DOI: https://doi.org/10.1070/RM9744

Topics in sub-Riemannian geometry

A. A. Agrachev

Abstract. Sub-Riemannian geometry is the geometry of spaces with nonholonomic constraints. This paper presents an informal survey of some topics in this area, starting with the construction of geodesic curves and ending with a recent definition of curvature.

Bibliography: 28 titles.

Keywords: geodesic, curvature, sub-Riemannian geometry.

Contents

Introduction

This survey is concerned with the recent development of the old idea of the distance between two points as the length of the shortest path in the class of all possible paths connecting the points. Here the adjective 'possible' is not accidental: the indicated development has mainly to do with that.

In Euclidean geometry shortest (length-minimizing) paths are straight-line intervals, satisfying the usual axioms. In the Riemannian world, Euclidean geometry is just one of countless possibilities. Nevertheless, any Riemannian metric can be well approximated by a Euclidean metric for very small distances: when looked at through a more and more powerful microscope, a neighbourhood of any point becomes less and less distinguishable from a Euclidean neighbourhood. In the limit, instead of the original space we obtain the space of initial velocities of paths issuing from the given point, and this space is Euclidean.

Riemann's construction was based on previous work by Gauss, who had investigated surfaces in 3-dimensional Euclidean space. The distance between two points on a surface is equal to the length of the shortest path on the surface that connects these points. The initial velocities of smooth curves on the surface that issue from a given point form the tangent plane to the surface, which is a Euclidean plane.

AMS 2010 *Mathematics Subject Classification.* Primary 53C17.

The planes tangent to the surface at two distinct points are isometric, but in the general case small neighbourhoods of these points on the surface are non-isometric. They are certainly non-isometric if the Gaussian curvature takes distinct values at these two points.

Riemann generalized Gauss' construction to higher dimensions and explained that everything could be done intrinsically, and embedding in a Euclidean space was not needed. In fact, to measure the lengths of curves, it is sufficient to know the Euclidean lengths of their velocities. A Riemannian manifold is a smooth manifold such that its tangent space at each point carries its own Euclidean structure, and these structures depend smoothly on the point.

For an inhabitant of a Riemannian space residing at some point, tangent vectors are directions in which he can move or send information or from which he can receive information. He can measure the lengths of vectors and the angles between vectors tangent to the same point in accordance with the Euclidean rules, and in general, this is all that he can do. Still, in principle, this inhabitant can recover the geometry of the space by making these simple measurements along different smooth curves.

In a sub-Riemannian space we can neither move nor send information in all the directions, nor can we receive information from everywhere. There are constraints (imposed by God, by a moral imperative, by a government, or just by the laws of Nature). A sub-Riemannian space is a smooth manifold such that some admissible subspace, endowed with a Euclidean structure, is selected in the tangent space to each point, and the admissible subspace and the Euclidean structure in it depend smoothly on the point.

Admissible paths are paths with admissible velocities. The distance between two points is the greatest lower bound of the lengths of paths connecting these points. We will assume that any two points on a connected component of the manifold can be connected by an admissible path. At first sight this condition may look odd and difficult to satisfy, but in actual fact it is not. The crucial thing is that the admissible subspaces vary from point to point, and our condition holds for a more-or-less general dependence of the subspace on the point. More precisely, it fails only for a very special choice of the admissible subspaces.

Let k be the dimension of the admissible subspaces and assume that the manifold is connected and has dimension $n > k$. As inhabitants of such a sub-Riemannian world, we move and send and receive information in accordance with the rules of the k-dimensional Euclidean space, but notwithstanding, we can in principle reach each point in the n-dimensional manifold. Some coordinates in the sub-Riemannian space are 'clandestine' and poorly discernible at short distances. This is a remarkable, but not yet fully exploited source for occult speculations based on mysterious hidden dimensions, and also for theoretical physicists, who are perpetually searching for new wild formalisms!

In mechanics this is the natural geometry of systems with non-holonomic constraints: skates, wheels, rolling balls, bearings, and so on. This geometry could also be useful in models for social behaviour within a restrictive bureaucratic and legal system, by showing how the degree of freedom can be increased without violating the existing rules.

All this relates to certain natural applications, and the reader himself can indicate many others. On the other hand, a mathematician's business is to develop the geometry itself, for otherwise we will have nothing in our hands to apply. Currently this is an extensive and rapidly developing area, though it has only been in recent years that mathematicians have appeared who specialize in sub-Riemannian geometry. Authors from the most diverse areas have left their mark here: hyperbolic, conformal, and CR-geometry, hypoelliptic operators, non-commutative harmonic analysis, variational calculus and geometric measure theory, optimal control, and non-holonomic mechanics. This paper does not pretend to be an exhaustive survey of the basics of sub-Riemannian geometry: we have only touched on topics which are close to this author.

I have tried to write in an informal narrative manner, hopefully without infringing on accuracy. References are deferred to the comments after each section, in order to keep the presentation fluent. The aim of this survey is to expose this beautiful and fascinating field to as wide a circle of mathematicians as possible and perhaps to attract new researchers or at least bring some intellectual pleasure to the reader.

1. Geodesic curves

Let $\Delta \subset TM$ be a smooth vector distribution on a manifold M, and let $\overline{\Delta}$ be the space of smooth vector fields on M taking values in Δ. Vector fields in $\overline{\Delta}$ are also said to be *horizontal.* Let $\Delta_q = \Delta \cap T_qM$ for $q \in M$. A smooth curve $\gamma\colon [t_0, t_1] \to M$ is called a *horizontal* or *admissible* curve if $\dot\gamma(t) \in \Delta_{\gamma(t)}$ for $t_0 \leqslant t \leqslant t_1$. It is also natural to regard a concatenation of several admissible curves as an admissible curve. (Generally speaking, it is not smooth, but only piecewise smooth.) An even more flexible and convenient class of admissible curves comprises the Lipschitz admissible curves, which are Lipschitz curves $\gamma\colon [t_0, t_1] \to M$ such that $\dot\gamma(t) \in \Delta_{\gamma(t)}$ for almost all $t \in [t_0, t_1]$. These curves are solutions of non-autonomous differential equations of the form $\dot q = V_t(q)$, where $V_t \in \overline{\Delta}$, $t_0 \leqslant t \leqslant t_1$, and the map $(q, t) \mapsto V_t(q)$ is measurable and locally bounded.

We define an equivalence relation on M by saying that two points are equivalent if they can be connected by an admissible curve. Sussmann's remarkable 'orbit theorem' says that equivalence classes are immersed submanifolds of M. This result, which is not very hard to prove but is very important, provides a kind of description of the tangent spaces to these submanifolds, and we now present it.

Recall that a vector field V on M is said to be complete if solutions of the differential equation $\dot q = V(q)$ are defined on the whole real line $\mathbb{R}$, and that all fields with compact support are complete. Let V be a complete field and let the map $P^t\colon q(0) \mapsto q(t)$ shift each point in M by the time t along the solution of the equation $\dot q = V(q)$ going through this point. Then $P^t\colon M \to M$ is a diffeomorphism, and

$$P^{t+s} = P^t \circ P^s \qquad \forall\, t, s \in \mathbb{R}.$$

The one-parameter diffeomorphism group P^t, $t \in \mathbb{R}$, is the *flow generated by the field* V. It can be conveniently denoted as an exponential: $P^t \doteq e^{tV}$. The flow generated by a horizontal vector field is said to be horizontal.

Let us now consider the subgroup generated by the horizontal flows in the diffeomorphism group of M, that is:

$$\mathscr{P} \doteq \{e^{t_1 f_1} \circ \cdots \circ e^{t_k f_k}(\,\cdot\,) : f_i \in \overline{\Delta},\ t_i \in \mathbb{R},\ i = 1, \dots, k,\ k = 1, 2, \dots\}.$$

The orbits of the action of $\mathscr{P}$ on M clearly lie in our equivalence classes. In fact, the orbits coincide with equivalence classes, and the tangent space to the orbit through a point $q \in M$ has the form

$$T_q\mathscr{P}(q) = \operatorname{span}\{(P_* V)(q) \colon P \in \mathscr{P},\ V \in \overline{\Delta}\}.$$

Indeed, let $P_{1*}V_1, \dots, P_{k*}V_k$ be a basis of the space on the right-hand side of this equality. Then the germ of the k-dimensional submanifold

$$\{e^{s_1 P_{1*}V_1} \circ \cdots \circ e^{s_k P_{k*}V_k} \colon \text{ the } s_i \text{ are close to zero}\} \tag{1.1}$$

lies in the orbit $\mathscr{P}(q)$, and in fact, $e^{sP_*V} = P \circ e^{sV} \circ P^{-1}$.

Moreover, the germs of the form (1.1) at all the points $q \in M$ serve as a basis of a certain topology on M which is in general stronger than the original topology. The connected components of M with this strong topology are immersed submanifolds, essentially by definition. In fact, the ‘neighbourhoods’ (1.1) are coordinate charts on these submanifolds. Horizontal curves are continuous in the strong topology, so the connected components coincide with our equivalence classes. This is just Sussmann’s orbit theorem.

Since horizontal vector fields are tangent to orbits at all points, the commutators of these fields are also tangent. In the final analysis we get that

$$\operatorname{Lie}_q \Delta \doteq \operatorname{span}\{[V_1, [V_2, [\cdots, V_k]] \cdots](q) \colon V_i \in \overline{\Delta},\ i = 1, \dots, k,\ k = 1, 2, \dots\}$$

lies in $T_q\mathscr{P}(q)$. In particular, if $\operatorname{Lie}_q \Delta = T_q M$, then the orbit containing q is an open subspace of M (in the original ‘weak’ topology). Throughout what follows we assume that

$$\operatorname{Lie}_q \Delta = T_q M \qquad \forall\, q \in M. \tag{1.2}$$

Experts in different fields have different names for distributions satisfying (1.2): ‘bracket-generating’, ‘totally non-holonomic’, or ‘satisfying Hörmander’s condition’.

In view of the foregoing, the condition (1.2) ensures that the action of the group $\mathscr{P}$ is transitive on M. This is called the Rashevskii–Chow theorem.

Remark 1. If both M and the distribution Δ are real-analytic, then we have not only the inclusion $\operatorname{Lie}_q \Delta \subset T_q\mathscr{P}(q)$ but even the equality $\operatorname{Lie}_q \Delta = T_q\mathscr{P}(q)$. This can be seen, for example, from the analyticity of solutions of ordinary differential equations with analytic right-hand side. In this case, the condition (1.2) is not only sufficient but also necessary for $\mathscr{P}$ to act on M transitively.

Definition 1. A smooth vector subbundle $\Delta \subset TM$ endowed with a Euclidean structure and satisfying (1.2) is called a *sub-Riemannian structure.*

The scalar product of two vectors $v_1, v_2 \in \Delta_q$ will be denoted by $\langle v_1|v_2\rangle$, and the length of a vector $v \in \Delta_q$ by $|v| = \langle v|v\rangle^{1/2}$. The length of a horizontal curve $\gamma\colon [t_0, t_1] \to M$ is given by

$$\operatorname{length}(\gamma) = \int_{t_0}^{t_1} |\dot{\gamma}(t)|\, dt.$$

It is clear that the length of a curve is unchanged by a reparametrization.

The sub-Riemannian distance between two points $q_0, q_1 \in M$ (also called the Carnot–Carathéodory distance, following the trend started by Gromov) is defined as follows:

$$\delta(q_0, q_1) = \inf\{\operatorname{length}(\gamma)\colon \gamma\colon [0,1] \to M \text{ is horizontal},\ \gamma(0) = q_0,\ \gamma(1) = q_1\}.$$

If $\Delta = TM$, then this is the usual Riemannian distance, but if $\Delta \subsetneqq T_qM$, then it has some quite unusual properties. In any case, a sub-Riemannian metric, like a Riemannian metric, induces the standard topology on M. In fact, analysing (1.2), we easily see that points close to q_0 can be joined to q_0 by short horizontal paths.

Hence small sub-Riemannian balls are compact. Moreover, as in the Riemannian case, a sub-Riemannian metric space is complete if and only if all the balls are compact. Furthermore, in each compact ball its centre can be connected with each point in the ball by a length-minimizing horizontal path. In other words, if the ball with radius r and centre q_0 is compact and q_1 lies at a distance at most r from q_0, then we can replace inf by min in the definition of $\delta(q_0, q_1)$.

We see that it is important to know how to describe shortest paths. As in the Riemannian case, it is slightly easier to describe geodesic curves, that is, horizontal curves $\gamma\colon [0,1] \to M$ such that $\gamma\big|_{[t,s]}$ is a length-minimizing path between $\gamma(t)$ and $\gamma(s)$ for any sufficiently close t and s, although such a γ is not necessarily length minimizing from $\gamma(0)$ to $\gamma(1)$.

