2(0)$ so that $z(0) = dS(\gamma(0))$ also.

Proof of Claim 2. We will assume that γ_2 is differentiable. We leave it to the reader to fill in the functional analytic details necessary to massage our proof into an integrated form that will cover the possibility that γ_2 is merely absolutely continuous.

Hamilton's equations are

$$\dot{q} = \sum P_a X_a(q(t)), \tag{1.10}$$

$$\dot{p}_i = -\sum P_a \frac{\partial}{\partial q^i} P_a. \tag{1.11}$$

Moreover, in these local coordinates

$$P_a(q,p) = \sum p_i X_a^i(q). \tag{1.12}$$

We make the choice of covector p_i corresponding to $dS(q)$ so that we are taking $P_a = dS(\gamma(t))(X_a(\gamma(t)))$ along our curve. According to equation 1.9, these are the requisite components

$$v_a = P_a$$

along our curve, so that the first of Hamilton's equations (1.10) is satisfied. We must check equation 1.11. This follows from what is essentially the lynchpin of the Hamilton-Jacobi theory, applied to our situation.

Claim 3. Suppose that a horizontal curve $q(t)$ has components

$$v_a(t) = dS(q(t))(X_a(q(t)))$$

for some smooth function S. Then $z(t) = (q(t), dS(q(t)))$ satisfies Hamilton's equations if and only if S satisfies the Hamilton-Jacobi equation $H(q, dS(q)) =$ constant.

Proof of Claim 3. We have already seen that 1.10 is necessarily satisfied: $\dot{q}^i = \sum v_a X^i_a$ with $v_a = dS(q)(X_a(q))$. We check 1.11. In coordinates we have $p_i = \partial S/\partial q^i$ so that

$$\dot{p}_i = \sum \frac{\partial^2 S}{\partial q^j \partial q^i} v_a X^j_a.$$

From $P_a = \sum p_i X^i_a(q)$ we get

$$\frac{\partial P_a}{\partial q^i} = \sum_j \frac{\partial S}{\partial q^j}\frac{\partial X^j_a}{\partial q^i},$$

so that Hamilton's equations 1.11 read

$$\dot{p}_i = -\sum v_a \frac{\partial S}{\partial q^j}\frac{\partial X^j_a}{\partial q^i}.$$

Comparing the two equations we see that Hamilton's equations are satisfied if and only if

$$\sum v_a \left(\frac{\partial S}{\partial q^j}\frac{\partial X^j_a}{\partial q^i} + \frac{\partial^2 S}{\partial q^j \partial q^i} X^j_a \right) = 0.$$

But since $v_a = dS(X_a)$ this is simply the derivative $\partial/\partial q^i$ of the equation

$$\frac{1}{2}\sum dS(q)(X_a(q))^2 := H(q, dS(q)) = \frac{1}{2}.$$

This proves claim 3, and hence the lemma is proved.

QED

1.10 Examples

Penalty Metrics. Penalty metrics were used in the heuristic proof in section 1.9.1. Suppose that the tangent bundle of a manifold Q admits a splitting $TQ = \mathcal{H} \oplus V$ that is orthogonal for a family of metrics g_ϵ, $\epsilon > 0$. If $g_\epsilon|_{\mathcal{H}} = \langle \cdot, \cdot \rangle_{\mathcal{H}} + O(\epsilon)$ and if $\lim_{\epsilon \to 0} g_\epsilon(v, v) = +\infty$ for $v \in V$, $v \neq 0$, then the distance functions d_ϵ for g_ϵ converge to the subriemannian distance function for the subriemannian structure $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$. Also the Riemannian Hamiltonians H_ϵ converge to the subriemannian Hamiltonian. A simple example of such a family is $g_\epsilon = \langle \cdot, \cdot \rangle_{\mathcal{H}} + \langle \cdot, \cdot \rangle_V/\epsilon^2$.

Examples of this type also arise in the phenomenon of rigidity of locally symmetric spaces. Here $Q \subset TM$ is the collection of unit-length tangent vectors of a manifold M whose universal cover is a Riemannian symmetric space of non-positive curvature and $\mathcal{H}$ is the negative, or contracting, sub-bundle of the tangent bundle of Q. See [Pansu 1989; Mostow 1973].

Lie groups. Let G be a Lie group and $V \subset \mathfrak{g}$ a linear subspace of its Lie algebra. We can view the Lie algebra as the space of left-invariant vector fields on the group, in which case V corresponds to a left-invariant distribution. The bracket-generating condition is that V Lie-generates $\mathfrak{g}$. The restriction to V of an inner product on $\mathfrak{g}$ yields a subriemannian metric for which the action of G by left translation is an action by isometries.

Carnot Groups. This is a subexample of the previous example. Take G to be a graded nilpotent Lie group. This means that the Lie algebra of G has the form $\mathfrak{g} = V \oplus V_2 \oplus \ldots \oplus V_r$, where $V = V_1$ and $[V_i, V_j] = V_{i+j}$, and $V_s = 0$ for $s > r$. Thus all s-fold brackets are zero for $s > r$, which is to say that $\mathfrak{g}$ is nilpotent of step r. Assume further that V Lie-generates. Then using V we obtain a special version of the previous example. Any such G is called a *Carnot group.* An inner product on V endows G with a subriemannian structure.

Carnot groups enjoy the property of admitting dilations. A *dilation* δ_t, $t > 0$, on a metric space (Q, d) is a mapping such that $d(\delta_t x, \delta_t y) = td(x, y)$ for all $x, y \in Q$. To construct the Carnot dilation, consider the family of linear operators $\delta_t : \mathfrak{g} \rightarrow \mathfrak{g}$ which act by scalar multiplication by t^i on V_i. These operators are Lie algebra automorphisms which preserve V, and act on V by scaling by t. Consequently, assuming as we will now, that G is simply connected, the δ_t extend to group automorphisms to which we give the same name, $\delta_t : G \rightarrow G$. These are the Carnot dilations. Carnot groups arise as the "tangent cones" – the closest objects there are to the Riemannian tangent space – for a general subriemannian manifold (see chapter 8). Many of the papers in subriemannian geometry, and in particular in subriemannian geometric analysis, are devoted to Carnot groups.

The Heisenberg group of this chapter is the simplest noneuclidean Carnot group; it has step 2.

Bundles. Let $\pi : Q \rightarrow M$ be a Riemannian submersion, that is, Q and M are endowed with Riemannian metrics such that the restriction of $d\pi_q$ to the orthogonal complement to the fiber through q is an isometry of inner-product spaces. These orthogonal complements, $\ker(d\pi_q)^{\perp}$, form the horizontal spaces for what is known as an Ehresmann connection, and these will form the distribution $\mathcal{H}$ for a subriemannian structure on Q. The metric is obtained by restricting the Riemannian one to Q. We call such a metric a *subriemannian structure of bundle type.*

Principal Bundles. Apply the construction above to the special case where $Q \rightarrow M$ is a principal G-bundle. The submersion is Riemannian if and only if G acts on Q by isometries. In this way we get subriemannian structures on principal bundles for which G acts by subriemannian isometries. The underlying distribution for such a subriemannian structure is that of a connection on the bundle. Many physical problems can be put into this framework. Some of them are discussed in Part II of the book.

Suppose that in any local trivialization $\phi_U : U \times G \to \pi^{-1}(U)$ the pull-back $\phi_U^* ds_Q^2$ of the Riemannian metric on Q has the property that its restriction to the group factor $G \subset U \times G$ is a fixed bi-invariant metric on the Lie group G. In this case we say that the metric is of *constant bi-invariant type.* The main theorem of chapter 11 is the following:

Theorem 1.10.1 (on projected geodesics) *If Q is a principal bundle with a subriemannian structure of bundle type whose corresponding Riemannian structure is of constant bi-invariant type, then the normal subriemannian geodesics on Q are precisely the horizontal lifts of the projections of the Riemannian geodesics on Q.*

Homogeneous Bundles. We further specialize to the case where $Q = G$ is itself a Lie group and M is a homogeneous space for this group. Then $Q \to M$ is the quotient projection $G \to G/K$ where $K \subset G$ is the isotropy subgroup of the action of G on M. This defines on G the structure of a principal K-bundle. A G-invariant connection for this bundle is defined by choosing a subspace $\mathcal{H}_e \subset \mathfrak{g}$ that is complementary to the Lie algebra of K. In this way we get a subriemannian structure of bundle type on a Lie group.

If G admits a bi-invariant Riemannian metric, we can use it to define the complement $\mathcal{H}$, and hence the subriemannian structure on G. The theorem on projected geodesics now implies that the normal subriemannian geodesics on G are precisely the horizontal lifts of the orbits of its one-parameter subgroups $\exp(t\xi)$ acting on $M = G/K$.

Contact distributions. A *contact distribution* on a manifold Q is a distribution $\xi \subset TQ$ defined by the vanishing of a single one-form θ with the property that the restriction $d\theta|_{\xi_q}$ is symplectic on each distribution plane ξ_q, $q \in Q$. We recall that "symplectic" means "non-degenerate". In other words, if $X \in \xi$ and if $d\theta(X, v) = 0$ for all $v \in \xi$ then in fact $X = 0$. This condition of being symplectic implies that the distribution has even rank. Put a metric on the contact planes ξ_q and we have a subriemannian metric of contact type. These have been studied in great detail in dimension 3. (See [Agrachev et al. 1996] and references therein.)

Contact distributions are the most studied of the non-integrable distributions. They arise in complex analysis, where they are closely related to the notion of a CR manifold. They arise in the study of quantization. A symplectic manifold (see Appendix A for the definition) can be "pre-quantized". The result is a contact manifold of bundle type which forms a circle bundle over the original symplectic manifold. Due largely to the stability of contact manifolds, the field of *contact topology* is the subject of very active research.

Heisenberg group revisited. This example fits within the confines of every category of example so far discussed. The Heisenberg group G is a Lie group with a left-invariant subriemannian structure. The center H of this group is the

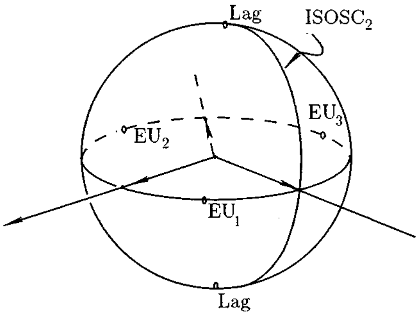

Figure 1.4: Spherical Heisenberg geometry.

real line generated by Z. The quotient $G \to G/H$ is, in our exponential coordinates, the projection $\mathbb{R}^3 \to \mathbb{R}^2$ which we have been using. Any Riemannian metric on G for which X, Y, Z are orthogonal has the property that this projection is a Riemannian submersion, relative to the usual Euclidean structure on the plane. Thus the subriemannian metric is of homogeneous bundle type. Since every translation-invariant metric on $\mathbb{R}$ is bi-invariant, Theorem 1.10.1 on projected geodesics applies and could be used to obtain the subriemannian geodesics.