Recall that in the Riemannian case a geodesic is characterized by the initial point and the initial velocity. On the other hand, if $\Delta \subsetneqq T_qM$, then geodesic curves issuing from q cannot be characterized by their initial velocities, as is at once clear. Indeed, we know that the length-minimizing paths from q fill a whole neighbourhood of q, while their initial velocities lie in Δ_q. We just do not have enough initial velocities! To save the situation, we can pass from velocities to momenta, that is, from the tangent bundle to the cotangent bundle.

Let $p \in T_q^*M$ and let $h_q(p) = \max\{\langle p, v\rangle\colon v \in \Delta_q,\ |v| \leqslant 1\}$ be the norm on $\Delta^* = T_q^*M/\Delta_q$ dual to the Euclidean norm on Δ_q. Varying now not only p but also q, we obtain a non-negative function $h\colon T^*M \to \mathbb{R}$. Furthermore, $h^{-1}(0) = \Delta^{\perp}$ is the orthogonal complement of Δ, and the restriction of h to $T^*M \backslash \Delta^{\perp}$ is a smooth function.

It is easy to see that $h^2\big|_{T_q^*M}$ is a non-negative quadratic form. It is important to note that in the general case this form is degenerate, and $\ker h^2\big|_{T_q^*M} = \Delta_q^{\perp}$. Thus the level sets $h^{-1}(c) \cap T_q^*M$, where $c > 0$, are homothetic elliptic cylinders with generating subspace $\Delta_q^{\perp}$.

Recall that T^*M carries a canonical symplectic structure. Let $\pi\colon T^*M \to M$ be the standard projection: $\pi(T_q^*M) = q$. First we define the tautological differential

1-form s on T^*M: $s_\lambda = \lambda \circ \pi_* \in T^*_\lambda(T^*M)$, $\lambda \in T^*M$, and then we define the symplectic structure $\sigma = ds$. In the local coordinates (p, q) on T^*M, where $q = (q^1, \dots, q^n)$ and $\lambda = p_1\,dq^1 + \dots + p_n\,dq^n$, the tautological form can be expressed by $s = p_1\,dq^1 + \dots + p_n\,dq^n$, in accordance with its name, and the symplectic form is $\sigma = dp_1 \wedge dq^1 + \dots + dp_n \wedge dq^n$.

Recall also that a characteristic (or a characteristic curve) of a differential form is a curve whose velocity at each point lies in the kernel of the form. The symplectic form σ is non-degenerate and thus has no characteristics, but in general the restriction of σ to a level set of h has characteristics. Characteristics of $\sigma\big|_{h^{-1}(c)}$ are called *sub-Riemannian extremals.* Here we distinguish between *normal extremals* $(c > 0)$ and *abnormal extremals* $(c = 0)$. These extremals are curves in T^*M.

Theorem 1. *Each geodesic curve is the projection on M of some extremal.*

A geodesic is said to be normal if it is the projection of a normal extremal, and it is abnormal if it is the projection of an abnormal extremal. Generally speaking, a geodesic can be normal and abnormal at the same time.

Let us now look more closely at extremals, starting with normal ones. For $r > 0$ the dilation $\lambda \mapsto r\lambda$, $\lambda \in T^*_qM$, $q \in M$, takes characteristics of $\sigma\big|_{h^{-1}(c)}$ to characteristics of $\sigma\big|_{h^{-1}(rc)}$, and hence we can confine ourselves to normal extremals lying in $h^{-1}(1)$. Note that $h^{-1}(1)$ is a codimension-one submanifold of T^*M, and its tangent space at λ is the kernel of the linear form $d_\lambda h$. Thus, $\ker \sigma\big|_{h^{-1}(1)}$ is the skew-orthogonal complement of $\ker d_\lambda h$. This is the straight line generated by $\vec{h}(\lambda)$, where $\sigma(\,\cdot\,, \vec{h}(\lambda)) = dh$.

The vector field $\vec{h}$ on T^*M is just the Hamiltonian field with Hamiltonian $h\colon T^*M \to \mathbb{R}$. Thus, normal extremals are trajectories of the Hamiltonian system $\dot{\lambda} = \vec{h}(\lambda)$. The theorem stated above says that any geodesic is the projection of some extremal. For normal extremals we also have the converse result.

Proposition 1. *The projection of any trajectory of the Hamiltonian system $\dot{\lambda} = \vec{h}(\lambda)$ is a geodesic curve.*

The codimension of $h^{-1}(0) = \Delta^\perp$ in T^*M is greater than 1, and the rank of $\sigma\big|_{h^{-1}(0)}$ can vary from point to point, so abnormal extremals are more difficult to describe. Note that, in accordance with their definition, abnormal extremals depend only on the distribution Δ, but not on the Euclidean structure on it. In addition, their projections are not necessarily geodesic curves.

It is time to give concrete examples of sub-Riemannian structures and their geodesics. The simplest class of sub-Riemannian spaces comes from planar isoperimetric problems when they are properly interpreted.

Let ω be a smooth differential 1-form on $\mathbb{R}^2$:

$$\omega_x = a_1(x)\,dx^1 + a_2(x)\,dx^2, \qquad x = (x^1, x^2) \in \mathbb{R}^2. \tag{1.3}$$

We consider the problem of minimizing the lengths of curves $\gamma\colon [0,1] \to \mathbb{R}^2$ that connect two fixed points in the plane and satisfy the additional condition $\int_\gamma \omega = c$, where c is a fixed constant. This problem is related in a natural way to a certain special sub-Riemannian structure in

$$\mathbb{R}^3 = \{(x, y) \colon x \in \mathbb{R}^2,\ y \in \mathbb{R}\}.$$

Let $q = (x, y) \in \mathbb{R}^3$, and let $\Delta_q = \ker(dy - \omega_x)$. The 2-dimensional subspaces Δ_q, $q \in \mathbb{R}^3$, form a distribution Δ in $\mathbb{R}^3$. The restriction to Δ_q of the projection $(x, y) \mapsto x$ is an invertible map, so the standard Euclidean structure $(dx^1)^2 + (dx^2)^2$ in $\mathbb{R}^2$ induces a sub-Riemannian structure in $\mathbb{R}^3$ with Δ as the distribution.

A curve $t \mapsto (x(t), y(t))$ is horizontal if and only if $\dot{y}(t) = \langle \omega_{x(t)}, \dot{x}(t) \rangle$. That is, a horizontal curve has the form

$$t \mapsto \biggl(\gamma(t), y_0 + \int_{\gamma|_{[0,t]}} \omega\biggr),$$

where γ is an arbitrary Lipschitz curve in the plane. Furthermore, the sub-Riemannian length of the horizontal curve is equal to the length of γ. We see that the original isoperimetric problem is equivalent to the problem of minimizing the sub-Riemannian lengths of horizontal curves connecting two fixed points in $\mathbb{R}^3$.

We can describe the geodesics of this sub-Riemannian structure. Of course, it is sufficient to describe their projections on $\mathbb{R}^2$. Note first of all that for two forms ω whose difference is a closed form the result will be the same. In fact, the difference

$$\int_\gamma (\omega + d\varphi) - \int_\gamma \omega = \varphi(\gamma(1)) - \varphi(\gamma(0))$$

depends only on the endpoints of γ.

We see that to describe the projections of geodesic curves on the plane it is sufficient to know the form

$$d\omega = b(x)\, dx^1 \wedge dx^2, \quad \text{where} \quad b = \frac{\partial a_2}{\partial x^1} - \frac{\partial a_1}{\partial x^2}\,.$$

In fact, for this problem the condition of total non-holonomicity (1.2) is equivalent to the following: at each point $x \in \mathbb{R}^2$ at least one partial derivative $\dfrac{\partial^{i+j} b}{(\partial x^1)^i (\partial x^2)^j}$ is non-zero.

Proposition 2. *A curve $\gamma\colon [0,1] \to \mathbb{R}^2$ is the projection of a normal geodesic on the plane if and only if for all $t \in [0,1]$ and some constant $\nu \in \mathbb{R}$ the curvature of γ at $\gamma(t)$ is equal to $\nu b(\gamma(t))$.*

The curve γ is the projection of an abnormal extremal if and only if $b(\gamma(t)) = 0$ for $0 \leqslant t \leqslant 1$. If in addition $d_{\gamma(t)} b \neq 0$ for $0 \leqslant t \leqslant 1$, then γ is the projection of an abnormal geodesic.

We can consider a slightly more general problem by replacing $\mathbb{R}^2$ by some Riemannian 2-manifold N. Then Proposition 2 remains valid after the replacement of the curvature of the plane curve by the geodesic curvature of the curve in N.

We can generalize the problem further by taking a principal 1-dimensional bundle over N (with structure group $\mathbb{R}$ or S^1) instead of the Cartesian product $N \times \mathbb{R}$.

Then the role of ω will be played by a connection on the principal bundle, and horizontal curves will be parallel translations along curves on the base. Proposition 2 will still be valid with b the curvature of the connection.

We see that the projections of normal geodesics on N are trajectories of point charges in the magnetic field b, and the constant ν plays the role of the charge of a particle. It is important to note that the abnormal geodesics described in Proposition 2 are independent of the choice of a Riemannian metric on the surface!

In the next example we will again minimize the length of a curve connecting two points in the plane, but under different additional conditions. Consider a ball which rolls along the plane without slipping or spinning. The instantaneous state of this system is given by the point of tangency of the ball and the plane and by the orientation of the ball. Thus, the state space M is $\mathbb{R}^2 \times \mathrm{SO}(3)$. The no-slipping condition means that the point of tangency has the same velocity vectors on the plane and on the sphere, and the no-spinning condition means that at each moment of time the angular velocity vector of the rotation of the ball is parallel to the plane on which the ball rolls. These conditions define a 2-dimensional distribution $\Delta \subset TM$ of admissible velocities.

It is easy to see that the ball with fixed initial orientation can roll in a unique admissible way along an arbitrary plane curve. In other words, there is a natural bijection between admissible (horizontal) curves starting at a fixed point of the state space M and plane curves starting at the projection of this point on $\mathbb{R}^2$. That is, the terminal orientation of the ball is determined by its initial orientation and the path in $\mathbb{R}^2$ along which it rolls. The problem is to find the length-minimizing path in the class of plane paths with given endpoints such that after rolling along these paths, the ball takes the prescribed orientation.

In this problem, projections of normal geodesics on $\mathbb{R}^2$ are *Euler elasticas*, very well-known curves. Euler obtained them as equilibrium states of an elastic rod, and in the problem of a ball rolling on the plane they were discovered by Jurdjevic. Geometrically, elasticas can be characterized as smooth curves in the Euclidean plane whose curvature is an affine function of the coordinates. In other words, $\gamma\colon [0,1] \to \mathbb{R}^2$ is an Euler elastica if and only if there exist $\overline{a} \in \mathbb{R}^2$ and $a \in \mathbb{R}$ such that the curvature of γ at a point $\gamma(t)$ is $\langle \overline{a} | \gamma(t) \rangle + a$ for all $t \in [0,1]$. We note that elasticas with this characterization are a particular case of the curves we encountered above in our consideration of the isoperimetric problem. It turns out that normal geodesics correspond to a ball rolling along trajectories of charged particles in an affine magnetic field.

Elasticas can also be characterized in a more traditional way. To do this we identify $\mathbb{R}^2$ with the complex plane $\mathbb{C}$. Let $\gamma\colon [0,t_1]$ be a curve parametrized by arc length. Then $\dot\gamma(t) = e^{i\theta(t)}$, $\theta \in \mathbb{R}$. The curve γ is an elastica if and only if $\ddot\theta = a\sin(b\theta + c)$, $0 \leqslant t \leqslant t_1$, for some constants a, b, and c. In other words, elasticas parametrized by arc length are curves whose velocities have directions satisfying the equation of motion of a mathematical pendulum. Figure 1, which gives various types of elastica, was taken from Euler's book [16]. Elasticas with points of inflection correspond to oscillatory motion of the pendulum (when the pendulum changes its direction of motion), and elasticas without points of inflection correspond to rotational motion.

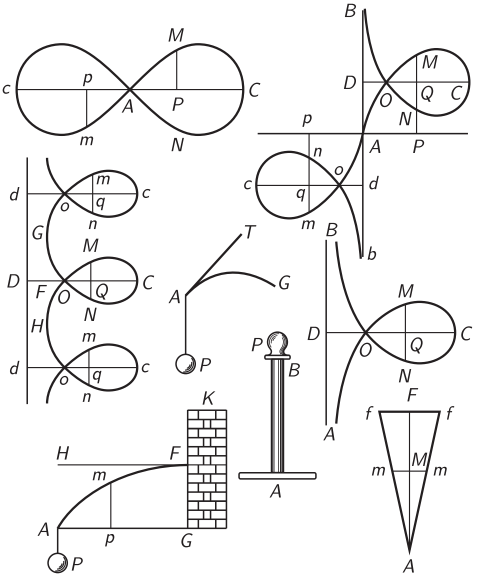

Figure 1

In this problem abnormal geodesics are not very interesting: they correspond to the ball rolling simply along straight lines. Thus, abnormal geodesics are at the same time normal geodesics, the most degenerate ones.