Higher Heisenberg groups. Let V be a symplectic vector space with symplectic form ω and with dimension 2ℓ. Set $\mathfrak{g} = V \oplus \mathbb{R}$ and define a bracket on $\mathfrak{g}$ by $[(v,s),(w,t)] = (0, \omega(v,w))$. This is a 2-step graded nilpotent Lie algebra. It has a basis $\{X_i, Y_i, Z, i = 1, 2, \ldots \ell\}$, with nontrivial bracket $[X_i, Y_j] = \delta_{ij} Z$. The corresponding simply connected Lie group is the $(2\ell + 1)$-dimensional Heisenberg group. All the properties of the three-dimensional Heisenberg group carry over to this general case.

Spherical Heisenberg and the Hopf fibration. This is my favorite example. Start with the Heisenberg group. Replace the base space, which is the Euclidean plan, by the round two-sphere. Replace the z variable by a circular variable θ that evolves according to spherical area swept out. Scale the size of the sphere so that its total area is 2π. This scaling is consistent with θ taking values in the unit circle. The resulting global subriemannian geometry is that induced from the Hopf fibration

$$S^1 \to S^3 \to S^2.$$

Here S^3 is the round three-sphere, the Hopf fibration is a Riemannian submersion onto a round two-sphere, and the subriemannian geometry is the one induced from this Riemannian submersion.

S^3 with its standard "round" Riemannian metric is isometric to the Lie group $SU(2) = Sp(1)$ endowed with a *bi-invariant* metric. So the subriemannian structure for S^3 is of homogeneous bundle type as described above.

The geodesics on the sphere are of course great circles. Their projections to S^2 form the "small circles" – the geometric circles on S^2. As in the Heisenberg case, these circles are precisely the curves of constant curvature. Using the theorem on projected geodesics (Theorem 1.10.1), we see that the horizontal lifts of these small circles to S^3 exhaust the supply of subriemannian geodesics on S^3.

The geometry of the Hopf fibration is fundamental to algebraic topology [Bott and Tu 1995], Riemannian geometry, quantum mechanics (see [Feynman and Vernon 1957; Shapere and Wilczek 1989], or our chapter 13), and is important in celestial mechanics ([Stiefel and Scheifele 1971; Chenciner and Montgomery 2000]).

The $(2\ell + 1)$-dimensional version of the spherical Heisenberg group is not a group; it is a subriemannian homogeneous space structure on the $(2\ell + 1)$-dimensional sphere, and is associated with the Hopf fibration in that dimension. (See section 11.3.2.)

Chapter 2

Chow's Theorem: Getting from A to B

Chow's theorem, also known as the Chow-Rashevskii theorem, is the first theorem in subriemannian geometry [Chow 1939; Rashevskii 1938]. It asserts that any two points in a connected manifold endowed with a bracket-generating distribution can be connected by a horizontal path – a path tangent to the distribution. In other words, such a subriemannian manifold is "horizontally path connected". The definition of *bracket generating* and the theorem are found here in section 2.1, and were already presented in section 1.6.

In section 2.1 we also state various allied theorems and a counterexample. Theorem 2.1.3 asserts that on a subriemannian manifold endowed with a bracket-generating distribution the topology defined by the subriemannian distance function agrees with the usual manifold topology. We introduce the endpoint map, which plays a crucial role throughout the book. Theorem 2.1.5 asserts that the endpoint map is an open mapping. In section 2.1 we also present a counterexample to the converse to Chow's theorem. In section 2.2 we give a heuristic proof of Chow's theorem, following Hermann [1962], which provides more insight into the theorem than the standard proof.

In section 2.3 we define the growth vector and canonical flag associated to a distribution. These are used in section 2.4 where we state the ball-box theorem (Theorem 2.4.2), and prove half of it. The ball-box theorem is a stronger, quantitative version of Chow's theorem. Our proof of the ball-box theorem is standard; it is based on the implicit function theorem applied to "flow coordinates" defined via a frame for the entire tangent bundle built out of a frame for the distribution. In section 2.5 we prove the theorem on topologies (Theorem 2.1.3). In sections 2.6 and 2.7 we prove the other half of the ball-box theorem, following the exposition of Bellaiche [1996] and his definition and use of privileged coordinates. Finally, in section 2.8 we apply the ball-box theorem to prove Mitchell's theorem concerning the Hausdorff dimension and Hausdorff measure of a subriemannian manifold.

2.1 Bracket-generating distributions

Bracket-generating distributions stand at the opposite extreme from involutive distributions.

Definition 2.1.1 *A distribution $\mathcal{H} \subset TQ$ is called* bracket generating *if any local frame X_i for $\mathcal{H}$, together with all of its iterated Lie brackets $[X_i, X_j]$, $[X_i, [X_j, X_k]]$, $\ldots$, spans the tangent bundle TQ. Bracket-generating distributions are sometimes also called* completely nonholonomic *or distributions* satisfying Hormander's condition.

Remarks. If a frame is bracket generating in a neighborhood of a point, then so is any other frame. For any distribution $\mathcal{H}$, a slight perturbation of $\mathcal{H}$ in a generic direction is also bracket generating. Said in the language of jets, the germs of distributions that are bracket generating form an open dense subset of the space of distribution germs. (For some discussion in this direction see chapter 6.)

Theorem 2.1.2 (Chow's theorem) *If $\mathcal{H}$ is a bracket-generating distribution on a connected manifold Q then any two points of Q can be joined by a horizontal path.*

The theorem is proved in section 2.4.

Example: The Heisenberg group. For the Heisenberg group, X and Y frame $\mathcal{H}$, and with their bracket $[X, Y] = Z = \partial/\partial z$, they span the whole tangent space. Chow's theorem guarantees that we can connect any point to any other by a horizontal path.

Example: The Martinet distribution. The distribution on $\mathbb{R}^3$ defined by the vanishing of the form $dz - y^2 dx$ is called the Martinet distribution. It is spanned by vector fields $X = \partial/\partial x = y^2\partial/\partial z$ and $Y = \partial/\partial y$. We compute that $[X, Y] = -2y\partial/\partial z$ and $[[X, Y], Y] = -2\partial/\partial z$. The surface $y = 0$ is called the *Martinet surface.* Off of this surface one bracket suffices to span the entire tangent bundle. On the surface we need two brackets to span. But the distribution is bracket generating everywhere, so any two points can be connected by a horizontal path.

We now present two other theorems, which are essentially equivalent to the formulation of Chow's theorem in Theorem 2.1.2 and which provide additional insight into the theorem.

Theorem 2.1.3 (Topological theorem) *If $\mathcal{H}$ is a bracket-generating distribution on Q then the topology on Q induced by the subriemannian distance function is the usual manifold topology.*

This theorem is proved in section 2.5.

Definition 2.1.4 *The* endpoint map *associated to a distribution $\mathcal{H}$ on Q and based at the point $A \in Q$ is the map that takes each horizontal curve beginning at A to its endpoint.*

We will prove in Appendix E that the endpoint map is a smooth map from an infinite dimensional manifold to our finite dimensional manifold Q.

Theorem 2.1.5 (Open mapping version of Chow) *If $\mathcal{H}$ is bracket generating then the endpoint map is an open mapping.*

This theorem is a consequence of the ball-box theorem of section 2.4.

Definition 2.1.6 *If $\mathcal{H}$ is a (not necessarily bracket-generating) distribution on Q and $A \in Q$, then the* accessible set *associated to $\mathcal{H}$ and A, denoted* Acc(A)*, is the image of the endpoint map based at A.*

In other words, the accessible set is the locus swept out by all horizontal paths that pass through A. The terminology "accessible set" is borrowed from control theory. Chow's theorem asserts that if $\mathcal{H}$ is bracket generating on a connected manifold then the accessible set of any point is the entire manifold.

The converse to Chow fails. The converse to Chow is valid for analytic distributions, but false in general. We now give a class of counterexamples. These are distributions on $\mathbb{R}^3$ which are not bracket generating but are horizontally path-connected.

Consider the one-form $\Theta = dz - \alpha_1(x,y)dx - \alpha_2(x,y)dy$ on $\mathbb{R}^3$ with $\alpha_i(x,y)$ smooth functions on the plane. Our distribution will be $\mathcal{H} = \{\Theta = 0\}$ on $\mathbb{R}^3$. Write $B = \partial\alpha_2/\partial x - \partial\alpha_1/\partial y$ for the "magnetic field". Then $d\alpha = Bdx \wedge dy$ and $\Theta \wedge d\Theta = -Bdxdydz$.

If $B \neq 0$ at some point, then only one Lie bracket is needed to generate all of $\mathbb{R}^3$ at that point. The distribution is contact there. So suppose that the vanishing locus $Z = \{(x,y) : B = 0\}$ of the magnetic field is non-empty. If Z is not the entire plane, then the distribution provides horizontal connectivity. To see this, observe, as in section 1.3, that any horizontal curve $\gamma(t) = (x(t), y(t), z(t))$ for $\mathcal{H}$ can be characterized by its planar projection $c(t) = (x(t), y(t))$ and starting height $z(0)$ according to the rule

$$z(t) = z(0) + \int_{c([0,t])} \alpha.$$

The planar curve c can be any absolutely continuous curve in the plane. Thus, for example to connect two points q_0 and q_1 lying over Z by a horizontal curve, draw the line segment c between their planar projections $\pi(q_0)$ and $\pi(q_1)$ and take the horizontal lift γ of this segment, starting from q_0. Unless we are very lucky, we will not have hit q_1. The z coordinate of the endpoint $\tilde{q_1}$ of this lift will be wrong. To fix it, leave the zero locus Z, exiting into the open set $\{B \neq 0\}$ along the horizontal lift of any curve, say another line segment. Fiddle around in this open set, by traveling along the horizontal lifts of small circles in order to climb or descend to the required z-height, and then return to $\pi(q_1)$. The concatenation of the original γ with this new horizontal "fiddling" will connect q_0 to q_1.

If Z is a non-empty planar domain then the bracket-generating condition fails in the domain. Indeed, the distribution is integrable over this domain. Yet any two points can still be connected by a horizontal path.

If the magnetic field vanishes to infinite order at a single point then Z consists of that single point. Horizontal connectivity holds easily for the distribution, but the distribution fails to be bracket generating at the vanishing point.

2.2 A heuristic proof of Chow's theorem

We follow Hermann's proof of Chow [Hermann 1962]. This proof is flawed because it assumes that the accessible set $\mathrm{Acc}(A)$ from the point A is an embedded submanifold.

For a vector field X write $t \mapsto \exp(tX)$ for its local flow. If X is horizontal, we will call its flow a *horizontal flow.* Applying a horizontal flow to any point yields a horizontal curve $t \mapsto \exp(tX)q$. If we stop such a curve at t_1 and then begin again with a different horizontal flow we obtain a horizontal curve $t \mapsto \exp(tY)\exp(t_1X)q$. Continuing in this manner, we see that if $q \in \mathrm{Acc}(A)$ then so is $\exp(t_1X_1)\exp(t_2X_2)\cdots\exp(t_dX_d)q$, for any list $X_1, \ldots, X_d$ of horizontal vector fields, and any list $t_1, \ldots, t_d$ of stopping times small enough so that the flows make sense.

Now let f be a smooth function that is constant on $\mathrm{Acc}(A)$. The previous observation implies that for $q \in \mathrm{Acc}(A)$,

$$f(q) = f(\exp(t_1X_1)\exp(t_2X_2)\cdots\exp(t_dX_d)q).$$

Differentiating this equality with respect to t_1, with the rest of the $t_i = 0$, we have $df(X_i) = 0$ for any horizontal vector field X_i. Differentiating with respect to t_2, and then t_1 with the rest of the $t_i = 0$, we find $X_1X_2[f] = 0$ for any two horizontal vector fields. Here the X_i are thought of as first-order partial differential operators so that $X_i[f] = df(X_i)$, and X_1X_2 is a second-order partial differential operator. Continuing in this manner we find that $X_1 \ldots X_d[f] = 0$ for any list of horizontal vector fields.

But $[X_1, X_2] = X_1X_2 - X_2X_1$ so that $df([X_1, X_2]) = 0$. And $[X_1, [X_2, X_3]] = X_1(X_2X_3 - X_3X_2) - (X_2X_3 - X_3X_2)X_1$ is the sum of products of third-order horizontal operators, so that $df([X_1, [X_2, X_3]]) = 0$. Continuing in this manner we find $df(Y) = 0$ for any Y that is an iterated Lie bracket of horizontal vector fields. But such vector fields Y span the entire tangent space. Hence $df = 0$ and f is constant as claimed.

The problem with this proof is our starting assumption. In fact, the accessible set is generally not an embedded submanifold. For example, an irrational line field on a two-torus has for its accessible sets the leaves, which are immersed dense submanifolds.

The assumption can be altered to make it true. By the *ϵ-accessible set*, for a positive number ϵ, we will mean the image of the endpoint map restricted to horizontal curves of length less than ϵ. Sussmann [1973] has proved that the ϵ-accessible sets are embedded submanifolds for sufficiently small ϵ. (See Appendix D.) This result allows us to turn Hermann's proof into a rigorous one. But Sussmann's proof is significantly harder than the rigorous proof of Chow that we give next, so we do not follow this approach.

2.3 The growth vector and canonical flag

We have been using $\mathcal{H}$ to denote the distribution $\mathcal{H} \subset TQ$. We will now also use the same symbol to denote the sheaf of smooth vector fields tangent to $\mathcal{H}$. Sheaves provide a convenient and necessary language for discussing local properties of systems of vector fields. They are necessary because they allow the rank of the subspaces spanned by the fields to jump, without requiring the fields themselves to be globally defined. (See for example Warner [1971] for the definition of sheaf; a detailed understanding of the definition is not required here.)

Thought of as a sheaf, $\mathcal{H}$ assigns to each open set $U \subset Q$ the collection $\mathcal{H}_U$ of all smooth horizontal vector fields defined on U. $\mathcal{H}$ is a subsheaf of the tangent sheaf TQ of all smooth vector fields on Q.

The Lie brackets of vector fields in $\mathcal{H}$ generate a flag of subsheaves

$$\mathcal{H} \subset \mathcal{H}^2 \subset \cdots \subset \mathcal{H}^r \subset \cdots \subset TQ,$$

with

$$\mathcal{H}^2 = \mathcal{H} + [\mathcal{H}, \mathcal{H}], \quad \mathcal{H}^{r+1} = \mathcal{H}^r + [\mathcal{H}, \mathcal{H}^r]$$

where

$$[\mathcal{H}, \mathcal{H}^k] = \text{span}\left\{[X, Y] : X \in \mathcal{H}, Y \in \mathcal{H}^k\right\}$$

with the span taken over the smooth functions on Q. In other words, $\mathcal{H}^2$ is generated by vector fields in $\mathcal{H}$ and their two-fold brackets $[X, Y]$, $\mathcal{H}^3$ adds the three-fold brackets, and so on. The assumption that $\mathcal{H}$ is bracket generating is equivalent, at least in the case of Q compact, to the assumption that there is an r such that $\mathcal{H}^r = TQ$. Henceforth we will make this assumption.

At a point $q \in Q$, the flag of subsheaves gives a flag of subspaces of T_qQ (the "stalks" of the sheaves):

$$\mathcal{H}_q \subset \mathcal{H}_q^2 \subset \mathcal{H}_q^3 \subset \cdots \subset \mathcal{H}_q^r = T_qQ. \tag{2.1}$$

Definition 2.3.1 *Set $n_i(q) = \dim \mathcal{H}_q^i$. The integer list $(n_1(q), n_2(q), \ldots, n_r(q))$ of dimensions is called the* growth vector *of $\mathcal{H}$ at q. The smallest integer $r = r(q)$ such that $\mathcal{H}_q^r = T_qQ$ is called the* step *or* degree of nonholonomy *of the distribution at q.*

The dimensions $n_i(q)$ may vary from point to point, which is to say that $\mathcal{H}^i$ need not correspond to an actual distribution. In that case $\mathcal{H}^i$ is a sheaf of vector fields which does not arise as the sheaf of sections of a distribution.

Definition 2.3.2 *A distribution $\mathcal{H}$ on a manifold Q is called* regular *at a point $q \in Q$ if the growth vector is constant in a neighborhood of q.*

Note that $n_1 = k$ is the rank of the distribution and $n_r = n$ is the dimension of the manifold. *The growth vector is the most basic numerical invariant associated with a distribution.*

We give three examples now. The first two were described in section 2.1.

Example: The Heisenberg distribution. The Heisenberg distribution on $\mathbb{R}^3$ is spanned by the vector fields $\partial/\partial y$ and $\partial/\partial x + y\partial/\partial z$ and is annihilated by the one-form $dz - ydx$. Its growth vector is $(2, 3)$. To say that a distribution on a three-manifold has this growth is equivalent to saying that it is contact.

Example: The Martinet distribution. The Martinet distribution on $\mathbb{R}^3$ is spanned by the vector fields $\partial/\partial y$ and $\partial/\partial x + (y^2/2)\partial/\partial z$ and is annihilated by the one-form $dz - (y^2/2)dx$. Its growth vector is $(2, 3)$ away from the plane $y = 0$ and is $(2, 2, 3)$ at points of this plane. This example plays a central role in chapter 4.

Example. The distribution on $\mathbb{R}^3$ annihilated by $dz - y^r dx$ has step r, with growth vector $(2, 2, \ldots, 2, 3)$ along the plane $y = 0$, and is of contact type off that plane.

2.4 Chow and the ball-box theorem

We now give a complete, albeit standard, proof of Chow's theorem. It is based on the inverse function theorem and a frame for the whole tangent bundle of the manifold that is built out of a frame for the distribution. It yields a result somewhat stronger than Chow's theorem, which is called the ball-box theorem.

The proof boils down to the idea that commutators of flows of horizontal vector fields allow us to move transversally to the horizontal. Let X_1 and X_2 be smooth vector fields, with (local) flows $\Phi_i(t) = \exp(tX_i)$. We recall that for small t,

$$\Phi_1(t) \circ \Phi_2(t) \circ \Phi_1(t)^{-1} \circ \Phi_2(t)^{-1}(q) = q + t^2[X_1, X_2](q)$$

in any coordinate system. For brevity we will write

$$[\Phi_1(t), \Phi_2(t)] = \Phi_2(t)^{-1} \circ \Phi_1(t)^{-1} \circ \Phi_2(t) \circ \Phi_1(t)$$

for this commutator of flows. We will also use the fact that

$$\Phi_i(t)^{-1} = \Phi_i(-t).$$

Fix a base point $q_0 \in Q$ and choose a local orthonormal frame X_i, $i = 1, \ldots, k$, for our distribution $\mathcal{H}$. Let Φ_i be the corresponding local flows. We use them to move easily in the horizontal directions,

$$\Phi_i(t)(q) = q + tX_i(q) + O(t^2),$$

for t small. Let us call the curves $t \mapsto \Phi_i(t)(q)$ *simple horizontal curves.* Note that the length of a simple horizontal curve with $0 \le t \le \epsilon$ is simply ϵ. By applying $\Phi_k(t_k) \circ \cdots \circ \Phi_2(t_2) \circ \Phi_1(t_1)$ to q_0 and letting the t_i vary over the k-cube $|t_i| \le \epsilon$ we sweep out a k-dimensional curvilinear cube tangent to $\mathcal{H}_{q_0}$ at $t = 0$ with k-dimensional volume approximately $(2\epsilon)^k$. Each point in this cube is the endpoint of a concatenation of k or fewer simple curves: first travel along $\Phi_1(t)(q_0)$ from 0 to t_1, set $q_1 = \Phi_1(t_1)$, then travel along $\Phi_2(t)(q_1)$ from $t = 0$ until $t = t_2$, and so on. Each of these simple horizontal curves has length less than ϵ, so the cube sits in the subriemannian ball of radius $k\epsilon$.

We may move in the $\mathcal{H}^2/\mathcal{H}$ directions along horizontal paths by applying the commutators $\Phi_{ij}(t) = [\Phi_i(t), \Phi_j(t)]$ to q_0. Now $\Phi_{ij}(t)(q_0) = q_0 + t^2[X_i, X_j](q_0)$, so if we restrict to $|t| \le \epsilon$ we will move by a *Euclidean* amount ϵ^2 in the $\mathcal{H}^2/\mathcal{H}$ direction.

We continue the process of taking commutators and brackets until we have exhausted the tangent space. The implementation of this idea is mostly a matter of notation. For multi-indices $I = (i_1, i_2, \ldots, i_m)$, $1 \le i_j \le k$, define vector fields X_I inductively by $X_I = [X_{i_1}, X_J]$, where $J = (i_2, i_3, \ldots, i_m)$. We'll write $i_1 J = I$ and denote the length of a multi-index I by $|I|$, so $|J| = m-1$. Similarly define flows Φ_I by $\Phi_I(t) = [\Phi_{i_1}(t), \Phi_J(t)]$. Observe that

$$\Phi_I(t) = \mathbf{1} + t^m X_I + O(t^{m+1}).$$

By the bracket-generating assumption we can select a local frame for the entire tangent bundle from amongst the X_I. We choose such a frame and relabel it Y_i, $i = 1, \ldots, n$, to respect the canonical filtration: $\{Y_1 = X_1, \ldots, Y_k = X_k\}$ span $\mathcal{H}$ near q_0; $\{Y_1, \ldots, Y_{n_2}\}$ span $\mathcal{H}^2$ near q_0; $\{Y_1, \ldots, Y_{n_2}, \ldots, Y_{n_3}\}$ span $\mathcal{H}^3$; and so on, where $(k, n_2, n_3, \ldots, n_r)$ is the growth vector of the distribution at q_0.

Weighting. For each chosen Y_i of the form X_I, let w_i be the length $|I|$. Thus $w_i = m$ if and only if $Y_i \in \mathcal{H}^m$ and $Y_i \notin \mathcal{H}^{m-1}$. The assignment $i \mapsto w_i$ is called the *weighting* associated to the growth vector.