The last example in this section is a ball rolling on the plane without slipping, but with possible spinning. The state space is the same as in the previous example, but the horizontal distribution is 3-dimensional. As before, we can roll the ball along an arbitrary plane curve, but the angular velocity of the rotation can be arbitrary. By definition, the length of the velocity vector of a horizontal curve is equal to the length of the angular velocity of the rotation.

In this problem we can no longer recover horizontal curves in M from their initial points and their projections on $\mathbb{R}^2$, in any case if we are talking about non-parametrized curves. But if we confine ourselves to horizontal curves parametrized by arc length, then under a certain reasonable additional condition we can recover them from the initial point and the parametrized projection on $\mathbb{R}^2$. In fact, the projection on $\mathbb{R}^2$ is a curve $\gamma\colon [0,1] \to \mathbb{R}^2$ such that $|\dot{\gamma}(t)| \leqslant 1$. The non-slipping condition means that the length of the orthogonal projection of the angular velocity of the rotation at time t on the horizontal plane is equal to $|\dot{\gamma}(t)|$. Since we have assumed that this angular velocity has length 1, we can uniquely recover the length of the vertical component of the angular velocity. And if we have fixed the direction of rotation about the vertical axis, then we can uniquely recover the whole angular velocity.

In summary: we can roll the ball without slipping and with unit angular velocity in a unique way along each curve $\gamma\colon [0,t_1] \to \mathbb{R}^2$ with $|\dot{\gamma}(t)| \leqslant 1$, provided that the ball can spin about the vertical axis only clockwise (or anticlockwise). It remains to describe the geodesics in this problem, which are very simple.

A curve $\gamma\colon [0,t_1] \to \mathbb{R}^2$ is the projection of a geodesic curve parametrized by arc length if and only if by using a rotation and a translation of the plane it can be transformed into a curve of the form

$$\overline{\gamma}(t) = \bigl(at, b\sin(ct+d)\bigr), \qquad 0 \leqslant t \leqslant t_1,$$

where $a^2 + (bc)^2 \leqslant 1$.

The ball rolls along a sinusoid, or (if $a = 0$) oscillates along a straight-line interval while simultaneously spinning about the vertical axis. The spinning compensates for the variable rolling velocity, but the direction of spin always remains the same. In this problem the abnormal geodesics are the same as in the preceding problem: rolling without spinning along a straight line.

Comments. Sussmann's orbit theorem was proved in [27], but its corollary, the Rashevskii–Chow theorem, appeared much earlier (see [23] and [15]). The analytic version of the orbit theorem is due to Nagano [21]. Theorem 1 is an easy consequence of Pontryagin's maximum principle [22], and details of its proof together with the proofs of Propositions 1 and 2 can be found in [3], for instance. The book [20] treats the same questions from a slightly different point of view. The fact that geodesics in the problem of a ball rolling on the plane without spinning are Euler elasticas was discovered by Jurdjevic (see [18]), and further information can be found in [26]. Figure 1 with elasticas was taken from Euler's book [16]. Geodesics in the problem of rolling with spinning were investigated in [10].

2. Balls

Now we turn away from Euclidean balls rolling on the plane and try to figure out what balls look like in the most general sub-Riemannian metric space. First of all we are interested in balls with small radius, which are quite different from small Riemannian balls. Recall that the ball (sphere) with radius $r > 0$ and centre $q \in M$ is the set $B_q(r) = \{x \in M\colon \delta(q,x) \leqslant r\}$ (the set $S_q(r) = \{x \in M\colon \delta(q,x) = r\}$), where δ is the sub-Riemannian distance. Since the sub-Riemannian metric induces the standard topology in M, we have $S_q(r) = \partial B_q(r)$ as in the Riemannian case, but then we encounter significant differences.

Following Gromov, we look at a small ball 'through a microscope'. Namely, we look at the ball $B_q(\varepsilon r)$ and multiply all distances in it by the factor $1/\varepsilon$, after which we let ε go to zero. In other words, we look at the family of metric spaces $\bigl(B_q(\varepsilon r), \delta/\varepsilon\bigr)$ and its Gromov–Hausdorff limit as $\varepsilon \to 0$. By definition, this limit (which always exists in the case of a sub-Riemannian space) is the ball of radius r in the metric tangent space of (M,δ) at a point q. We denote this ball, together with the metric in it, by $(\widehat{B}_q(r), \widehat{\delta}\,)$. Thus,

$$(\widehat{B}_q(r), \widehat{\delta}\,) = \lim_{\varepsilon\to 0}\biggl(B_q(\varepsilon r), \frac{1}{\varepsilon}\delta\biggr).$$

Clearly, $(\widehat{B}_q(r_1), \widehat{\delta}\,) \subset (\widehat{B}_q(r_2), \widehat{\delta}\,)$ for $r_1 < r_2$. The union $\widehat{M}_q = \bigcup_{r>0} \widehat{B}_q(r)$ equipped with the metric $\widehat{\delta}$ is the tangent space of (M,δ) at q.

In the Riemannian case the tangent space T_qM of M carries a Euclidean structure. The Euclidean space T_qM coincides with the metric tangent space $(\widehat{M}_q, \widehat{\delta}\,)$, as is obvious. To describe the metric tangent space in the general sub-Riemannian case we must start by defining the *flag of the distribution* Δ at q.

The flag of the distribution is the sequence of subspaces

$$\Delta_q = \Delta_q^1 \subset \dots \subset \Delta_q^m = T_qM$$

which are defined as follows:

$$\Delta_q^k = \operatorname{span}\{[V_1, [V_2, \dots, V_l] \dots](q) \colon V_i \in \overline{\Delta},\ i = 1, \dots, l,\ l \leqslant k\}, \qquad 0 \leqslant k \leqslant m.$$

The dimension of Δ_q^k can depend on q, but it is clear that $\dim \Delta_q^k$ is lower semicontinuous as a function of q. The distribution Δ is said to be *equiregular* at q if the quantities $\dim \Delta_x^k$, $k = 1, \dots, m$, are constant in a neighbourhood of q. Since the functions $\dim \Delta_x^k$ are semicontinuous and integer-valued, Δ is equiregular on a dense open subset of M.

We shall describe $\widehat{M}_q$ in detail only at equiregular points of Δ, though the construction of $\widehat{M}_q$ at other points is not much more complicated. First of all we note that we do not need all sections of the distribution to calculate the flag: we can just take some local basis of this distribution. In fact, let $V_i \in \overline{\Delta}$, $i = 1, \dots, d$, and let $\Delta_q = \operatorname{span}\{V_i(q),\ i = 1, \dots, d\}$. Then any horizontal field V in a neighbourhood of q has a representation $V = \sum_{i=1}^d a_i V_i$, where the a_i are smooth functions. By the Leibniz rule $[V, aW] = a[V, W] + (Va)W$, we now easily deduce that

$$\Delta_q^k = \operatorname{span}\{[V_{i_1}, [V_{i_2}, \dots, V_{i_l}] \dots] \colon 1 \leqslant i_1, \dots, i_l \leqslant d,\ l \leqslant k\}.$$

Here $(Va)(x) \doteq \langle d_x a, V(x) \rangle$ is the derivative of the function a in the direction of V.

Assume now that Δ is equiregular at q. Then we have a whole flag of distributions $\Delta^1 \subset \Delta^2 \subset \dots \subset \Delta^m$ defined in some neighbourhood of q. Let $V \in \overline{\Delta}^i$ and $W \in \overline{\Delta}^j$. Then we use the same Leibniz rule to show that $[V, W](q) \in \Delta_q^{i+j}$, and moreover the projection of $[V, W](q)$ on $\Delta_q^{i+j}/\Delta_q^{i+j-1}$ depends only on the projection of $V(q)$ on $\Delta_q^i/\Delta_q^{i-1}$ and the projection of $W(q)$ on $\Delta_q^j/\Delta_q^{j-1}$.

In other words, the commutator of vector fields induces a Lie algebra structure on the graded space

$$L_q = \bigoplus_{i=1}^m (\Delta_q^i/\Delta_q^{i-1})$$

(we set $\Delta_q^0 = 0$ by definition). It is easy to see that L_q is a nilpotent Lie algebra. Furthermore, this graded nilpotent Lie algebra is generated by the first term of the grading $\Delta_q^1 = \Delta_q$. We also recall that, by the definition of a sub-Riemannian structure, Δ_q is endowed with a Euclidean scalar product.

A finite-dimensional graded nilpotent Lie algebra generated by the first term of the grading, which is endowed with a Euclidean structure, is called a *Carnot algebra*. The simply connected Lie group corresponding to a Carnot algebra is called a *Carnot group*.

Let G_q be the Carnot group associated with the Carnot algebra L_q. Each subspace of a Lie algebra is a left-invariant distribution on the corresponding Lie group.

Thus, $\Delta_q \subset L_q$ is a left-invariant distribution on the Lie group G_q, and it defines on this group a left-invariant sub-Riemannian structure (and therefore a left-invariant sub-Riemannian metric). We call this metric the canonical metric on the Carnot group.

Theorem 2 (Gromov–Mitchell Theorem). *Assume that Δ is equiregular at a point $q \in M$. Then the metric tangent space $(\widehat{M_q}, \widehat{\delta}\,)$ is isometric to the Carnot group G_q endowed with the canonical metric.*

This is a fairly accurate description of the metric tangent space, although it is too abstract so far. We now explain how we obtain this space by passing to the limit as we look at a neighbourhood of the point q in (M, δ) through a more and more powerful microscope.

Let $x^1, \dots, x^n$ be local coordinates on M in a neighbourhood of q such that $x^\iota(q) = 0$ for $\iota = 1, \dots, n$. We say that these coordinates are compatible with the flag $\Delta^1_q \subset \dots \subset \Delta^m_q = T_qM$ if in these coordinates the subspaces Δ^i_q are coordinate subspaces of $\mathbb{R}^n$, more precisely, if Δ^i_q is represented by the subspace $\mathbb{R}^{k_1} \oplus \dots \oplus \mathbb{R}^{k_i}$ for $i = 1, \dots, m$, where

$$\mathbb{R}^{k_i} = \{(0, \dots, 0, x^{k_1+\dots+k_{i-1}+1}, \dots, x^{k_1+\dots+k_i}, 0, \dots, 0) \colon x^\iota \in \mathbb{R}\},$$

$k_i = \dim(\Delta^i_q/\Delta^{i-1}_q)$. In what follows we assume that we have chosen coordinates compatible with the flag.

We know that horizontal paths can take us to all the points, but different directions have different weight: for greater i it is more difficult to go in the direction of $\mathbb{R}^{k_i}$. Using horizontal curves of small length ε, we advance by a (Euclidean) distance of order ε in the direction of $\mathbb{R}^{k_1}$, but only by a distance of order ε^2 in the direction of $\mathbb{R}^{k_2}$, of order ε^3 in the direction of $\mathbb{R}^{k_3}$, and so on.

This can be formalized as follows. We introduce a grading in the algebra of real polynomials in n variables by assigning the following weights to the variables: $w(x^\iota) = i$ for $k_1 + \dots + k_{i-1} < \iota \leqslant k_1 + \dots + k_i$, $i = 1, \dots, m$. Correspondingly, $w(x^{\iota_1} \cdots x^{\iota_l}) = w(x^{\iota_1}) + \dots + w(x^{\iota_l})$. We regard polynomials as differential operators of order zero with polynomial coefficients, and we extend this grading to the whole algebra of linear differential operators with polynomial coefficients by setting $w\biggl(\dfrac{x^\alpha \partial^l}{(\partial x)^\beta}\biggr) = w(x^\alpha) - w(x^\beta)$, where $x^\alpha = x^{\alpha_1} \cdots x^{\alpha_j}$ and $x^\beta = x^{\beta_1} \cdots x^{\beta_l}$ are arbitrary monomials and $(\partial x)^\beta = \partial x^{\beta_1} \cdots \partial x^{\beta_l}$.

We say that a linear differential operator with polynomial coefficients is quasi-homogeneous with weight ν if all the monomial operators in it have weight ν. In this way the whole space of differential operators with polynomial coefficients is decomposed into the direct sum of the spaces of quasi-homogeneous operators with different weights. It is easy to see that the composition of two quasi-homogeneous operators D_1 and D_2 is quasi-homogeneous and $w(D_1 \circ D_2) = w(D_1) + w(D_2)$.

Vector fields are differential operators of order 1. For a pair of quasi-homogeneous vector fields we have $w([V_1, V_2]) = w(V_1) + w(V_2)$. Note that a quasi-homogeneous vector field has weight at least $-m$, and if its weight is non-negative, then the field vanishes at q. Moreover, if $w(V) \geqslant -i$, then $V(q) \in \Delta^i_q$.

Now we define a decreasing filtration on the Lie algebra of smooth vector fields on $\mathbb{R}^n$:

$$\begin{aligned}\operatorname{Vec}(\mathbb{R}^n) = \operatorname{Vec}^{-m}(k_1,\dots,k_m) &\supset \operatorname{Vec}^{1-m}(k_1,\dots,k_m)\\ &\supset \cdots \supset \operatorname{Vec}^{\nu}(k_1,\dots,k_m) \supset \cdots,\end{aligned}$$

where $\operatorname{Vec}^{\nu}(k_1,\dots,k_m)$ is the subspace of fields whose Taylor series expansion contains only monomial fields with weight at least ν.

Definition 2. Coordinates compatible with the flag are said to be *privileged* if $\overline{\Delta} \subset \operatorname{Vec}^{-1}(k_1,\dots,k_m)$.

Exercise. Show that for $m = 2$ any coordinates compatible with the flag are privileged, but this is not so for $m \geqslant 3$.

Theorem 3. *For any totally non-holonomic germ of a distribution there are privileged coordinates.*

As we promised, 'looking through a microscope' is an anisotropic dilation of privileged coordinates according to their weights. Namely, let the *dilation*

$$\eta_s \colon \mathbb{R}^{k_1} \oplus \cdots \oplus \mathbb{R}^{k_m} \to \mathbb{R}^{k_1} \oplus \cdots \oplus \mathbb{R}^{k_m}, \qquad s > 0,$$

be defined by

$$\eta_s(x_1 \oplus \cdots \oplus x_m) = s x_1 \oplus s^2 x_2 \oplus \cdots \oplus s^m x_m.$$

A polynomial ϕ is quasi-homogeneous with weight ν if and only if $\phi \circ \eta_s = s^{\nu}\phi$ for $s > 0$. A polynomial vector field V is quasi-homogeneous with weight ν if and only if $\eta_{s*}V = s^{-\nu}V$ for $s > 0$.

Let $V_1,\dots,V_{k_1} \in \overline{\Delta}$ be an orthonormal basis of our sub-Riemannian structure in a fixed coordinate neighbourhood, and let $\delta(\,\cdot\,,\,\cdot\,)$ be the sub-Riemannian distance corresponding to this structure. Then for each $\varepsilon > 0$ the fields $\varepsilon V_1,\dots,\varepsilon V_{k_1}$ form an orthonormal basis of the rescaled sub-Riemannian structure corresponding to the distance δ/ε. Furthermore, the limit $\lim_{\varepsilon\to 0} \varepsilon(\eta_{1/\varepsilon})_* V = \widehat{V}$ exists for each $V \in \overline{\Delta}$, where $\widehat{V}$ is the quasi-homogeneous field with weight -1 which is equal to the sum of the monomial fields with weight -1 in the Taylor expansion of V. In fact, this is because we are working in privileged coordinates, and the Taylor expansion of V contains no monomial fields with weight < -1.

Rescaling the distance while simultaneously dilating a neighbourhood of q, we get by passing to the limit that the metric tangent space $(\widehat{M_q}, \widehat{\delta}\,)$ is the sub-Riemannian space represented by the sub-Riemannian structure in $\mathbb{R}^n$ with orthonormal basis $\widehat{V}_1,\dots,\widehat{V}_{k_1}$ formed by the quasi-homogeneous fields with weight -1.

As we know, a commutator of quasi-homogeneous fields is quasi-homogeneous, and upon taking the commutator the weights are added. Hence the fields $\widehat{V}_1,\dots,\widehat{V}_{k_1}$ generate a graded nilpotent Lie algebra such that all the iterated commutators of these fields with order greater than m vanish. It is easy to show that if the distribution Δ is equiregular, then this graded Lie algebra has dimension n and is isomorphic to the algebra $L_q = \bigoplus_{i=1}^{m} (\Delta_q^i/\Delta_q^{i-1})$ described above. In the general, non-equiregular case the fields $\widehat{V}_1,\dots,\widehat{V}_{k_1}$ can generate a Lie algebra of dimension greater than n.

Let

$$\Pi(k_1,\dots,k_m;\varepsilon)=\{x_1\oplus\cdots\oplus x_m\colon x_i\in\mathbb{R}^{n_i},\ |x_i|\leqslant\varepsilon^i,\ i=1,\dots,m\}.$$

As a crude conclusion from our looking through a microscope, we can say that for small $\varepsilon>0$ a sub-Riemannian ball has the following upper and lower estimates:

$$c\Pi(k_1,\dots,k_m;\varepsilon)\subset B_q(\varepsilon)\subset c'\Pi(k_1,\dots,k_m;\varepsilon),$$

where c and c' are constants independent of ε.

To fill a cube of fixed size we require about $(1/\varepsilon)^{k_1+2k_2+\cdots+mk_m}$ translated 'boxes' $\Pi(k_1,\dots,k_m;\varepsilon)$ (as $\varepsilon\to 0$). Hence the Hausdorff dimension of the sub-Riemannian space is $k_1+2k_2+\cdots+mk_m$. We see that the Hausdorff dimension is strictly greater than the topological dimension in all cases except the Riemannian case. We can say that a sub-Riemannian metric space has a fractal nature, while topologically it is a quite ordinary manifold.

Example 1. Consider a sub-Riemannian structure in $\mathbb{R}^3$ with a distribution of rank 2. It is easy to see that such a distribution is equiregular at a point q if and only if it defines a contact structure in a neighbourhood of q. Then $\Delta_q^2=\mathbb{R}^3$. Elementary calculations show that, up to a metric-preserving isomorphism, the Heisenberg group is the unique 3-dimensional Carnot group. This is the simplest non-Riemannian sub-Riemannian space. Let $V_1,V_2\in\Delta$ be a left-invariant orthonormal basis of the distribution Δ in this case. Then $V_1,V_2,[V_1,V_2]$ is a basis of the Heisenberg Lie algebra, and $[V_1,V_2]$ commutes with both V_1 and V_2.

In § 1 we considered 3-dimensional sub-Riemannian structures connected with planar isoperimetric problems. It is easy to see that the Heisenberg group corresponds to the most famous of these problems, Dido's problem, in which the function b characterizing the isoperimetric problem is a constant. In Fig. 2 we show how a sub-Riemannian sphere looks in a section on the Heisenberg group.

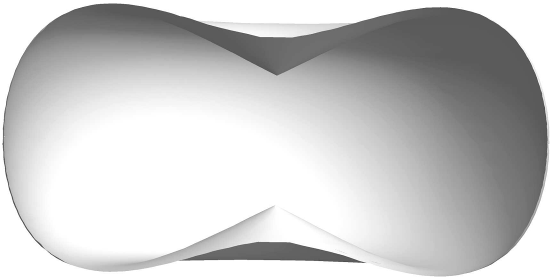

Figure 2

Because of quasi-homogeneity and left invariance, all the spheres in a Carnot group have the same structure and can be obtained one from another using a dilation and a left translation on the group. We see that these spheres, which can have an arbitrarily small radius, have non-smooth points, the points where they intersect the vertical axis; this axis is the translation of the centre of the group that contains the centre of the ball. The presence of non-smooth points is not accidental: the simplest properties of the set of geodesic curves starting from the centre of the ball imply that there must exist non-smooth points.

We fix $q_0 \in M$ and consider the function $c(q) = \delta(q_0, q)$. The balls with centre q_0 are Lebesgue sets of c, and the corresponding spheres are level sets of c. We can show that if $T_{q_0}M \neq \Delta_{q_0}$, then any sphere $S_{q_0}(r)$ with sufficiently small radius contains non-smooth points of c. On the other hand, c is smooth on a dense open subset of a neighbourhood of q_0. This is perhaps all we know about the distance function c in the case of an arbitrary sub-Riemannian structure, apart from the obvious Hölder property with exponent $1/m$ in the case when $\Delta^m = TM$. Of course, in special cases we know much more.

Let us explain the connection between the analytic properties of the distance function and geodesic curves. Recall that a horizontal path $\gamma\colon [0,t] \to M$ is said to be length minimizing if the length of γ is equal to the distance between $\gamma(0)$ and $\gamma(t)$. We say that a length-minimizing path γ is extendable if there exists a length-minimizing path $\overline{\gamma}\colon [0,\bar{t}\,] \to M$, $\bar{t} > t$, such that $\gamma = \overline{\gamma}\big|_{[0,t]}$ and $\overline{\gamma}(\bar{t}\,) \neq \gamma(t)$.

Theorem 4. *Let q be a point in a compact ball with centre q_0. Then the function c is smooth at q if and only if a unique length-minimizing path joins q to q_0, this path is extendable, and it is a strictly normal geodesic.*

At points where c fails to be smooth the nature of its non-smoothness depends essentially on whether the length-minimizing paths terminating at this point are normal or abnormal. Recall that a function on M is said to be semiconcave (semiconvex) if in local coordinates it can be represented as the sum of a concave (convex) function and a smooth function. It is easy to see that this property is independent of the choice of local coordinates. Clearly, any semiconcave (semiconvex) function is locally Lipschitz, and furthermore, it has a second derivative almost everywhere according to Alexandrov's well-known result.

Theorem 5. *If a point q in a compact ball with centre q_0 can be joined to q_0 only by strictly normal length-minimizing paths, then c is semiconcave in some neighbourhood of q.*

In the picture of a ball in the Heisenberg group (see Fig. 2) we easily see that the distance function is semiconcave: the tangent cone of the ball at a non-smooth point is the complement of a convex cone.

Theorem 6. *If all non-constant length-minimizing paths starting at a point q_0 are strictly normal, then the sphere $S_{q_0}(r)$ is a Lipschitz submanifold of M for almost all $r \geqslant 0$ such that the ball $B_{q_0}(r)$ is compact.*

As a counterpoint to the 'normal' situation, we give the following result.

Proposition 3. *If a point q lies in a compact ball with centre q_0 and is joined to q_0 only by strictly abnormal length-minimizing paths, then $|d_{q_n}c| \to \infty$ for any sequence $\{q_n\}$ of smooth points of c that converges to q. In particular, the function $c(q)$ is not locally Lipschitz in a neighbourhood of q.*

We obtain the simplest sub-Riemannian space with abnormal geodesics by considering the isoperimetric problem in $\mathbb{R}^2$ with the 1-form $\omega = (x^1)^2\,dx^2$ (see (1.3)).

Then the function $b\colon \mathbb{R}^2 \to \mathbb{R}$ determined by the conditions $d_x\omega = b(x)\,dx^1 \wedge dx^2$ is linear: $b(x) = 2x^1$, and by Proposition 2 the projections of abnormal geodesics on $\mathbb{R}^2$ are intervals of the line $\{(0,x^2)\colon x^2 \in \mathbb{R}\}$. It is easy to see that the abnormal geodesics themselves are intervals of straight lines $\{(0,x^2,c) \in \mathbb{R}^3\colon x^2 \in \mathbb{R}\}$, where $c \in \mathbb{R}$ can be an arbitrary constant. In this case the abnormal geodesics are at the same time normal. In fact, straight-line intervals are length-minimizing paths for the Euclidean metric, even without imposing an isoperimetric condition.

The vector fields $V_1 = \dfrac{\partial}{\partial x^1}$ and $V_2 = \dfrac{\partial}{\partial x^2} + (x^1)^2\dfrac{\partial}{\partial y}$ form an orthonormal basis of this structure in $\mathbb{R}^3$. The structure is not equiregular at points in the plane $\{(0,x^2,y) : x^2, y \in \mathbb{R}\}$. At the point $q = (0,x^2,y)$ the flag of the distribution has the form

$$\widehat{\Delta}^1_q = \widehat{\Delta}^2_q = \{(\xi_1,\xi_2,0)\colon \xi_1,\xi_2 \in \mathbb{R}\}, \qquad \widehat{\Delta}^3_q = \mathbb{R}^3.$$

This example is in some way universal: while the Heisenberg group describes the metric tangent space at any equiregular point of a sub-Riemannian metric on a 3-manifold, this example describes the metric tangent space at each point q such that $\Delta^1_q = \Delta^2_q \neq \Delta^3_q$. In accordance with our notation above, this is the case $k_1 = 2$, $k_2 = 0$, $k_3 = 1$.