Similarly, we relabel the flows Φ_I as Φ_i, $i = 1, \ldots, n$, for those multi-indices arising in the construction of our frame Y_i. Now each point $\Phi_i(t)(q_0) = \Phi_I(t)(q_0)$ is the endpoint of a horizontal path consisting of the concatenation of w_i simple horizontal paths, each one of length t. So if we restrict t to $|t| \le \epsilon$ then $\Phi_i(t)(q_0)$ lies in the ball of radius $w_i\epsilon$ about q_0. On the other hand, since

$$\Phi_i(t)(q_0) = q_0 + t^{w_i} Y_i(q_0) + O(t^{w_i+1}),$$

this point lies in the Euclidean box whose dimensions are of the order ϵ^{w_i} in the $\mathcal{H}^{w_i}$ directions. This suggests that the subriemannian ball of radius ϵ contains a Euclidean coordinate box whose sides are of order ϵ^{w_i} in the i-th coordinate direction. This result, called the ball-box theorem, will be proved after the following definition.

Definition 2.4.1 *Coordinates $y_1, \ldots, y_n$ are said to be* linearly adapted *to the distribution $\mathcal{H}$ at q_0 if $\mathcal{H}^i(q_0)$ is annihilated by the differentials $dy_{n_i+1}, \ldots, dy_n$*

The x-axis is a horizontal curve for $\mathcal{H}$. Consider a finite arc γ_0 of this axis from the origin to a point $(x_0, 0, 0)$, parameterized in the standard way: $\gamma_0(t) = (t, 0, 0)$, $0 \leq t \leq x_0$. Let $\gamma(t) = (x(t), y(t), z(t))$, $0 \leq t \leq \tau$, be any other horizontal curve with the same endpoints. Integrating the differential constraint $\Theta = 0$ yields

$$0 = z(\tau) - z(0) = \frac{1}{2}\int y^2 dx,$$

where the line integral is over the projection of the curve to the xy-plane. Now suppose that $dx/dt > 0$ along γ. Then the integrand $y^2 dx = y^2 \dot{x} dt$ is positive unless y is identically zero. If y is identically zero then the curve is simply a reparameterization of our segment γ_0 of the x-axis. This proves that there is a C^1-neighborhood of our original curve γ_0 such that every horizontal curve in this neighborhood that shares endpoints with γ_0 is a reparameterization of γ_0. Curves γ_0 with this property will be called *C^1-rigid*, following Bryant and Hsu [1993].

3.2 Martinet's genericity result

We put the example of a rigid curve in its general context. Let ξ be a two-plane field on a three-manifold. Let θ be any nonvanishing one-form that annihilates ξ. Then the contact condition is $\theta \wedge d\theta \neq 0$.

Suppose instead that $\theta \wedge d\theta(p) = 0$ at some point p. Choose a volume form d^3x in a neighborhood of p, and use it to define a function f according to $\theta \wedge d\theta = f d^3x$. Then we have $f(p) = 0$. Let us make the *nondegeneracy assumption* $df \wedge \theta(p) \neq 0$. This implies in particular that $df(p) \neq 0$ so that the set $\Sigma := \{f = 0\}$ is a smooth surface near p, by the implicit function theorem. The nondegeneracy assumption asserts that the plane ξ_p is transverse to the tangent space to Σ at p, and consequently the planes ξ_q for nearby $q \in \Sigma$ remain transverse. Their intersections with $T_q\Sigma$ define a line field on Σ near p.

Definition 3.2.1 *We call θ a* Martinet form, *the distribution $\{\theta = 0\}$ a* Martinet distribution, *and the surface $\Sigma = \{f = 0\}$ the* Martinet surface. *The integral curves of the line field obtained by intersecting the tangent planes to this surface with the distribution planes are called the* Martinet curves.

Martinet [1970] proved the following theorem.

Theorem 3.2.2 (Martinet normal form) *Let θ be a Martinet form near p. Then there exist coordinates (x, y, z) centered at p and a positive function g such that*

$$g\theta = dz - \frac{1}{2}y^2 dx.$$

In the normal form coordinates, the Martinet surface is given by $y = 0$ and the Martinet curves are parallels to the x-axis which lie in this surface. By the analysis of section 3.1, these Martinet curves are C^1-rigid, at least in any neighborhood in which Martinet's normal form theorem holds.

3.3 The minimality theorem

Being C^1-isolated curves, the Martinet curves γ_0 are automatically local minima *in the C^1-topology* for *any* functional. Hence our main theorem may not be a surprise.

Theorem 3.3.1 (Minimality theorem) *Let g be a subriemannian structure whose underlying distribution is a Martinet distribution, and let p be any point on the Martinet surface. Then there is an $x_* > 0$ (depending on g and p) such that the arc γ_0 of the Martinet curve starting at p and of length x_* is the unique minimizing geodesic joining p to its endpoint.*

The proof of this minimization theorem is surprisingly difficult. The problem is that we are obliged to consider *all* curves joining the given endpoints and for which the length makes sense, and there is no reason to restrict ourselves, a priori, to some C^1-neighborhood of a given candidate minimizer. In other words, the C^1-topology is not the correct topology for the calculus of variations.

Theorem 3.3.2 *For a generic metric g on the Martinet distribution, the minimizing curve of Theorem 3.3.1 is not the projection of any solution to Hamilton's equations for the subriemannian Hamiltonian.*

Theorems 3.3.1 and 3.3.2 say that there are geodesics that do not satisfy the geodesic equations.

Definition 3.3.3 *A minimizing subriemannian geodesic that is not the projection of an integral curve for the subriemannian Hamiltonian is called a* singular minimizer.

By Martinet's normal form theorem, his distribution is topologically stable: small perturbations of it are still Martinet and hence locally diffeomorphic to its normal form. Thus the singular minimizers of the theorem are topologically stable. In other words, they persist under perturbations of the subriemannian structure.

3.4 The minimality proof of Liu and Sussmann

We prove Theorem 3.3.1 following Liu and Sussmann [1995]. Consider any subriemannian metric g on a Martinet distribution. Choose a g-orthonormal horizontal frame X, Y, with X tangent to the Martinet curves. Then Y is orthogonal and, in particular, transverse to the Martinet surface Σ. Given any surface Σ in a three-manifold, and any transverse vector field Y, we can alway find local coordinates (x, y, z) such that $\Sigma = \{y = 0\}$ and $Y = \partial/\partial y$ in a neighborhood of any point of Σ. These can be constructed using flow-box coordinates, with the box chosen to have sides parallel to Σ.

The defining property of the coordinates is unchanged under coordinate changes $\bar{x} = \bar{x}(x, z)$, $\bar{y} = y$, $\bar{z} = \bar{z}(x, z)$ independent of the y variable. We can use such a coordinate change to "rectify" X along Σ, so that $X|_\Sigma = \partial/\partial x$ and

$$X = (1 + y\psi_1)\frac{\partial}{\partial x} + y\psi_2\frac{\partial}{\partial y} + y\psi_3\frac{\partial}{\partial z}$$

for some smooth functions ψ_i. Now $[Y, X]|_\Sigma = \psi_1\partial/\partial x + \psi_2\partial/\partial y + \psi_3\partial/\partial z$, from which it follows that the Martinet surface is also defined by $\psi_3 = 0$. Therefore $\psi_3 = y\phi$ for some smooth function ϕ. The Martinet nondegeneracy condition is $d\psi_3(Y) \neq 0$, hence $\phi \neq 0$. Finally, a scaling $z \to \lambda z$, with λ a (possibly negative) constant, insures that $\phi(0) = 1$. This is the partial normal form that we will use to prove the minimality theorem. Note that in these coordinates the x-axis is still the Martinet curve passing through the origin, and it is parameterized by arc length. The minimality theorem is now restated as the optimality lemma:

Lemma 3.4.1 (Optimality lemma) *There is a positive constant x_* depending only on the ψ_i and ϕ of the normal form such that for all positive $x_1 \leq x_*$ the arc of the x-axis from $(0,0,0)$ to $(x_1, 0, 0)$ is the unique minimizing geodesic joining its endpoints.*

Proof. Let $\gamma(t) = (x(t), y(t), z(t))$, $0 \leq t \leq \tau$, be an a.e. horizontal curve leaving the origin (i.e. a possible competitor). Without loss of generality, we may assume that its subriemannian speed is 1 a.e. Thus, $\dot\gamma = u_1X + u_2Y$, with $u_1^2 + u_2^2 = 1$ a.e., and the length of γ is τ.

In our normal coordinates, the derivative $\dot\gamma$ is

$$\begin{aligned} \dot x &= (1 + y\psi_1)u_1 \\ \dot y &= y\psi_1 u_1 + u_2 \\ \dot z &= y^2\phi u_1 \end{aligned} \tag{3.1}$$

We will sometimes refer to the u_i as the *controls* of our curve γ. Integrating, and using the condition that $\gamma(0) = 0$, we find

$$\begin{aligned} x(t) &= \int^t (1 + y\psi_1)u_1 \\ y(t) &= \int^t y\psi_1 u_1 + u_2 \\ z(t) &= \int^t y^2\phi u_1 \end{aligned}$$

where we have used the shorthand $\int^t f$ for $\int_0^t f(x(t), y(t), z(t))dt$. We will use this shorthand throughout the rest of the section.

Choose some relatively compact neighborhood K of the origin that is contained in our coordinate chart, for example a small subriemannian ball, and set $k = d(0, \partial K)$, the subriemannian distance from the origin to the boundary of K. Fix positive constants C_1, C_2, C_3 such that for all points q of $\bar K$ we have

$$\begin{aligned} &|\psi_1(q)| \leq C_1, \\ &|y\psi_1 u_1 + u_2| \leq C_2 \quad \text{when } u_1^2 + u_2^2 \leq 1, \\ &|1 - \phi(q)| \leq \epsilon_3. \end{aligned}$$

We may assume that $\epsilon_3 < 1$, by shrinking K if necessary, because $\psi_1(0) = 1$. Note that the second inequality bounds $\dot{y}$. At the end of the proof the constant x_* will be determined in terms of the C_i and ϵ_3. It will also be taken to be less than k so that all competing curves γ may be taken to lie within K. (If they left K they would be longer than k and hence than γ_0.)

Impose the endpoint conditions $y(\tau) = 0$ and $z(\tau) = 0$ on our competing curve. (Remember that we assume $x_* < k$ and so we may assume that $\ell(\gamma) \leq k$ and hence that all the estimates above involving the C_i hold on γ.)

Claim 0: $x(\tau) \leq \tau$ provided $\tau \leq x_*$.

Establishing this claim will complete the proof because $\ell(\gamma) = \tau$, but for our singular curve γ_0 we have $x_1 = t_1 = \ell(\gamma_0)$, so the endpoint condition $x(\tau) = x_1$ yields $\ell(\gamma_0) \leq \ell(\gamma)$.

Set

$$h(t) = \int^t u_1,$$

where u_1 is the first control for our competing curve, and set $\alpha = \tau - h(\tau)$. Note that $\alpha \geq 0$ since $|u_1| \leq 1$. Also set

$$\beta = \sup_{y \in I} |y(t)|,$$

where $I = [0, \tau]$ is the domain of our competing curve.