The above model is called the *Martinet flat sub-Riemannian structure*, and the plane $\{(0,x^2,y)\colon x^2, y \in \mathbb{R}\}$ is called the *Martinet plane.* All spheres with centres on the Martinet plane have the same shape, and the set of non-smooth points of the sub-Riemannian distance from a point on the Martinet plane is the plane itself. Let us take the sphere with radius r and centre at the origin. It is smooth outside the Martinet plane, and its intersection with the Martinet plane appears as depicted in Fig. 3.

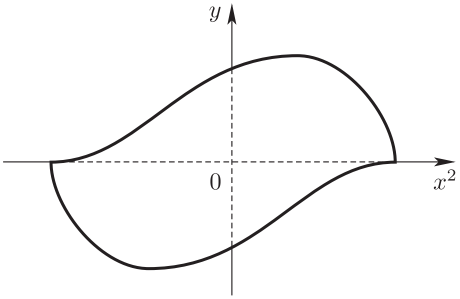

Figure 3

This intersection is a closed curve which fails to be smooth at two points on the sphere that are joined to the centre by abnormal geodesics. Each of the two smooth pieces of the curve is tangent to an abnormal geodesic at one end-point and makes a non-zero angle with it at the other. Note that the distance to the centre increases at a constant rate along each length-minimizing curve when it is parametrized by arc length, for instance, along our abnormal geodesic. On the other hand, the abnormal geodesic is tangent to a curve on the sphere (which is a level set of the distance to the centre). Hence it is immediate that the distance to the centre cannot be a Lipschitz function in a neighbourhood of the corresponding point of the sphere. On the other hand, the sphere itself is a Lipschitz manifold in this case.

Metric tangent spaces are certainly very important, but still very special examples of sub-Riemannian spaces. What can be said about generic spaces? Here we say that a certain property is generic if it holds for each structure in some open dense subset in the Whitney C^∞-topology of the space of sub-Riemannian structures on the given manifold.

It is easy to suppose that a geodesic cannot be both normal and abnormal in the generic case. Moreover, if two abnormal extremals have the same projection, then these extremals are proportional in T^*M and differ by a multiplicative constant. In this case we say that all the extremals (normal and abnormal alike) *have corank* 1. In fact, this property holds not only for individual generic structures, but also for families of sub-Riemannian structures which depend smoothly on a finite number of real parameters. That is taken to mean that all extremals of all structures in a generic family have corank 1, which means (according to the usual terminology of singularity theory) that the set of sub-Riemannian structures not all of whose extremals have corank 1 is a subset of infinite codimension in the space of all sub-Riemannian structures.

The condition that all extremals have corank 1 has certain implications for the local structure of spheres, which we will now describe. Let $T_qM \supset E_0$ be a co-oriented hyperplane in the tangent space. (By definition, this means that one of the two half-spaces into which the hyperplane partitions T_qM is regarded as positive and the other as negative.) Thus, $T_qM = E_- \cup E_0 \cup E_+$, where $E_\pm$ are open half-spaces.

Let $q \in S_{q_0}(r)$. We will say that E_0 is a tangent hyperplane to the ball $B_{q_0}(r)$ at the point q if for each smooth curve $\phi\colon (-1,1) \to M$ with $\phi(0) = q$ that is transversal to E_0 there exists an $\varepsilon > 0$ such that $\phi(-\varepsilon, 0) \subset B_{q_0}(r)$ and $\phi(0,\varepsilon) \cap B_{q_0}(r) = \varnothing$. Clearly, a ball can have at most one tangent hyperplane at q, but it may also have none, for instance, as at points of the abnormal length-minimizing curve in the Martinet flat structure considered above.

Proposition 4. *If all the geodesics have corank* 1 *and* q_0 *is joined to* q *by an extendable length-minimizing path, then the ball* $B_{q_0}(r)$ *has a tangent hyperplane* E_0 *at* q*. Let* γ *be a length-minimizing path with endpoints* $\gamma(0) = q_0$ *and* $\gamma(r) = q$ *which is parametrized by arc length. If* γ *is a normal geodesic, then* $\dot\gamma(r) \in E_+$*. If* γ *is an abnormal geodesic differentiable at the point* r*, then* $\dot\gamma(r) \in E_0$*.*

We look more closely at the case of generic sub-Riemannian structures in $\mathbb{R}^3$. For an orthonormal basis $V_1(x)$, $V_2(x)$, $x \in \mathbb{R}^3$, the equation $\det(V_1, V_2, [V_1, V_2]) = 0$ defines a smooth surface Σ, called the Martinet surface:

$$\Sigma = \{x \in \mathbb{R}^3 \colon \det(V_1(x), V_2(x), [V_1, V_2](x)) = 0\}.$$

Our structure is equiregular at all points in $\mathbb{R}^3 \setminus \Sigma$, and not equiregular at the points in Σ.

At all points in Σ except for a discrete subset of the surface, the distribution

$$\Delta_x = \operatorname{span}\{V_1(x), V_2(x)\}$$

is transversal to Σ, so the subspace $\Delta_x \cap T_x\Sigma$ is 1-dimensional. The non-exceptional points x are characterized by the condition $\Delta^1_x = \Delta^2_x \neq \Delta^3_x$ and are called Martinet points. The lines $\Delta_x \cap T_x\Sigma$, $x \in \Sigma$, form a direction field on Σ (with isolated singularities). The integral curves of this direction field are abnormal geodesics: a non-constant admissible curve is an abnormal geodesic if and only if it lies entirely on the Martinet surface.

Let x be a Martinet point; by Proposition 4 small spheres with centre x have tangent planes at points of the abnormal geodesic passing through x, and the intersection of the sphere with the Martinet surface appears roughly as shown in Fig. 4.

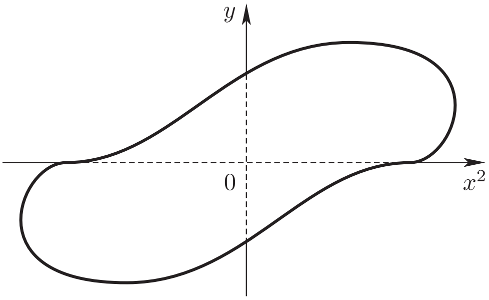

Figure 4

In the generic case this intersection is distinct from the set of non-smooth points of the sphere. Moreover, simple topological considerations suggest that in this case the set of non-smooth points of the sphere cannot be a closed curve, but must consist of two contractible connected components.

Let us now consider small spheres with centre at a point $x \in \mathbb{R}^3 \setminus \Sigma$. The metric tangent space at x is isometric to the Heisenberg group discussed above, and it is highly symmetric. In passing from a sphere in the metric tangent space to a small generic sphere, the symmetry breaks down and the structure of the singularities changes.

The structure of the singularities of a sphere is closely connected with the singularities of the family of geodesic curves issuing from its centre. Let us explain this. In our case all the geodesic curves are normal, and therefore are the projections on $M = \mathbb{R}^3$ of the trajectories of the Hamiltonian system

$$\dot{\lambda} = \vec{h}(\lambda), \quad \lambda \in T^*M, \qquad h(\lambda) = 1$$

(see § 1). It is easy to see that these projections are parametrized by arc length. Geodesics from x are the projections of trajectories with the initial condition $\lambda(0) = T^*_x M \cap h^{-1}(1)$.

Let $\pi\colon T^*M \to M$ be the standard projection: $\pi(T^*_x M) = x$. We set $H_x = T^*_x M \cap h^{-1}(1)$ and consider the *exponential map*

$$\mathscr{E}_x\colon (t, \lambda) \mapsto \pi \circ e^{t\vec{h}}(\lambda), \qquad t > 0, \quad \lambda \in H_x. \tag{2.1}$$

Note that the curves $t \mapsto \mathscr{E}_x(t,\lambda)$ are geodesics issuing from x, and each compact ball $B_x(r)$ is the image of the set $[0,r] \times H_x$ under the exponential map: $B_x(r) = \mathscr{E}_x([0,r] \times H_x)$. Furthermore, $S_x(r) \subset \mathscr{E}_x(r, H_x)$. This inclusion is always strict, because not all geodesic curves with length r are length minimizing no matter how small r is.

How can we distinguish between a normal geodesic that is length minimizing and one that is not? The recipe is generally the same as in classical Riemannian geometry. A small interval of a geodesic curve

$$\gamma(t) = \mathscr{E}_x(t,\lambda), \qquad t > 0, \tag{2.2}$$

is length minimizing by definition. A strictly normal geodesic ceases to be length minimizing after going through a *cut point* or the first *conjugate point*. These points are very important for us, also because, according to Theorem 4, the distance function fails to be smooth just at the endpoints of non-extendable length-minimizing curves.

Definition 3. The *cut time along the geodesic curve* (2.2) is the number

$$\bar{t} = \min\{t > 0\colon \exists\, \lambda' \in H_x,\ \lambda' \neq \lambda,\ \gamma(t) = \mathscr{E}_x(t,\lambda')\},$$

and the point $\gamma(\bar{t}) = \mathscr{E}(\bar{t},\lambda)$ is called a *cut point*.

A *conjugate time along* γ is a number $\widehat{t} > 0$ such that $(\widehat{t},\lambda)$ is a critical point of the map $\mathscr{E}_x$, and the point $\gamma(\widehat{t}) = \mathscr{E}(\widehat{t},\lambda)$ is called a *conjugate point*.

That is, a cut point is a point at which the geodesic first meets another geodesic from x which has the same length. Conjugate points are points on the 'enveloping surface' of the family of geodesics from x. Using a slightly outdated language, we can say that at a conjugate point the geodesic meets an infinitesimally close geodesic from x. These two types of points play slightly different roles: a geodesic stops being length-minimizing past a cut point, but it remains locally length minimizing (that is, shorter than all the C^0-close admissible paths with the same endpoints) up to the first conjugate point. Past the first conjugate point it is no longer locally length minimizing.

On the Heisenberg group geodesic curves are circular helices, and their 'enveloping surface' degenerates into a straight line (the vertical coordinate axis), and furthermore their cut points and first conjugate points coincide, so that a geodesic ceases to be globally and locally length minimizing at the same time. In the generic case this is not so, and the endpoints of almost all non-extendable length-minimizing curves are cut points which are not conjugate points.

The set of terminal points of all non-extendable length-minimizing curves issuing from x is called the *cut locus*, and the set of first conjugate points is called the *first caustic*. We can show that the cut locus lies in the closure of the set of cut points, and this holds not just in the 3-dimensional problem under consideration, but also for any sub-Riemannian structure all of whose length-minimizing paths are strictly normal. We note that the initial point x also lies in the closure of the set of cut points, provided that $\Delta_x \neq T_xM$.

In the next section we will consider details of the structure of the cut locus and the first caustic in a neighbourhood of x for a germ of a generic 3-dimensional sub-Riemannian structure (see Fig. 9 in the next section).

Comments. Theorem 2 was proved in [19] (see also [17]), and Theorem 3 was proved in [8] and [11] (see also the subsequent paper [9]). The proofs of Theorem 4 and Proposition 3 can be found [3], as well as the proof of Theorem 6, which is due to Rifford [25]. Theorem 5 is a consequence of a more general result due to Cannarsa and Rifford [12]. A sphere in the Heisenberg group was described in [28], one of the first papers on sub-Riemannian geometry, and a sphere in a flat Martinet structure was described in [6]. That sub-Riemannian structures all of whose extremals have corank 1 are typical was shown in [14], and Proposition 4 was proved in [2].

3. Curvature

For who can make that straight,
which he hath made crooked?[1]

The definition of sub-Riemannian curvature is based on an idea going back essentially to Gauss with his geodesic triangles. Let A, B, and C be three sufficiently close points on a sub-Riemannian manifold. We connect A and B with C by length-minimizing paths. Assume that these paths are normal geodesics. Then in the case of negative curvature the picture must be roughly as in Fig. 5, and for positive curvature it must be as in Fig. 6.

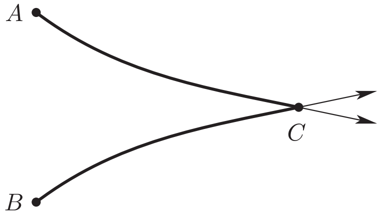

Figure 5

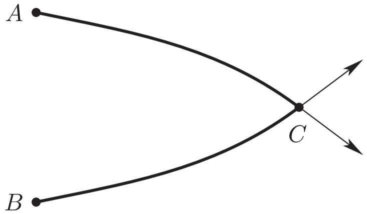

Figure 6

In other words, the greater the curvature, the larger the vector difference between the velocities of the corresponding geodesics at C. This idea can be rigorously expressed using the following construction.

This construction does not involve all normal geodesics, but only *ample* ones. We explain the meaning of this term a little later; here it is sufficient to know that for any $q_0 \in M$ almost all geodesics from q_0 are ample. Recall that normal geodesics are the projections on M of trajectories of the Hamiltonian system $\dot{\lambda} = \vec{h}(\lambda),\ h(\lambda) = 1$, and these geodesics are automatically parametrized by arc length.

The Hamiltonian h is homogeneous of degree 1 on fibres of the cotangent bundle, so the trajectories of the system $\dot{\lambda} = \vec{h}(\lambda)$, $h(\lambda) = c$, with an arbitrary $c > 0$ have the same projections on M, with the same parametrization. In many respects it is more convenient to work with the Hamiltonian $h^2(\lambda)/2$, which is quadratic on

[1] Ecclesiastes 7:13.

fibres, and with the corresponding Hamiltonian system

$$\dot{\lambda} = h(\lambda)\vec{h}(\lambda), \qquad \lambda \in T^*M. \tag{3.1}$$

The projections of trajectories of this system on M are the same geodesics, and the parameter on them is proportional to arc length but does not necessarily coincide with it. These projections are easy to express in terms of the exponential map (2.1). If $\lambda(0) \in T^*_{q_0}M$ and $h(\lambda(0)) = c > 0$, then the projection of a trajectory $t \mapsto \lambda(t)$ is the geodesic $t \mapsto \mathscr{E}_{q_0}(ct, \lambda(0)/c)$. On the other hand, if $h(\lambda(0)) = 0$, then $\lambda(0)$ is a fixed point.

Thus, let $\gamma\colon t \mapsto \gamma(t)$, $\gamma(0) = q_0$, be an ample geodesic with parameter proportional to arc length. Fix a sufficiently small $t > 0$; then for each q in some neighbourhood of q_0 in M there exists a unique length-minimizing path $\gamma_q\colon [0,t] \to M$ with parameter proportional to arc length such that $\gamma_q(0) = q$ and $\gamma_q(t) = \gamma(t)$ (see Fig. 7).

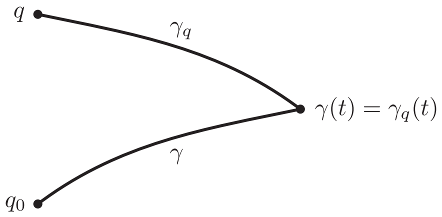

Figure 7

Consider the function $q \mapsto b_t(q) \doteq |\dot{\gamma}_q(t) - \dot{\gamma}(t)|^2/2$, which is a smooth function defined in a neighbourhood of q_0. In addition, it clearly takes its minimum value at q_0. Thus, $d_{q_0}b_t = 0$, and the Hessian $D^2_{q_0}b_t$ is a well-defined non-negative quadratic form on $T_{q_0}M$. We will extract the curvature from the asymptotic behaviour of the family of quadratic forms $D^2_{q_0}b_t\big|_{\Delta_{q_0}}$ as $t \to 0$.

To understand the nature of this behaviour we start by calculating b_t at points of the geodesic γ. We obtain

$$\gamma_{\gamma(s)}(\tau) = \gamma\left(s + \frac{t-s}{t}\tau\right), \qquad 0 \leqslant \tau \leqslant t,$$

$$b_t(\gamma(s)) = \frac{s^2}{2t^2}|\dot{\gamma}(t)|^2 = \frac{s^2}{2t^2}|\dot{\gamma}(0)|^2.$$

Hence $D^2_{q_0}b_t(\dot{\gamma}(0)) = |\dot{\gamma}(0)|^2/t^2$.

Exercise. Let M be a Euclidean space (that is, $\mathbb{R}^n$ with the Euclidean metric). Then the function b_t is the same for all the geodesics starting from q_0:

$$b_t(q) = \frac{|q - q_0|^2}{2t^2}, \qquad D^2_{q_0}b_t(v) = \frac{1}{t^2}|v|^2, \quad v \in \mathbb{R}^n.$$

Next we take the general Riemannian case. According to Riemann's original construction, the Riemannian curvature measures the infinitesimal deviation of the square of the Riemannian distance from the square of the Euclidean distance, so the formula below should come as no surprise. In the Riemannian case all geodesics are ample, and the following asymptotic relation holds:

$$D^2_{q_0} b_t(v) = \frac{1}{t^2}|v|^2 + \frac{1}{3}\langle R(v|\dot\gamma)\dot\gamma|v\rangle + O(t), \qquad v \in T_{q_0}M,$$

as $t \to 0$. Here R is the Riemann tensor: if $|\dot\gamma(0)| = |v| = 1$ and $\langle\dot\gamma(0)|v\rangle = 0$, then $\langle R(v|\dot\gamma)\dot\gamma|v\rangle$ is the sectional curvature in the direction of $\operatorname{span}\{\dot\gamma(0), v\}$.

Finally, we consider the most general sub-Riemannian case.

Theorem 7. *Let γ be an ample geodesic. Then there exist quadratic forms Q and R_γ on Δ_{q_0} such that $Q(v) \geqslant |v|^2$ for $v \in \Delta_{q_0}$ and*

$$D^2_{q_0} b_t(v) = \frac{1}{t^2}Q_\gamma(v) + \frac{1}{3}R_\gamma(v) + O(t), \qquad v \in \Delta_{q_0},$$

as $t \to 0$.

The quadratic form R_γ is called the curvature form in the direction of the geodesic γ, and the positive-definite form Q_γ involved in the first term of the asymptotic formula is a dimension invariant and characterizes the anisotropy of the sub-Riemannian metric at small distances. The equality

$$Q_\gamma(v) = |v|^2 \qquad \forall\, v \in \Delta_{q_0}$$

holds if and only if $\Delta_{q_0} = T_{q_0}M$, that is, only in the Riemannian case.

Now we give the definition of an ample geodesic and at the same time find out how to calculate the spectrum of the form Q and try to clarify its geometric meaning. In fact, ampleness depends only on the distribution and the germ of a horizontal curve: it has nothing to do with the Euclidean structure on Δ and the fact that γ is a normal geodesic.

Definition 4. The germ of a smooth horizontal curve at a point q_0 is said to be *ample* if there exists a $k > 0$ such that the k-jet of the germ is distinct from the k-jets of the projections on M of abnormal extremals.

This is too abstract a definition, so we make it concrete and find effective checks for ampleness. Let $\Phi_t\colon M \to M$, $t \in \mathbb{R}$, be a horizontal flow such that $\Phi_t(q_0) = \gamma(t)$, $t \geqslant 0$. In other words, γ is a trajectory of Φ_t. Taking the subspaces $\Delta_{\gamma(t)} \subset T_{\gamma(t)}M$, we shift them to q_0 with the help of the flow. This gives us the family of subspaces

$$\Delta^t_{q_0} = (\Phi_t)^{-1}_* \Delta_{\gamma(t)}, \qquad \Delta^t_{q_0} \subset T_{q_0}M.$$

It is easy to show that $\gamma\big|_{[0,t]}$ is the projection on M of an abnormal extremal if and only if $\operatorname{span}\{\Delta^\tau_{q_0},\ 0 \leqslant \tau \leqslant t\}$ is a proper subspace of $T_{q_0}M$.

We now define the flag of Δ along the horizontal curve γ:

$$\Delta^{(i+1)}_\gamma = \frac{d^i}{dt^i}\Delta^t_{q_0}\bigg|_{t=0}, \qquad i = 0, 1, 2, \dots .$$

Here by definition the ith derivative of a family of subspaces is understood to be the linear span of the ith derivatives of all the sections $t \mapsto v(t) \in \Delta^t_{q_0}$, $t \geqslant 0$. It is easy to see that

$$\Delta_{q_0} = \Delta^{(1)}_\gamma \subset \Delta^{(2)}_\gamma \subset \dots \subset \Delta^{(i)}_\gamma \subset \cdots, \tag{3.2}$$

and the subspaces $\Delta^{(i)}_\gamma$ depend only on Δ and γ, and not on the choice of the flow Φ_t. Moreover, the germ of γ is ample if and only if there exists a k such that $\Delta^{(k)}_\gamma = T_{q_0}M$.

Of course, we need not differentiate all the sections $\Delta^t_{q_0}$ to calculate Δ^i_γ, nor is it necessary to find the family $\Delta^t_{q_0}$ itself. Let $\Phi_t = e^{tX}$ and $\Delta_q = \operatorname{span}\{V_1(q), \dots, V_d(q)\}$, where X and $V_1, \dots, V_d$ are horizontal vector fields. Recall that $\dfrac{d}{dt} e^{-tX}_* V = e^{-tX}_*[X, V]$, from which it is easy to get that

$$\Delta^{(i+1)}_\gamma = \operatorname{span}\{\underbrace{[X, \dots, [X}_{j}, V_l] \dots] \colon 0 \leqslant j \leqslant i,\ 1 \leqslant l \leqslant d\}.$$

We see that the flag (3.2) is a microlocal version of the flag of a distribution, which we considered in § 2. In the construction of the flag of a distribution we take all the iterated commutators of horizontal vector fields, but for the flag along a curve we iterate only commutators with the fixed field generating this curve. Of course, the flags along distinct horizontal curves going out from a given point can be distinct. We are mostly interested in flags along normal geodesic curves.

Let $\xi \in T^*_{q_0}M$, and let $\bar{\xi}$ denote the geodesic corresponding to the solution of (3.1) with the initial condition $\lambda(0) = \xi$. In other words, $\bar{\xi}(t) = \mathscr{E}\left(h(\xi)t, \dfrac{1}{h(\xi)}\xi\right)$ if $h(\xi) \neq 0$, and $\bar{\xi}(t) \equiv q_0$ if $h(\xi) = 0$. With each $\xi \in T^*_{q_0}M$ we associate the flag

$$\Delta_{q_0} = \Delta^{(1)}_{\bar{\xi}} \subset \Delta^{(2)}_{\bar{\xi}} \subset \dots \subset T_{q_0}M.$$

Proposition 5. *The integer-valued functions*

$$\xi \mapsto \dim \Delta^{(i)}_{\bar{\xi}}, \qquad \xi \in T^*_{q_0}M, \quad i = 1, 2, \dots, \tag{3.3}$$

are lower semicontinuous on $T^*_{q_0}M$. In addition, there exists a Zariski-open subset $\mathscr{O}_{q_0}$ of $T^*_{q_0}M$ such that for each $\xi \in \mathscr{O}_{q_0}$ the geodesic curve $\bar{\xi}$ is ample, and the functions $\xi \mapsto \dim \Delta^{(i)}_{\bar{\xi}}$, $i = 1, 2, \dots$, are constant on $\mathscr{O}_{q_0}$ and are equal to their maximum values on $T^*_{q_0}M$.

Next we turn to the spectrum of the quadratic form Q_γ. We calculate it not for an arbitrary ample germ, but only for germs satisfying the additional (but not too restrictive) condition of *equiregularity*. Let us explain this condition.

So far we have fixed a point $q_0 \in M$ and considered the germs of all normal geodesics starting from this point. Now we fix a geodesic $\gamma\colon [0, t_1] \to M$ and consider its germs at different points. Namely, for each $s \in [0, t_1)$ we set $\gamma_s(t) = \gamma(s + t)$. Then $\Delta^{(1)}_{\gamma_s} \subset \Delta^{(2)}_{\gamma_s} \subset T^*_{\gamma(s)}M$. It is easy to see that the integer-valued functions $s \mapsto \dim \Delta^{(i)}_{\gamma_s}$ are lower semicontinuous. They are locally constant on an open dense subset of the half-open interval $[0,t_1)$. The germ of γ at zero is said to be *equiregular* if the functions $s\mapsto \dim\Delta^{(i)}_{\gamma_s}$, $i=1,2,\dots$, are locally constant in a neighbourhood of zero, so that $\dim\Delta^{(i)}_{\gamma_s}=\dim\Delta^{(i)}_{\gamma}$ for all sufficiently small $s>0$.

We illustrate these concepts using the sub-Riemannian structures in $\mathbb{R}^3$ that we considered in § 2. If the initial point $q_0=x$ lies outside the Martinet surface, then all non-constant geodesics are ample and equiregular, and the flags along their germs coincide with the flag of the distribution at q_0. In particular, $\Delta^{(2)}_{\overline{\xi}}=\mathbb{R}^3$. But if $q_0=x$ is a Martinet point, then all geodesics which are not abnormal are ample, but none of them are equiregular. For $\xi\in\mathscr{O}_{q_0}$ we get that $\Delta^{(2)}_{\overline{\xi}}=\Delta^{(1)}_{\overline{\xi}}=\Delta_{q_0}$ and $\Delta^{(3)}_{\overline{\xi}}=\mathbb{R}^3$. To pass to equiregular germs we must move slightly along the geodesic, thereby leaving the Martinet surface.