Claim 1: $x(\tau) \leq \tau - \alpha + C_1\beta\tau$.

Proof of Claim 1:

$$\begin{aligned} |x(t)| &= \int^t u_1 + \int y\psi_1 u_1 \\ &\leq h(t) + \left| \int^t y\psi_1 u_1 \right| \\ &\leq h(t) + \beta C_1 \int^t |u_1| \\ &\leq h(t) + \beta C_1 t \end{aligned}$$

where the last inequality results from the fact that $|u_1| \leq 1$. Now evaluate at $t = \tau$ and use the definition of α.

Looking at Claim 1, we see that to prove Claim 0, and hence the lemma, it suffices to show that $-\alpha + C_1\beta\tau$ is negative, i.e. that $C_1\beta\tau \leq \alpha$. Now if $\tau \leq x_*$ we certainly have $C_1\beta\tau \leq C_1\beta x_*$. Thus it suffices to prove that $C_1\beta x_* \leq \alpha$. Multiplying both sides by β^2 yields Claim 2:

Claim 2: $C_1 x_* \beta^3 \leq \beta^2\alpha$.

This claim will be proved by comparing both sides with $\int^\tau y^2$.

Claim 3:

$$k_-\beta^3 \leq \int^\tau y^2 \leq k_+\beta^2\alpha,$$

with

$$k_- = \frac{2}{3C_3}, \quad k_+ = \frac{1}{1-\epsilon_3}.$$

Claim 0 and the lemma then follow by setting $C_1 x_* = k_-/k_+$, i.e. $x_* = (1-\epsilon_3)/(C_1C_3)$.

Proof of the upper bound in Claim 3: We will use the z endpoint condition $\int^\tau y^2\phi u_1 = 0$. Then

$$\begin{aligned} \int^\tau y^2 &= \int^\tau y^2(1-u_1) + \int^\tau y^2 u_1 \\ &= \int^\tau y^2(1-u_1) + \int^\tau y^2(u_1 - \phi u_1) \\ &\leq \beta^2 \int^\tau (1-u_1) + \int^\tau y^2|1-\phi||u_1|. \end{aligned}$$

Now $\int^\tau (1-u_1) = h(\tau) = \alpha$ and $|1-\phi||u_1| \leq \epsilon_3$. Thus

$$\int^\tau y^2 \leq \beta^2\alpha + \epsilon_3 \int^\tau y^2$$

or

$$(1-\epsilon_3)\int^\tau y^2 \leq \beta^2\alpha.$$

which is the desired upper bound.

Proof of the lower bound in Claim 3: Since $y(t)$ is continuous there is a $t_0 \in I$ such that $|y(t_0)| = \beta$. For simplicity, assume $y(t_0) = \beta$. (The argument is nearly identical in case $y(t_0) = -\beta$. The sign and direction of the "tent function" f below needs to be reversed.) Since $\dot{y}_3 \leq C_3$ we have that $y(t) \geq f(t)$ where $f(t)$ is the piecewise linear "tent function" whose maximum is at β and is increasing with slope C_3 up to $t = t_0$ and decreasing with slope $-C_3$ afterwards. (See Figure 3.1.) The zeros of $f(t)$ are $t_\pm = t_0 \pm \beta/C_3$. These must be inside our interval $I = [0,\tau]$ because $y \geq f$ and $y(0) = y(\tau) = 0$. Write $J = J_- \cup J_+ = [t_-, t_0] \cup [t_0, t_+]$. Then

$$\int y^2 \geq \int_J f^2.$$

But

$$\int_J f^2 = 2\int_{J_-} f^2 = 2\int_0^{\beta/C_3} (C_3 t)^2 dt = \frac{2\beta^3}{3C_3},$$

yielding the desired result:

$$\frac{2\beta^3}{3C_3} \leq \int^\tau y^2.$$

QED

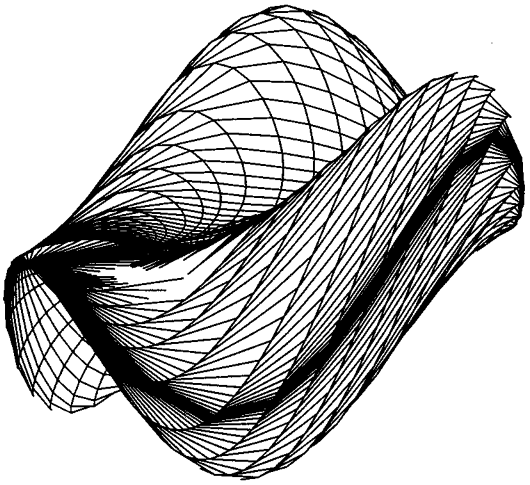

Figure 3.1: The Martinet sphere.

3.5 Failure of geodesic equations

Recall from the previous section that we have coordinates x, y, z such that our horizontal vector field X, tangent to the Martinet surface, has x-component $(1 + y\psi_1)$ with ψ_1 vanishing at the origin. Theorem 3.3.2 can be restated as the following lemma.

Lemma 3.5.1 *If $\psi_1(0) \neq 0$, then the minimal geodesic of Theorem 3.3.1 – a small arc of the x-axis – is not the projection of a solution to Hamilton's equations.*

Proof. We prove that if the x-axis is the projection of a solution to Hamilton's equations for H_0 then ψ_1 must be identically zero along this axis.

By construction, the axis is parameterized by arc length. It satisfies the equations $\dot{x} = 1$ and $\dot{y} = 0$. On the other hand, if it is a solution then we have $\dot{x} = \{x, H_0\}$ and $\dot{y} = \{y, H_0\}$, where $H_0 = (P_X^2 + P_Y^2)/2$ is the subriemannian Hamiltonian.

Now $P_X = p_x + y(\psi_1 p_x + \psi_2 p_y + y\phi p_z)$ so $P_X = p_x$ and $P_Y = p_y$ along the Martinet surface $y = 0$. It follows from the $\dot{x}$ and $\dot{y}$ equations that $P_X = 1$ and $P_Y = 0$ along the x-axis. For solutions parameterized by arc length, like our x-axis, we have $H_0 = \frac{1}{2}$ and so we can write $P_X = \cos(\phi)$ and $P_Y = \sin(\phi)$, thus defining the angle ϕ. Geometrically, ϕ is the angle between the projected solution curve γ and the vector field X. One computes

$$\frac{d}{dt}P_X = \{P_X, H\} = -P_Z P_Y, \quad \frac{d}{dt}P_Y = \{P_Y, H\} = P_Z P_X,$$

where P_Z is the momentum function for the vector field $Z = [X, Y]$. It follows from these equations that $\dot{\phi} = -P_Z$.

The x-axis, being an integral curve of X, satisfies $\phi \equiv 0$. So, for the axis to be the projection of a solution we must have that $P_Z = 0$. However, on the Martinet surface we have $Z = -(\psi_1 \partial/\partial x + \psi_2 \partial/\partial y + y\phi\partial/\partial z)$, so along the x-axis we have $P_Z = -(\psi_1 p_x) = -\psi_1$. This proves that we must have $\psi_1 \equiv 0$ along the axis in order for the axis to be the projection of a solution.

QED

3.6 Singular curves in higher dimensions

The Martinet curves generalize in higher dimensions.

Theorem 3.6.1 (Liù-Sussmann, Bryant-Hsu) *Let $\mathcal{H}$ be a rank two distribution on an n-manifold. Suppose that $\mathcal{H} \neq \mathcal{H}^2$ as sheaves. Then there exist singular minimizers that are C^1-rigid. Passing through points at which the growth vector is either $(2, 3, 4, \ldots)$ or $(2, 3, 5, \ldots)$ there is an $(n-4)$-parameter family of C^1-rigid singular minimizers.*

Liu and Sussmann [1995] proved the minimality statement in this theorem. In fact, the proof of Theorem 3.3.1 given in section 3.4 is their proof of Theorem 3.6.1 simplified to three dimensions. Hsu [1991] established the existence of the $(n-4)$-parameter family of rigid curves. More details concerning this theorem, and in particular the $n-4$ parameters, are discussed in section 5.4.

3.7 There are no H^1-rigid curves

Theorem 3.7.1 *A bracket-generating distribution admits no H^1-rigid curves.*

Proof. According to Theorem E.0.1 in Appendix E, the endpoint map is an open mapping from the space of H^1-curves to the manifold.

Given a curve γ, pick an intermediate point along the curve, say $P = \gamma(t_2)$. Divide γ into two arcs, γ_1 going from the initial point to P, and γ_2 from P to $\gamma(1)$. Apply the open mapping result separately to γ_1 and γ_2^{-1}. We get open neighborhoods of curves, N_1 containing γ_1 and N_2 containing γ_2^{-1}, whose endpoints fill up a small neighborhood of the midpoint P. Take curves γ_1^ϵ in N_1 and $(\gamma_2^\epsilon)^{-1}$ in N_2 that are ϵ-close in the H^1 norm to their respective arcs, and that both end at the same point P_ϵ, very close to P. Then the concatenation $\gamma_2^\epsilon * \gamma_1^\epsilon$ is a curve that is $O(\epsilon)$-close to γ in the H^1 topology, and that does not coincide with it.

QED

Question. Is this theorem true for any distribution, bracket generating or not? I believe so. Here is why. First, take the case of an involutive distribution. We may as well work on the leaf containing the curve in question, and the theorem is true there. Second, the theorem holds for general analytic distributions. To see this, apply Sussmann's theorem (Theorem D.1.4 of Appendix D), which asserts that the accessible set is a nice immersed submanifold. Any horizontal curve lies in the accessible set of its starting point. Restrict attention to this accessible submanifold. In the analytic case the restricted distribution is bracket generating within this submanifold. Apply Theorem 3.7.1.

3.8 Towards a conceptual proof?

Various proofs of minimality of the Martinet curves exist, but to my mind none of them are satisfactory. They do not explain the reason behind the phenomenon.

The most satisfactory "proof" is not a proof at all: The Martinet curve is isolated in the C^1 topology (C^1-rigidity). Consequently, it is a local minimum, relative to this topology, for any functional, and in particular for the length functional.

The problem with this "proof" is that the C^1-topology is not the correct topology for the calculus of variations. We are interested in all horizontal curves joining two given endpoints, not just those C^1-close to some given curve. The correct topology is the H^1 topology, in which two curves starting from the same point are close if their derivatives (with respect to a horizontal frame) are L^2-close. There are no H^1-isolated horizontal curves, and in particular the Martinet curve is not H^1-isolated.

A satisfactory proof might proceed as follows. Consider the variety of all horizontal curves that share endpoints with the Martinet curve, endowed with the H^1 topology. The Martinet curve is a singular point of this variety. What does the singularity look like? Can we show that it is a very "sharp" singularity, with this "sharpness" increasing as we take shorter initial segments of the Martinet curve? Imagine a very sharp cone. Then almost any linear function will have a minimum (or maximum) at the cone point (see Figure 3.2). A proof along these lines should add much understanding.

3.9 Notes

False theorems asserting that every subriemannian minimizer is normal appeared in the literature between 1967 and 1990, and almost certainly earlier (see for example [Rayner 1967; Hamenstädt 1990; Strichartz 1986, 1989; Taylor 1989; Bär 1998]). My counterexample to this assertion, the singular minimizer described in this chapter, was discovered in 1991. It did not appear in print for several years [Montgomery 1994a]. My construction was based on intuition about the motion of a particle in a magnetic field. Soon after, Liu and Sussmann [1995] published a much simpler proof and more general results, which we have presented here.

Strong hints of the existence of singular minimizers appeared much earlier. Carathéodory and Hilbert were quite familiar with the rigidity phenomenon (see [Young 1980]). Bismut [1984] clearly points out the existence of singular minimizers. Baillieul [1978] and Brockett [1984] left the question of existence

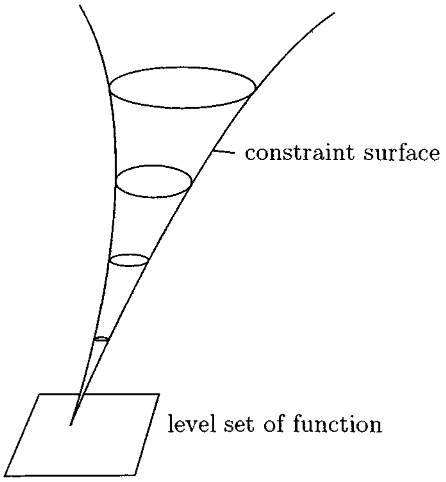

Figure 3.2: The singularity in path space.

open. Gaveau [1977] claimed to have found a singular minimizer, but his work was later shown to be in error by Brockett [1984].

The mistake made by most of those who claimed to have proved that all subriemannian minimizers are normal was a mis-application of the maximum principle of Pontryagin et al. [1962]. This principle, when stated as a theorem, becomes quite complicated and typically takes a full page of text. In essence the principle is a version of the method of Lagrange multipliers for finding constrained minimizers. The mistake boils down to a misuse of that method. Suppose we want to minimize F subject to a constraint $G = 0$. We tell our calculus students to find the critical points of $F + \lambda G$, where λ is the Lagrange multiplier. This is the wrong approach because it can miss the minimizer. Instead we should look for the critical points of the function $\lambda_0 F + \lambda G$, where the multipliers λ_0 and λ are not both zero. The possibility of a critical point with $\lambda_0 = 0$ allows for minimizers to be singular points of the constraint hypersurface, i.e. singular minimizers.

The affair of Gaveau is interesting. Gaveau [1977] claimed that the free two-step Carnot group over $\mathbb{R}^4$ admits singular geodesics. This Carnot group is a subriemannian geometry of bundle type on $\mathbb{R}^4 \oplus \bigwedge^2 \mathbb{R}^4$. Gaveau's "proof" that there is a singular minimizer was based on his faulty assertion that no normal geodesic can be found to connect $(0, 0) \in \mathbb{R}^4 \oplus \bigwedge^2 \mathbb{R}^4$ to an element of the form $(0, z)$; consequently the minimizer between these two points must be a singular one. Afterward, though, Brockett [1984] exhibited an explicit normal extremal connecting these points (see chapter 12). Decades later Golé and Karidi [1995] proved that in a two-step Carnot group every subriemannian geodesic is normal, and then constructed the first example of a Carnot group with a singular minimizer. The Golé-Karidi example is a three-step Carnot geometry in the fibers of $\mathcal{H}^{\perp}$. In section 5.6 we discuss a class of distributions, the *fat distributions*, for which there are no singular curves.

5.1 The space of horizontal paths

The endpoint map takes a horizontal rectifiable curve to its endpoint (see section 2.1). A *singular curve* is a critical point for the endpoint map. For this definition to make sense, we need to know that the endpoint map is differentiable with respect to curves. In particular, the space Ω of curves, which is the domain of the endpoint map, must be endowed with a differentiable structure. We describe this structure.

Denote the underlying manifold with distribution by $(Q, \mathcal{H})$. Fix an interval $I = [a, b]$ and a beginning point $q_0 \in Q$. Write $\Omega = \Omega(I, q_0; \mathcal{H})$ for the space of all curves that start at q_0 and whose derivative is square integrable. In other words, a curve $\gamma : I \to Q$ is in Ω if $\gamma(a) = q_0$, if the curve is absolutely continuous and if its derivative (guaranteed to exist a.e.) is square integrable with respect to some (and hence any) subriemannian metric on Q. (Recall that a continuous curve is *absolutely continuous* if it has a derivative almost everywhere and it can be reconstructed, in any coordinate chart, by integrating the curve of its derivatives. See for example [Royden 1968].) The *endpoint map* is then the map end : $\Omega([a, b], q_0; \mathcal{H}) \to Q$ defined by $\operatorname{end}(\gamma) = \gamma(b)$.

To describe the differentiable structure on Ω we suppose for the moment that $\{X_1, \dots, X_k\}$ form a global orthonormal frame for $\mathcal{H}$ and that the X_a are complete vector fields, meaning that their flows are defined for all time. Expand the derivative of a curve γ in terms of this frame: $\dot{\gamma} = \sum u^a X_a = u \cdot X$. The u are the coordinates of our curve. If $\dot{\gamma}$ is square integrable, then $u = (u^1, \dots, u^k)$ is in $L^2(I, \mathbb{R}^k)$.

To insure that the u are good coordinates we need to show that the map $\gamma \to u$ is invertible. This is done in Appendix E, where we prove that the initial value problem

$$\dot{\gamma} = u \cdot X(\gamma), \quad \gamma(a) = q_0 \tag{5.1}$$

is *well posed*. This means that for each u and q_0 there is a unique solution $\gamma = \gamma(u; q_0)$ that depends smoothly on u and q_0. Moreover, for each fixed $t \in I$ the map $L^2(I, \mathbb{R}^k) \times Q \to Q : (u, q_0) \mapsto \gamma(u; q_0)(t)$ is smooth. Note that the map $(u, q_0, t) \mapsto \gamma(u, q_0)(t)$ cannot be smooth, since the vector field $u \cdot X$ need only be L^2 in the time variable t. Freezing q_0 and solving the corresponding initial value problem defines the endpoint map relative to these X-based coordinates u for Ω. The result is the coordinate version

$$L^2(I, \mathbb{R}^k) \to Q : u \mapsto \gamma(b)$$

of the endpoint map.

We need to surmount various technical difficulties to insure that this idea yields a manifold structure for the horizontal paths on a general manifold with a general distribution. For example, $\mathcal{H}$ need not admit a global frame. Even if it did, the vector fields in the frame need not be complete. All of these difficulties are overcome in Appendix E.

Different choices of frame yield different charts for Ω. Two different charts are related by a smooth change of variables, regardless of whether or not the frames are orthonormal.

Definition 5.1.1 *A* singular curve *is a critical point for the endpoint map. A* regular curve *is a is a regular point for the endpoint map.*

At first glance, this definition may seem to depend on the choice of metric on the distribution planes, since we need an inner product on these planes to define square-integrability in the definition of Ω. However, a different subriemannian metric on the same $\mathcal{H}$ yields the same set Ω of curves. This is because any two metrics along a fixed curve are Lipschitz related. Therefor, the concept of *singular curve* depends only on the distribution of k-planes and not on the inner product put on these planes.

In Riemannian geometry, $\mathcal{H} = TQ$ and there are no singular curves. On the other hand, "most" subriemannian geometries have singular curves. The Martinet curves of chapter 3 are perfect examples of singular curves. We saw that they are minimizing geodesics regardless of how we measure length within the distribution planes. That is, they are geodesics by virtue of their singular nature alone. The existence of such singular minimizers is one of the chief differences between Riemannian and subriemannian geometry. Because of their existence the following two problems are *open* in subriemannian geometry. *Are all minimizing geodesics smooth? Are all sufficiently small subriemannian balls homeomorphic to the usual Euclidean ball?* For more about these problems, see chapter 10.

5.2 A microlocal characterization

5.2.1 Characteristics

We will need a computational method for finding singular curves. Hsu [1991] developed such a tool, the method of characteristics, in his thesis. One finds the same tool, in a less intrinsic form, in the work of Rayner [1967, app.] and Pontryagin et al. [1962].

The method of characteristics requires the subbundle $\mathcal{H}^{\perp} \subset T^*Q$ of all one-forms that annihilate $\mathcal{H}$. $\mathcal{H}^{\perp}$ is a linear subbundle of rank $n-k$. A typical element of $\mathcal{H}^{\perp}$ will be written λ, or (q, λ) when we want to emphasize the base point $q \in Q$ of $\lambda \in \mathcal{H}^{\perp}_q$. Recall that T^*Q admits a canonical symplectic form, written $\sum dp_i \wedge dq^i$ in canonical coordinates (see Definition A.4.2 in Appendix A). Let ω denote the restriction of this form to $\mathcal{H}^{\perp}$. This restriction need not be symplectic, and hence it might admit *characteristics.*

Definition 5.2.1 *A* characteristic *for $\mathcal{H}^{\perp}$ is an absolutely continuous curve $\lambda(t) \in \mathcal{H}^{\perp}$ that never intersects the zero section of $\mathcal{H}^{\perp}$ and that satisfies $i_{\dot{\lambda}(t)}\omega = 0$ on $T_{\lambda}\mathcal{H}^{\perp}$ at every point t for which the derivative $\dot{\lambda}(t)$ exists.*

Our computational tool is the following theorem of Hsu [1991] and Pontryagin et al. [1962].

Theorem 5.2.2 *A curve $\gamma \in \Omega$ is singular if and only if it is the projection of a characteristic λ for $\mathcal{H}^{\perp}$ with square-integrable derivative.*

The square-integrable conditions on Ω and on the characteristics are imposed for applications to the study of geodesics. We could just as easily talk about curves and characteristics with derivative in L^p.

Let us see what the proposition says in computational terms. Fix a local frame $\theta = (\theta^1, \ldots, \theta^s)$ for $\mathcal{H}^{\perp}$ defined in some neighborhood U of Q. Any $\lambda \in \mathcal{H}^{\perp}$ lying over U can be expanded uniquely as

$$\lambda = \sum \lambda_a \theta^a. \tag{5.2}$$

This equation defines fiber coordinates λ_a, $a = 1, \ldots, s$, on $\mathcal{H}^{\perp}$.