In the general case we can supplement Proposition 5 in the following way: *for all q_0 in some open dense subset of M, ample geodesics $\overline{\xi}$, $\xi\in\mathscr{O}_{q_0}$, are equiregular.*

We now concentrate on the equiregular case. Let $\dim\Delta_{q_0}=d$ and let $\Delta^m_\gamma=T_{q_0}M$. We set $d_1=d$ and $d_{i+1}=\dim\Delta^{(i+1)}_\gamma-\dim\Delta^{(i)}_\gamma$ for $i=1,\dots,m-1$. In the equiregular case the sequence of numbers $d_1,\dots,d_m$ is non-increasing: $d_{i+1}\leqslant d_i$ for $1\leqslant i\leqslant m-1$. We form the Young diagram with columns of lengths $d_1,\dots,d_m$: see Fig. 8.

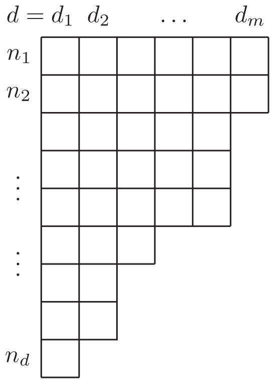

Figure 8

Let the rows of this diagram have lengths $n_1,\dots,n_d$. Then

$$\operatorname{spec} Q_\gamma=\{n_1^2,\dots,n_d^2\}. \tag{3.4}$$

Recall that Q_γ is a quadratic form on the d-dimensional Euclidean space Δ_{q_0}, and its spectrum consists of the eigenvalues of the corresponding symmetric operator on Δ_{q_0}. Some of the integers $n_1,\dots,n_d$ may be repeated, with the number of repetitions corresponding to the multiplicity of the eigenvalue.

Remark 2. For a non-equiregular germ of an ample geodesic the eigenvalues of the form Q_γ need not be integers. For instance, if $M=\mathbb{R}^3$, then for the germs of ample geodesics at a Martinet point we have $\operatorname{spec} Q_\gamma=\{1,9/4\}$.

The geometric meaning of the form Q_γ can be well illustrated by the notion of the *geodesic dimension* of the sub-Riemannian space M at the point q_0. Here is the definition.

Consider a compact subset C of M with non-empty interior (for example, a ball with small radius and centre q_0). We join all points in C to q_0 by length-minimizing geodesics, with parameter proportional to arc length and ranging in the same interval $[0,1]$ for all geodesics. Thus, a point $x \in C$ is connected with q_0 by a geodesic curve $\gamma_x\colon [0,1] \to M$ such that $\gamma_x(0) = q_0$ and $\gamma_x(1) = x$. If q_0 and x can be joined by several length-minimizing paths, then we take all of them, so that γ_x is not a single geodesic but a set of geodesics.

We now want to contract C to q_0 along the length-minimizing paths. Let $C_t = \{\gamma_x(t)\colon x \in C\}$ for $0 \leqslant t \leqslant 1$; then $C_1 = C$ and $C_0 = \{q_0\}$. Assume that a volume form is defined on M, so that the volumes $\operatorname{vol}(C_t) > 0$ are defined for $0 < t \leqslant 1$. The *geodesic dimension* of the sub-Riemannian space M at q_0 is equal to the limit

$$\lim_{t\to 0} \frac{1}{t} \log \operatorname{vol}(C_t)$$

if it exists and is independent of the choice of the set C and the volume form. In other words, the geodesic dimension is the order of convergence to zero of the volume of C_t as $t \to 0$.

Theorem 8. *Assume that for $\xi \in \mathscr{O}_{q_0}$ the geodesics $\overline{\xi}$ are equiregular and the abnormal length-minimizing paths from a point q_0 sweep out a subset of M of measure zero. Then the geodesic dimension of the sub-Riemannian space M at q_0 is equal to $\operatorname{tr} Q_{\overline{\xi}}$ for any $\overline{\xi} \in \mathscr{O}_{q_0}$.*

Some explanations: we have already mentioned that the condition of equiregularity of the geodesics $\overline{\xi}$ for $\xi \in \mathscr{O}_{q_0}$ holds for all q_0 in an open dense subset of M. As regards the condition on the abnormal length-minimizing paths issuing from q_0, it holds at all points in all the known examples of sub-Riemannian spaces, but we know of no proof that this is always the case. We know that such paths sweep out a closed nowhere dense subset of M, but do not know whether it can have positive measure. This is an important open question in sub-Riemannian geometry.

If the assumptions of the theorem are satisfied, then we have an explicit formula for the geodesic dimension (see (3.4)): it is equal to

$$\sum_{i=1}^{d} n_i^2 = \sum_{i=1}^{m} (2i-1)d_i.$$

It is interesting to compare this formula with the Hausdorff dimension calculated in § 2. We can do this for $m = 2$, that is, when $\Delta^{(1)}_{\xi} \neq \Delta^{(2)}_{\xi} = T_{q_0}M$ for $\xi \in \mathscr{O}_{q_0}$. Then the flags along geodesics coincide with the flag of the distribution at q_0. Thus, the Hausdorff dimension of the sub-Riemannian space M is equal to $d + 2(n-d)$, while its geodesic dimension is $d + 3(n-d)$, where n is the topological dimension of M. That is, the geodesic dimension is strictly greater than the Hausdorff dimension.

Recall that the Hausdorff dimension is equal to the order with which the volume of a ball tends to zero as its radius tends to zero. The inequality obtained is a very crude manifestation (on the level of dimensions) of a property of sub-Riemannian (but not Riemannian) spaces which we have already pointed out: there exist plenty of arbitrarily short geodesics that are not length-minimizing paths.

For example, as the ball $C = B_{q_0}(r)$ contracts to its centre along length-minimizing paths, the set C_t sweeps out only a certain part of the ball $B_{q_0}(tr)$, because as the length decreases there appear more and more geodesics starting from q_0 that are length minimizing (they do not attain the sphere $S_{q_0}(r)$, but do attain $S_{q_0}(tr)$). Moreover, C_t 'flattens' much more rapidly than the balls $B_{q_0}(tr)$ as $t \to 0$. All this is already quite clear in the simplest case of the Heisenberg group, when the topological dimension is three, the Hausdorff dimension is four, and the geodesic dimension is five.

Let us proceed to a discussion of the curvature forms R_γ. It is easy to show that $\dot\gamma(0) \in \ker R_\gamma$, so that the quadratic form R_γ is actually defined on the quotient space $\Delta_{q_0}/\mathbb{R}\dot\gamma(0)$. It is important to know how $R_{\bar\xi}$ depends on $\xi \in T^*_{q_0}M$.

We can show that $\xi \mapsto R_{\bar\xi}(v)$ is a rational function which is positive homogeneous of homogeneity degree 2 for each $v \in \Delta_{q_0}$. It is finite on $\mathscr{O}_{q_0}$ and can have poles outside $\mathscr{O}_{q_0}$. Recall that in the Riemannian case $R_{\bar\xi}(v) = \langle R(v, \dot{\bar\xi})\dot{\bar\xi}, v\rangle$, where R is the Riemannian curvature. That is, in the Riemannian case the function $\xi \mapsto R_{\bar\xi}(v)$ is just a quadratic form, but in the general case it is not.

In general, we know very little about the structure of these functions, and serious investigations of them have only just begun. If we write the rational function $\xi \mapsto R_{\bar\xi}(v)$ in local coordinates, then its coefficients are themselves rational functions of partial derivatives of the Hamiltonian h. However, the explicit formulae are very complicated and almost useless for understanding the geometry.

In this paper we confine ourselves to a description of the forms $R_{\bar\xi}$ in the well-understood case of a contact sub-Riemannian structure on a 3-manifold. In this case the functions $\xi \mapsto R_{\bar\xi}(v)$ are quadratic forms on $T^*_{q_0}M$.[2]

Thus, let Δ be a contact distribution on a 3-dimensional Riemannian manifold M. Then $\dim \Delta_{q_0} = 2$, and $R_{\bar\xi}$ is a quadratic form on the 1-dimensional Euclidean space $\Delta_{q_0}/\mathbb{R}\dot\gamma(0)$, that is, in essence a real number (the single eigenvalue of the symmetric operator corresponding to this quadratic form). We denote this number by $r(\xi)$. Then the function $\xi \mapsto r(\xi)$ is itself a quadratic form on $T^*_{q_0}M$. Calculations show that its restriction to the orthogonal complement of Δ_{q_0} is a positive-definite form: $r\big|_{\Delta^\perp_{q_0}} > 0$.

Therefore, the curvature of a contact sub-Riemannian structure cannot vanish identically. At first glance this seems to be in contrast to the usual Riemannian situation, but if we look more closely, we see that, on the contrary, it is in full agreement. In fact, in a sub-Riemannian but non-Riemannian space we always have arbitrarily short geodesic curves that are not locally length minimizing. Thus, the curvature not only cannot be identically equal to zero, but it even must not be uniformly bounded above on the space of geodesics of fixed length: this is what we actually have.

Recall that a geodesic $\bar\xi$ has parameter proportional to arc length, and $|\dot{\bar\xi}| = h(\xi)$ for $\xi \in T^*_{q_0}M$. Furthermore, h^2 is a quadratic form with kernel $\Delta^\perp_{q_0}$. If we add

[2] This is not so for contact sub-Riemannian structures on manifolds of higher dimension.

to a given $\xi \in T^*_{q_0}M$ a sufficiently large element of $\Delta^{\perp}_{q_0}$, then we increase the curvature without changing the initial velocity on the geodesic $\overline{\xi}$, and we make the first conjugate time (see Definition 3 at the end of § 2) closer to zero.

In actual fact there exists a very close not only qualitative but also quantitative connection between the quadratic form $r(\xi)$, $\xi \in T^*_{q_0}M$, and the structure of the first caustic and the cut locus near the point q_0. We describe this connection now. It will be convenient to renormalize the form $r(\,\cdot\,)$ by taking $5r(\,\cdot\,)/2$ instead.

A quadratic form on a finite-dimensional vector space can be represented (in many ways) as a linear combination of squares of linear forms. If the space is equipped with a Euclidean structure, then we obtain a canonical representation by taking the squares of the linear forms in a suitable orthonormal basis.

The quadratic form $r(\,\cdot\,)$ is defined on $T^*_{q_0}M$, and only the subspace Δ_{q_0} of $T_{q_0}M$ is endowed with a Euclidean structure. We know that $r\big|_{\Delta^{\perp}_{q_0}} > 0$. This information is sufficient to derive the following canonical representation of the form $5r/2$ as a linear combination of squares:

$$\frac{5}{2}r(\xi) = \langle \xi, f_0\rangle^2 + \alpha_1\langle \xi, f_1\rangle + \alpha_2\langle \xi, f_2\rangle,$$

where $\alpha_1 \geqslant \alpha_2$, and f_1 and f_2 lie in Δ_{q_0} and form an orthonormal basis in this Euclidean plane. The vector f_0 is transversal to the plane Δ_{q_0} and is determined up to a sign. If $\alpha_1 \neq \alpha_2$, then f_1 and f_2 are also determined up to a sign; otherwise we can take any orthonormal basis in Δ_{q_0}. We leave the simple derivation of this canonical form to the reader as an exercise.

Note that $h^2(\xi) = \langle \xi, f_1\rangle^2 + \langle \xi, f_2\rangle^2$ and $\dot{\overline{\xi}}(0) = \langle \xi, f_1\rangle f_1 + \langle \xi, f_2\rangle f_2$. We will use the special notation $\nu = 1/|\langle \xi, f_0\rangle|$ for the reciprocal of the absolute value of the third coordinate of ξ, which does not affect the initial velocity on the geodesic $\overline{\xi}$. We also set $\kappa = (\alpha_1 + \alpha_2)/2$ and $\chi = (\alpha_1 - \alpha_2)/6$; the quantities κ and χ are the principal numerical invariants of the sub-Riemannian structure.

Let $\operatorname{length}_{\mathrm{conj}}(\gamma)$ be the length of the interval of the geodesic γ up to the first conjugate point. Then the following asymptotic expression holds as $\nu \to 0$:

$$\operatorname{length}_{\mathrm{conj}}(\overline{\xi}) = 2\pi\nu - \pi\kappa\nu^3 + O(\nu^4).$$

Assume that $\chi \neq 0$, that is, $\alpha_1 > \alpha_2$. Figure 9 shows how the intersection of the first caustic and the cut locus with a small neighbourhood of q_0 looks in this case.

The tangent cone at q_0 to both the caustic and the cut locus is equal to the line $\mathbb{R}f_0$. The first caustic is the union of two pyramids with vertex q_0, one corresponding to the positive values of $\langle \xi, f_0\rangle$ and the other to the negative values. Each pyramid has four cuspidal edges, which asymptotically are semicubical parabolas. If we take $\mathbb{R}f_0$ to be the vertical axis, then asymptotically the horizontal sections of the first caustic are asteroids with centre on the axis $\mathbb{R}f_0$, and they are symmetric asteroids (whose two diagonals are equal). The vertices of these asteroids lie on straight lines parallel to the axes $\mathbb{R}f_1$ and $\mathbb{R}f_2$.

The first conjugate point on ξ has height $\pi\nu^2 + O(\nu^4)$, and the length of the diagonal of the 'asteroid' obtained as the section of the caustic at this height is $4\pi\chi\nu^3 + O(\nu^4)$. A horizontal section of the cut locus is asymptotically the diagonal

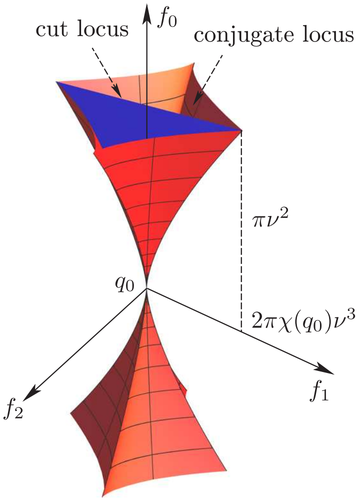

Figure 9

of the asteroid that connects the vertices lying on a line parallel to $\mathbb{R}f_1$. Here (as above) we write 'asymptotically' to avoid a lengthy and, in our opinion, rather cumbersome description, which is clear enough from the picture. Alternatively, we could write: the horizontal section of the cut locus at height $\pi\nu^2$ is a smooth curve connecting the opposite cuspidal edges of the caustic; after scaling the horizontal (but not the vertical!) coordinates with the factor $1/\nu^3$ and letting ν go to zero, this curve converges uniformly to the straight-line interval $(-2\pi\chi f_1, 2\pi\chi f_1)$.

We see that the curvature $R_{\overline{\xi}}$, $\xi \in T^*_{q_0}M$, completely determines the shape of the first caustic and the cut locus in a neighbourhood of q_0 not just qualitatively, but also quantitatively. However, we must admit that the mechanisms of this relationship are poorly understood at present: we have no general results deducing the properties of the caustic and the cut locus from the structure of the curvature form.

So far we have considered the curvature at a fixed point $q_0 \in M$, but since this can be an arbitrary point on the manifold, we have thus found the curvature and its normal form at all points. We obtain the smooth functions $q \mapsto \kappa(q)$ and $q \mapsto \chi(q)$ along with the vector field $q \mapsto f_0(q)$. If $\chi(q_0) = 0$, then the shape of the first caustic and the cut locus differ strongly from our description above, and they can be fairly accurately recovered from the derivatives of χ at q_0. On the other hand, if $\chi(q) = 0$ for each $q \in M$, then everything is determined by the derivatives of κ. If $\chi \equiv 0$ and $\kappa \equiv \text{const}$, then near q_0 the first caustic merges with the cut locus and coincides with a trajectory of the field f_0.

As a matter of fact, the class of contact sub-Riemannian structures satisfying the condition $\chi \equiv 0$ has a very transparent interpretation. This is equivalent to the condition that the flow generated by f_0 consists of sub-Riemannian isometries. By straightening the field f_0 in a neighbourhood of q_0 we can represent this neighbourhood as a product of a straight-line interval (a piece of a trajectory of f_0) and a 2-dimensional domain. The distribution Δ is transversal to trajectories of f_0, and the sub-Riemannian length of vectors in Δ depends only on their projections on the 2-dimensional domain, because a shift along the trajectories of f_0 is an isometry. Thus, the sub-Riemannian length of horizontal vectors defines a certain Riemannian structure on the 2-dimensional domain, turning it into a Riemann surface. Furthermore, the function κ, which is a sub-Riemannian invariant, is constant on trajectories of f_0 and in essence is a function on this Riemann surface.

It turns out that κ is none other than the Gaussian curvature of the Riemann surface, and the original sub-Riemannian structure is locally isometric to the structure corresponding to Dido's problem on the Riemann surface. We recall that Dido's problem is a particular kind of isoperimetric problem on a Riemann surface: in it the exterior differential of the 1-form defining the isoperimetric constraint is the area form on the Riemann surface. We recall also that the Heisenberg group corresponds to Dido's problem on the Euclidean plane, that is, to the case when $\chi(\,\cdot\,) \equiv \kappa(\,\cdot\,) \equiv 0$.

We see that a contact sub-Riemannian structure is locally isometric to its metric tangent space if and only if $\chi = \kappa = 0$ everywhere, or equivalently, the quadratic form $\xi \mapsto r(\xi)$ with $\xi \in T^*_{q_0}M$ has rank 1 for all q_0. We have already explained that this form cannot vanish.

Comments. The sub-Riemannian curvature was discovered quite recently (see [4], [5]). Its definition resembles Riemann's original construction of the curvature named after him [24]. Theorems 7 and 8 and Proposition 5 were proved in [4]. The typical 3-dimensional contact structures were thoroughly investigated in [1], [13], and [7], well before the discovery of sub-Riemannian curvature.

Bibliography

[1] A. A. Agrachev, "Exponential mappings for contact sub-Riemannian structures", *J. Dyn. Control Syst.* **2**:3 (1996), 321–358.

[2] A. A. Agrachev, "Tangent hyperplanes to subriemannian balls", *J. Dyn. Control Syst.* **22**:4 (2016), 683–692.

[3] A. Agrachev, D. Barilari, and U. Boscain, *Introduction to Riemannian and sub-Riemannian geometry*, Preprint SISSA 09/2012/M, 2016 (v1 – 2012), 457 pp., https://webusers.imj-prg.fr/~davide.barilari/2016-11-21-ABB.pdf.

[4] A. Agrachev, B. Barilari, and L. Rizzi, *Curvature: a variational approach*, Mem. Amer. Math. Soc., Amer. Math. Soc., Providence, RI (to appear); 2015 (v1 – 2013), 120 pp., arXiv: 1306.5318.

[5] A. Agrachev, B. Barilari, and L. Rizzi, "Sub-Riemannian curvature in contact geometry", *J. Geom. Anal.* **27**:1 (2017), 366–408; 2016 (v1 – 2015), 31 pp., arXiv: 1505.04374.

[6] A. Agrachev, B. Bonnard, M. Chyba, and I. Kupka, "Sub-Riemannian sphere in Martinet flat case", *ESAIM Control Optim. Calc. Var.* **2** (1997), 377–448.

[7] A. A. Agrachev, El-A. El-H. Chakir, and J. P. Gauthier, "Sub-Riemannian metrics on $\mathbf{R}^3$", *Geometric control and non-holonomic mechanics* (Mexico City, 1996), CMS Conf. Proc., vol. 25, Amer. Math. Soc., Providence, RI 1998, pp. 29–78.

[8] A. A. Agrachev, R. V. Gamkrelidze, and A. V. Sarychev, "Local invariants of smooth control systems", *Acta Appl. Math.* **14**:3 (1989), 191–237.

[9] A. Bellaïche, “The tangent space in sub-Riemannian geometry”, *Sub-Riemannian geometry*, Progr. Math., vol. 144, Birkhäuser, Basel 1996, pp. 1–78.

[10] И. Ю. Бесчастный, “Об оптимальном качении сферы с прокручиванием, без проскальзывания”, *Матем. сб.* **205**:2 (2014), 3–38; English transl., I. Yu. Beschastnyi, “The optimal rolling of a sphere, with twisting but without slipping”, *Sb. Math.* **205**:2 (2014), 157–191.

[11] R. M. Bianchini and G. Stefani, “Graded approximations and controllability along a trajectory”, *SIAM J. Control Optim.* **28**:4 (1990), 903–924.

[12] P. Cannarsa and L. Rifford, “Semiconcavity results for optimal control problems admitting no singular minimizing controls”, *Ann. Inst. H. Poincaré Anal. Non Linéaire* **25**:4 (2008), 773–802.

[13] El-A. El-H. Chakir, J.-P. Gauthier, and I. Kupka, “Small sub-Riemannian balls on $\mathbf{R}^3$”, *J. Dyn. Control Syst.* **2**:3 (1996), 359–421.

[14] Y. Chitour, F. Jean, and E. Trélat, “Genericity results for singular curves”, *J. Differential Geom.* **73**:1 (2006), 45–73.

[15] W.-L. Chow, “Über Systeme von linearen partiellen Differentialgleichungen erster Ordnung”, *Math. Ann.* **117** (1939), 98–105.

[16] L. Eulero, *Methodus inveniendi lineas curvas maximi minimive proprietate gaudentes, sive solutio problematis isoperimetrici lattissimo sensu accepti*, M. M. Bousquet, Lausanne–Geneva 1744, 322 pp.

[17] M. Gromov, “Carnot–Carathéodory spaces seen from within”, *Sub-Riemannian geometry*, Progr. Math., vol. 144, Birkhäuser, Basel 1996, pp. 79–323.

[18] V. Jurdjevic, *Geometric control theory*, Cambridge Stud. Adv. Math., vol. 52, Cambridge Univ. Press, Cambridge 1997, xviii+492 pp.

[19] J. Mitchell, “On Carnot–Carathéodory metrics”, *J. Differential Geom.* **21**:1 (1985), 35–45.

[20] R. Montgomery, *A tour of subriemannian geometries, their geodesics and applications*, Math. Surveys Monogr., vol. 91, Amer. Math. Soc., Providence, RI 2002, xx+259 pp.

[21] T. Nagano, “Linear differential systems with singularities and application to transitive Lie algebras”, *J. Math. Soc. Japan* **18**:4 (1966), 398–404.

[22] Л. С. Понтрягин, В. Г. Болтянский, Р. В. Гамкрелидзе, Е. Ф. Мищенко, *Математическая теория оптимальных процессов*, Физматгиз, М. 1961, 391 с.; English transl., L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The mathematical theory of optimal processes*, Interscience Publishers John Wiley & Sons, Inc., New York–London 1962, viii+360 pp.

[23] П. Рашевский, “О соединимости любых двух точек вполне неголономного пространства допустимой линией”, *Уч. зап. Моск. гос. пед. ин-та им. К. Либкнехта. Сер. физ.-матем.* **3**:2 (1938), 83–94. [P. Rashevsky, “Connecting two arbitrary point in a totally non-holonomic space by an admissible line”, *Uchenye Zap. Mosk. Gos. Pedagog. Inst. Ser. Fiz.-Mat.* **3**:2 (1938), 83–94.]

[24] B. Riemann, “Ueber die Hypothesen, welche der Geometrie zu Grunde liegen”, *Abhandlungen der Königlichen Gesellschaft der Wissenschaften zu Göttingen* **13** (1868), 133–150; English transl., B. Riemann, *On the hypotheses which lie at the bases of geometry*, Classic Texts in the Sciences (J. Jost, ed.), Birkhäuser/Springer, Basel 2016, 172 pp.

[25] L. Rifford, “À propos des sphères sous-riemanniennes”, *Bull. Belg. Math. Soc. Simon Stevin* **13**:3 (2006), 521–526.

[26] Ю. Л. Сачков, “Симметрии и страты Максвелла в задаче об оптимальном качении сферы по плоскости”, *Матем. сб.* **201**:7 (2010), 99–120; English transl.,

Yu. L. Sachkov, “Maxwell strata and symmetries in the problem of optimal rolling of a sphere over a plane”, *Sb. Math.* **201**:7 (2010), 1029–1051.

[27] H. J. Sussmann, “Orbits of families of vector fields and integrability of distributions”, *Trans. Amer. Math. Soc.* **180** (1973), 172–188.

[28] А. М. Вершик, В. Я. Гершкович, “Неголономные динамические системы. Геометрия распределений и вариационные задачи”, *Динамические системы* – 7, Итоги науки и техн. Сер. Соврем. пробл. матем. Фундам. направления, **16**, ВИНИТИ, М. 1987, с. 5–85; English transl., A. M. Vershik and V. Ya. Gershkovich, “Nonholonomic dynamical systems, geometry of distributions and variational problems”, *Dynamical systems*. VII, Encyclopaedia Math. Sci., vol. 16, Springer, Berlin 1994, pp. 1–81.

Andrei A. Agrachev
Steklov Mathematical Institute
of Russian Academy of Sciences;
International School for Advanced
Studies (SISSA), Trieste, Italy
E-mail: agrachev@mi.ras.ru, agrachev@sissa.it

Received 23/SEP/16
Translated by N. KRUZHILIN