On the other hand, we can think of the θ^a as one-forms on $\mathcal{H}^{\perp}$ by pulling them back from Q by the projection $\pi : \mathcal{H}^{\perp} \to Q$. Then equation 5.2 is the expression for the restriction Θ of the canonical one-form $\sum p_i dq^i$ on T^*Q to $\mathcal{H}^{\perp}$. Indeed, let us complete the θ^a to a frame for all of T^*Q by adding k independent one-forms η^i, $i = 1, \ldots, k$. Then $\Theta = \sum \lambda_a \theta^a + \sum \lambda_i \eta^i$ where (λ_a, λ_i) are the fiber coordinates on T^*Q defined by the full frame. The restriction of Θ to $\mathcal{H}^{\perp}$ is obtained by setting all the λ_i to zero. Since ω is the differential of the restriction of this one-form, we have

$$\omega = d\theta = \sum d\lambda_a \wedge \theta^a + \lambda_a d\theta^a. \tag{5.3}$$

Now, re-expand the $d\theta^a$ in terms of our frame,

$$d\theta^a = \sum c^a_{\mu\nu} \eta^\mu \eta^\nu + \sum c^a_{\mu c} \eta^\mu \theta^c + \sum c^a_{bc} \theta^b \theta^c, \tag{5.4}$$

thus defining the structure functions c^a_{bc}, $c^a_{\mu c}$, $c^a_{\mu\nu}$.

Let X_a, X_μ be the frame for TQ that is dual to our frame θ^a, η^μ. Thus the X_μ frame $\mathcal{H}$. Then $X_a, X_\mu, \partial/\partial\lambda_a$ form a basis for the vectors on $\mathcal{H}^{\perp}$, so that given any curve $\lambda(t)$, differentiable at t, we may expand

$$\dot{\lambda}(t) = \dot{\lambda}_a \frac{\partial}{\partial \lambda_a} + \dot{\gamma}^a X_a + \dot{\gamma}^\mu X_\mu. \tag{5.5}$$

Lemma 5.2.3 *An absolutely continuous curve λ with square-integrable derivative is a characteristic if and only if, in the notation of equations 5.2, 5.4, and 5.5, the following equations hold a.e.:*

$$\dot{\gamma}^a = 0 \tag{5.6}$$

$$\dot{\lambda}_a + \sum \lambda_b c^b_{\mu a} \dot{\gamma}^\mu = 0 \tag{5.7}$$

$$\sum \lambda_a c^a_{\mu\nu} \dot{\gamma}^\mu = 0 \tag{5.8}$$

Proof. Expand the one-form $i_{\dot\lambda}\omega$ in terms of the frame $d\lambda_a, \theta^a, \eta^\mu$ of covectors for $\mathcal{H}^\perp$, using equations 5.3 and 5.5. The coefficient of $d\lambda_a$ is $-\dot\gamma^a$, yielding equation 5.6. The coefficient of θ^a is the left-hand side of equation 5.7, and the coefficient of η^ν is the left-hand side of equation 5.8.

QED

The equations in Lemma 5.2.3 form a mixed algebraic-differential system of equations, called the *characteristic equations.* Equation 5.6 asserts that the projection γ of a characteristic is indeed horizontal. Write $\dot\lambda$ for the derivative of the characteristic curve at $\lambda \in \mathcal{H}^\perp$. We call it a *characteristic direction* at λ. Equation 5.7 says that this direction is completely determined by λ, and by the derivative $\dot\gamma \in \mathcal{H}$ of the projected curve. Finally, equation 5.8 says that this projected derivative lies in the kernel of $w(\lambda)$, where $w(\lambda)$ is the dual curvature map $w : \mathcal{H}^\perp \to \bigwedge^2 \mathcal{H}^*$ evaluated at λ (see section 4.2). Relative to our coframe,

$$w(\lambda) = \sum \lambda_a c^a_{\mu\nu} \eta^\mu \eta^\nu |_{\mathcal{H}}.$$

To see this, recall that λ stands for the covector $\lambda_a \theta^a \in \mathcal{H}^\perp$. View λ as a one-form (a section of $\mathcal{H}^\perp$). Then

$$d\lambda \equiv \sum \lambda_a d\theta^a \bmod \theta \equiv \sum \lambda_a c^a_{\mu\nu} \eta^\mu \eta^\nu.$$

But, as we saw in section 4.2, $w(\lambda) = d\lambda|_{\mathcal{H}}$.

The preceding discussion proves this lemma:

Lemma 5.2.4 *Let* $\lambda \in \mathcal{H}^\perp_q$ *and let* $w_q : \mathcal{H}^\perp_q \to \bigwedge^2 \mathcal{H}^*_q$ *be the dual curvature at* q. *Then the set of all characteristic directions at* λ *projects linearly isomorphically to the subspace* $\ker(w_q(\lambda))$ *of* $\mathcal{H}_q$ *under the projection* $d\pi_\lambda : T_\lambda \mathcal{H}^\perp \to T_q Q$.

Example: Contact manifolds. A contact manifold has no characteristics. Indeed, any contact distribution $\mathcal{H}$ is defined, at least locally, by the vanishing of a *contact form,* a one-form with the property that $d\theta|_{\mathcal{H}}$ is symplectic. The rank of $\mathcal{H}^\perp$ is one. The typical element of $\mathcal{H}^\perp$ is then $\lambda\theta$, where $\lambda \in \mathbb{R}$ is the fiber coordinate. Consequently $w(\lambda) = \lambda d\theta|_{\mathcal{H}}$ and this dual curvature form at λ is also symplectic for $\lambda \neq 0$. Since, by definition, symplectic forms have no kernels, there are no characteristic directions, and hence no characteristics on a contact manifold.

Example: The Martinet distribution. A Martinet distribution is a type of degenerate contact distribution on a three-manifold (see section 3.2). The normal form is the distribution annihilated by $\theta = dz - \frac{1}{2}y^2 dx$. The contact condition degenerates along the surface $y = 0$, where the growth vector is $(2, 2, 3)$ instead of $(2, 3)$. Again $w(\lambda\theta) = \lambda d\theta \bmod \theta$. By dimensional considerations, $w(\lambda\theta)$ is either 0 or symplectic on $\mathcal{H}_q$. It is zero exactly along the Martinet surface $y = 0$. By Lemma 5.2.4, a characteristic can pass through q if and only if q is on the Martinet surface. On this surface the characteristic equations (5.6, 5.7, 5.8) become $dz = 0$ and $d\lambda = 0$, showing that the characteristics are the lines lying on the Martinet surface and parallel to the x-axis.

5.2.2 The derivative of the endpoint map

To prove Theorem 5.2.2 on the equality of singular curves and characteristics, we will need formulae for the differential of the endpoint map and its transpose.

Recall that we are using $u \in L^2(I, \mathbb{R}^k)$ as coordinates for the space Ω of horizontal curves passing through q_0. These coordinates are found by solving the initial value problem 5.1 for the time-dependent vector field $u(t) \cdot X$. Let $\Phi_t = \Phi_t(u)$ denote the flow of this time-dependent vector field, which maps an open subset of Q to an open subset. $\Phi_t(u)(q)$ depends smoothly on u and q, and continuously on t. (This is shown in Appendix E.) We suppose, without loss of generality, that the domain I of the paths in Ω is the unit interval $[0,1]$. We suppose that the curve γ issuing from q_0 corresponds to the particular choice of control $u \in L^2$. Then $\Phi_t(u)(q_0) = \gamma(t)$ and $\mathrm{end}(\gamma) = \Phi_1(u)(q_0)$. Thus

$$d(\mathrm{end})_\gamma(v) = \left.\frac{\partial \Phi_1(u + \epsilon v)(q_0)}{\partial \epsilon}\right|_{\epsilon=0}.$$

Proposition 5.2.5 *The derivative of the endpoint map is given by*

$$d(\mathrm{end})_\gamma(v) = d\Phi_1(q_0) \int_0^1 d\Phi_t(q_0)^{-1}(v \cdot X)(t)dt. \tag{5.9}$$

Proof. Write $\gamma_\epsilon(t) = \Phi_t(u + \epsilon v)(q_0)$ for the curves corresponding to $u + \epsilon v$. We then derive a linear differential equation for $\delta\gamma(t) := \partial\gamma_\epsilon/\partial\epsilon$, taken at $\epsilon = 0$. The curve γ_ϵ satisfies $\partial\gamma_\epsilon/\partial t = (u(t) + \epsilon v(t)) \cdot X(\gamma_\epsilon(t))$. As usual, $\partial/\partial\epsilon$ and $\partial/\partial t$ commute. Thus

$$\frac{d}{dt}\delta\gamma(t) = (v(t) \cdot X)|_{\gamma(t)} + u(t) \cdot \frac{\partial X}{\partial x}\delta\gamma(t) \tag{5.10}$$

(To define $\partial X/\partial x$ we must work in a fixed but arbitrary coordinate system. The choice of this coordinate system will wash out at the end of the computation.) Equation 5.10 is an inhomogeneous linear differential equation for $W(t) = \delta\gamma(t)$. We solve it by the method of variation of parameters, which we now recall. Suppose the vector quantity W satisfies the inhomogeneous equation $dW/dt = j(t) + A(t)W(t)$. Suppose also that we know the fundamental matrix solution $\Psi(t)$ of the associated homogeneous equation $d\Psi/dt = A(t)\Psi(t)$, $\Psi(0) = \mathrm{Id}$. Using the standard variation of parameters method, we make the guess $W(t) = \Phi(t)w(t)$ for the original inhomogeneous equation. From this guess we derive $\Psi(t)dw(t)/dt = j(t)$, an equation for $w(t)$ that can be solved by quadrature. The result is the solution

$$W(t) = \Psi(t)\left(\int_0^t \Psi(s)^{-1}j(s)ds\right)$$

to the original inhomogeneous equation with initial value $W(0) = 0$.

The associated homogeneous linear equation is

$$\frac{d}{dt}y(t) = u(t) \cdot \frac{\partial X}{\partial x}y(t). \tag{5.11}$$

We now show that $\Psi(t) = d\Phi_t(q_0)$ is the fundamental solution. Recall that Φ_t is defined by the differential equation $d\Phi_t(x)/dt = u(t) \cdot X(t, \Phi_t(x))$, with initial condition $\Phi_0(x) = x$. Set $x = q_0 + \epsilon y(0)$ and differentiate this differential equation with respect to ϵ at $\epsilon = 0$. Switch the order of the ϵ and t derivatives. The result is the differential equation

$$\frac{d}{dt} d\Phi_t(q_0)(y_0) = u(t) \cdot \frac{\partial X}{\partial x} d\Phi_t(q_0)(y_0).$$

Since this homogeneous differential equation is the same as equation 5.11 satisfied by Ψ_t, and since $d\Phi_0 = \mathrm{Id}$, we must have that $\Psi(t) = d\Phi_t(q_0)$. Putting these results together yields the formula for the differential.

QED

5.2.3 The transpose of the differential

Let $\lambda_1 \in T^*_{q_1}Q$ be a covector at the endpoint $q_1 = \gamma(1) = \mathrm{end}(\gamma)$ of the horizontal curve γ, and let $\lambda_0 = d\Phi_1(q_0)^*(\lambda)$. Then

$$\langle \lambda_1, d(\mathrm{end})_\gamma(v) \rangle = \left\langle \lambda_0, \int_0^1 d\Phi_t(q_0)^{-1}(v \cdot X)(t)dt \right\rangle$$

where $\langle \cdot, \cdot \rangle$ denotes the pairing between dual vectors and vectors. Writing

$$\lambda(t) := d\Phi_t(q_0)^{-1*}(\lambda_0),$$

we see that

$$\langle \lambda_1, d(\mathrm{end})_\gamma(v) \rangle = \int_0^1 \langle \lambda(t), (v \cdot X)(t) \rangle dt,$$

where the pairing inside the integral is between dual vector and vector at the point $\gamma(t)$ on the curve. Note that $\lambda(0) = \lambda_0$ and $\lambda(1) = \lambda_1$. A basic computation in mechanics, presented in Appendix A, shows that the curve $(\gamma(t), \lambda(t))$ is the solution curve for the Hamiltonian system with initial condition (q_0, λ_0) and with Hamiltonian H_u equal to the momentum function for the vector field $u \cdot X$. (See section A.3 for the definition of the momentum function of a vector field.) That is,

$$H_u(q, p, t) = \sum u^a(t) P_a(q, p), \quad P_a(q, p) = \langle p, X_a(q) \rangle := p(X_a(q)). \qquad (5.12)$$

There is a technical problem here not found in the usual mechanics. The u_a need not be smooth, rather only L^2. This lack of smoothness requires some reworking of existence-uniqueness theory for solutions to Hamilton's equations for time-dependent Hamiltonians that are only L^2 in the time variable t (see Appendix E).

Now

$$\langle \lambda(t), (v \cdot X)(t) \rangle = \sum v^a(t) \lambda(t)(X_a(q(t))) = \sum v^a(t) P_a(q(t), \lambda(t)).$$

This shows that

$$d(\text{end})^*_\gamma \lambda = (P_1(t), \ldots, P_k(t)), \tag{5.13}$$

where by a slight abuse of notation we have written $P_a(t)$ for the power functions $P_a(q,p)$ evaluated at the moving point $(q,p) = (\gamma(t), \lambda(t)) \in T^*Q$. Note that this expression defines a linear map $T^*_{q_1}Q \to L^2(I, \mathbb{R}^k) = L^2(I, \mathbb{R}^k)^*$, as it should.

Now suppose that the curve γ is singular. Then the image of $d(\text{end})_\gamma$ is a proper subspace of $T_{q_1}Q$. Consequently there is a *nonzero* covector $\lambda \in T^*_{q_1}Q$ that annihilates this subspace: $\lambda(d(\text{end})_\gamma(v)) = 0$ for all $v \in L_2(I, \mathbb{R}^k)$. Since v is arbitrary, we must have that $P_a(\gamma(t), \lambda(t)) = 0$, $a = 1, \ldots, k$. But the X_a frame $\mathcal{H}$, and $P_a(\gamma(t), \lambda(t)) = \lambda(t)(X_a(\gamma(t)))$ so that the vanishing of the P_a is equivalent to the assertion that $\lambda(t)$ annihilates $\mathcal{H}$.

This proves the following proposition.

Proposition 5.2.6 *Let the curve $\gamma \in \Omega(q_0, [0,1])$ correspond to the controls $u(t) \in L^2([0,1], \mathbb{R}^k)$ via the frame X for $\mathcal{H}$. Then γ is singular if and only if there is a solution $(q(t), \lambda(t))$ to the Hamiltonian system with time-dependent Hamiltonian $H_u = \sum u^a(t) p(X_a(q)$, with initial condition $q(0) = q_0$, $\lambda(0) = \lambda_0$, $\lambda_0 \in T^*_{q_0}Q$, $\lambda_0 \neq 0$, which in addition satisfies $\lambda(t) \in \mathcal{H}^\perp$ for all t.*

In the classical control theory literature the curves $(q(t), \lambda(t))$ that satisfy the conditions of this proposition are called *abnormal extremals* (see for example [Pontryagin et al. 1962]). The proposition asserts that the singular curves are precisely the projections of the abnormal extremals.

5.2.4 Proof that singular equals characteristic

We prove Theorem 5.2.2. We have just seen that γ is a singular curve if and only if there is a solution $\zeta(t) = (\gamma(t), \lambda(t))$, $\lambda(t) \neq 0$, to the Hamiltonian equations with Hamiltonian H_u such that ζ satisfies the additional constraint $\lambda(t) \in \mathcal{H}^\perp_{\gamma(t)}$. We now show that these solution curves are precisely the characteristics for $\mathcal{H}^\perp$. Recall that the k conditions $P_a = 0$, $a = 1, \ldots, k$, locally define $\mathcal{H}^\perp$. Also recall the definition of the Hamiltonian vector field X_H for a Hamiltonian H, namely $i_{X_H}\Omega = dH$ where Ω is the symplectic form on all of T^*Q (see section A.4). Apply this definition, and remember that we do *not* differentiate with respect to time, but only with respect to q and p, in forming dH. (Indeed, H_u need not be differentiable in t.) We find that

$$i_{\dot\zeta}\Omega = \sum u_a(t) dP_a|_{\zeta(t)}$$

characterizes the solutions to the Hamiltonian H_u for a given u. Now if Y is tangent to $\mathcal{H}^\perp$ then $dP_a(Y) = 0$, $a = 1, \ldots, k$. Consequently if ζ is such a Hamiltonian solution then it is a characteristic: $\Omega(\dot\zeta, Y) = 0$ for all Y tangent to $\mathcal{H}^\perp$. This logic reverses, showing that a characteristic ζ over the curve γ which is generated by controls u (i.e. has components u relative to our frame) is a solution to the Hamilton equation for H_u.

The L^2 condition on the derivative of λ follows from the linear relation 5.7 between singular directions $\dot{\gamma}$ and characteristic directions $\dot{\lambda}$, and the condition that $\dot{\gamma}$ be square integrable.

QED

5.3 Singularity and regularity

A regular curve is by definition a horizontal curve that is not singular. Thus geodesics fall into two classes: regular and singular. In section 1.5 we defined a *normal* geodesic to be one which "satisfies the geodesic equations", meaning that it is the projection of a solution $\zeta \subset T^*Q$ to Hamilton's equations for the normal Hamiltonian H. We proved that any sufficiently short arc of such a solution curve is a minimizing geodesic (Theorem 1.5.7). On the other hand, in chapter 3 we gave an example of a singular minimizing geodesic that is not normal. What are the relations between singular curves and minimizing geodesics? We will prove that every regular minimizing geodesic is normal. But first let us let us look at the dichotomy between regular and singular from the point of view of the method of Lagrange multipliers.

5.3.1 Lagrange multipliers

In looking for minimizing geodesics joining q_0 to q_1 we are solving this constrained minimization problem: Minimize

$$E(\gamma) = \frac{1}{2}\int_I \|\dot{\gamma}\|^2 dt$$

subject to the constraint

$$F(\gamma) = q,$$

where F is the endpoint map based at q_0 (see section 1.3). According to the method of Lagrange multipliers, we are to form the associated functional

$$\lambda_0 E(\gamma) + \langle \lambda_1, F(\gamma)\rangle,$$

where the parameters $(\lambda_0, \lambda_1) \neq (0,0)$ are the Lagrange multipliers. Any minimizer γ must be a critical point for this functional, for some nonzero choice of the parameters. The parameter λ_0 is real, but λ_1 runs over the space dual to the range of F if that range is a linear space. In our nonlinear situation, this domain is the manifold Q and we have $\lambda_1 \in T^*_{q_1}Q$.

Often we only teach our calculus students this rule in the case $\lambda_0 \neq 0$, for which we may as well take $\lambda_0 = 1$, by scaling. This restriction is equivalent to the assumption that γ is a regular point for F, which means that γ is a regular curve. Minimizers that arise with $\lambda_0 = 0$ correspond to singular minimizers. To see this, note that when $\lambda_0 = 0$, the Lagrange multiplier test reduces to $\langle \lambda_1, dF\rangle = 0$, i.e. $\lambda_1 \neq 0$ annihilates the image of the differential of F. In this case the function E to be minimized has effectively disappeared.

Remark. Put properly into the control theoretic context, the method of Lagrange multipliers becomes the maximum principle of Pontryagin et al. [1962]. This principle has been frequently, and often incorrectly, applied to the problem of finding subriemannian geodesics. (See the notes in section 4.7.)

5.3.2 Regular implies normal

Theorem 5.3.1 *Every regular minimizing geodesic is normal.*

Proof. Suppose that γ is a regular minimizing geodesic joining q_0 to q_1. Write Ω for the space of horizontal paths starting at q_0. Write $\Omega(q_0, q_1) \subset \Omega$ for the space of horizontal curves joining q_0 to q_1 that have square-integrable derivatives. We have $\Omega(q_0, q_1) = \text{end}^{-1}(q_1) \subset \Omega$, a submanifold of Ω whose tangent space at γ is the kernel of $d(\text{end})_\gamma$. We have that γ minimizes the energy function E restricted to $\Omega(q_0, q_1)$. Consequently $dE = 0$ upon restriction to this kernel. For any continuous linear operator $A : V_1 \to V_2$, the annihilator of the kernel of A equals the image of the adjoint $A^* : V_2^* \to V_1^*$. Applying this fact to $d(\text{end})_\gamma$, we have that $dE(\gamma) \in \text{im}(d(\text{end})_\gamma^*)$, which is to say that there exists a $\lambda \in T_{q_1}^* Q$ such that

$$dE - d(\text{end})_\gamma^*(\lambda) = 0. \tag{5.14}$$

This is just the standard argument justifying the method of Lagrange multipliers.

We now recall our coordinate formula 5.13 for the adjoint. Let u be the controls (components) of our curve γ relative to the orthonormal horizontal frame X that determines our coordinates, and form the Hamiltonian H_u as in equation 5.12. Let $(q(t), p(t)) \in T_\gamma^* Q$ denote the solution to the corresponding Hamilton's equations in T^*Q with terminal point (q_1, λ). Then $d(\text{end})_\gamma^*(\lambda) = P(\gamma(t), p(t)) \in L^2(I, \mathbb{R}^k)$ where $P = (P_1, \ldots, P_k)$ is the vector of momentum functions corresponding to the horizontal frame $X = (X_1, \ldots, X_k)$. Consequently, if $v \in L^2$ denotes any variation of u, and hence of the path γ, then

$$\left(d(\text{end})_\gamma^*(\lambda)\right)(v) = \int P(\gamma(t), p(t)) \cdot v(t) dt.$$

Now $E = \int \frac{1}{2} \sum u_a(t)^2 dt$, so that $dE(\gamma)(v) = \int u \cdot v dt$. It follows that

$$dE - d(\text{end})_\gamma^* \lambda = \int (u(t) - P(q(t), p(t)) \cdot v(t) dt.$$

Since the left-hand side must be zero for all v, we have

$$u_a(t) = P_a(q(t), p(t)) \text{ a.e.}, \quad a = 1, 2, \ldots, k. \tag{5.15}$$

Recall that the vector field defining Hamilton's equations for any time-dependent function $f : T^*Q \times I \to \mathbb{R}$ is obtained by using the symplectic form to invert the one-form $d_{q,p} f$, where the differential $d_{q,p}$ means differentiate only with respect to the T^*Q-variables (see Appendix E). We have $f(q, p, t) = H_u =$