BOOK REVIEWS

BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 57, Number 4, October 2020, Pages 657–677
https://doi.org/10.1090/bull/1701
Article electronically published on June 5, 2020

A comprehensive introduction to sub-Riemannian geometry. From the Hamiltonian viewpoint, by Andrei Agrachev, Davide Barilari, and Ugo Boscain, with an Appendix by Igor Zelenko, Cambridge Studies in Advanced Mathematics, Vol. 181, Cambridge University Press, 2020, xviii+745 pp., ISBN 978-1-108-47635-5

1. Cats and magnetic fields

I fell into sub-Riemannian geometry through the problem of the falling cat [26], [27]. Dropped from upside down with no angular momentum, the cat changes her shape so as to land right-side up. How? Is there an optimal sequence of shape changes she can choose to follow in order to right herself? This last question is a particular case of the problem of finding a sub-Riemannian geodesic.

Open Problem 1. Is every sub-Riemannian geodesic smooth?
This problem remains open despite 30 years of effort.

The dominant feature of living in a sub-Riemannian geometry is that the space of directions in which you can move is restricted. A sub-Riemannian structure is a type of differential-geometric structure, we can put on a smooth manifold Q, which we will define carefully in due time. But for now, its dominant feature is that the space of directions $D(q)$ you can move in, starting from $q \in Q$, forms a k-dimensional subspace $D(q) \subset T_qQ$ of the n-dimensional tangent space at q. These k-planes vary smoothly with q. **We call such a field of k-planes on an n-manifold a distribution.**

For our falling cat, $k = n - 3$ and the k-plane fields $D(q)$ are given by the condition that the total angular momentum $J \in \mathbb{R}^3$ is zero. Recall that the total angular momentum of a configuration of N particles $q_1, \dots, q_N \in \mathbb{R}^3$ travelling with velocities $v_1, \dots, v_N$ is the sum $J(q, v) = \sum_a m_a q_a \times v_a \in \mathbb{R}^3$, where the m_a are the particle masses. For the cat, the q_a can be thought of as representative marker points on the cat—foot, tail, hips, …, with constraints (rigid rods), bones, connecting them. So $Q \subset (\mathbb{R}^3)^N$ represents the cat's *configuration space*—the cat's shape, including how that shape is oriented within the inertial space of the surrounding room within which she is falling. For each configuration $q = (q_1, \dots, q_N) \in Q$ of the cat, the space of physically allowable velocites $v = (v_1, \dots, v_N)$ must lie in the codimension 3 subspace $D(q) = \{v : J(q, v) = 0\}$ of the space of all possible velocities.

For the simplest nontrivial example imagine a two-plane field on $\mathbb{R}^3$; see Figure 1.1. So, at each point $q = (x, y, z) \in \mathbb{R}^3$, we attach a two-plane $D(x, y, z)$ and

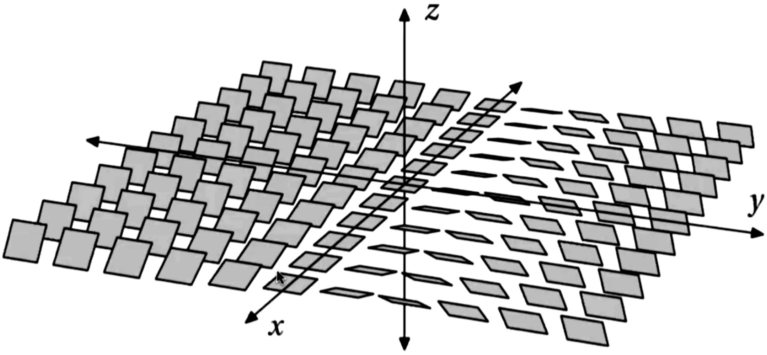

FIGURE 1.1. A field of two-planes ($A_1 = y, A_2 = 0$ in (1.1)).

these planes vary smoothly with q. We can only move away from q along curves tangent to the two-plane $D(q)$. For the cat, the vectors tangent to D represent the physically allowable infinitesimal changes of the cat's configuration. If our two-plane field never goes vertical, meaning it never contains the vertical direction $(0, 0, 1)$, and is invariant under translation in the z-direction, then we can express it as the vanishing of a one-form,

$$\theta := dz - A_1(x, y)dx - A_2(x, y)dy,$$

which means that $D(x, y, z) = \{(v_1, v_2, v_3) : v_3 - A_1(x, y)v_1 - A_2(x, y)v_2 = 0\}$. The A_i are smooth functions of x and y describing the orientation of the two-planes. A basis for $D(x, y, z)$ is

$$X(x, y, z) = (1, 0, A_1(x, y)), \qquad Y(x, y, z) = (0, 1, A_2(x, y)). \tag{1.1}$$

We call a smooth path $q(t) = (x(t), y(t), z(t))$ *horizontal* (for D) if it is everywhere tangent to D, i.e., $\dot{q}(t) \in D(q(t))$ where the dot denotes time derivative.

Question 1. Starting from the origin $q_0 = (0, 0, 0)$, can we get to any point $q_1 = (x_1, y_1, z_1)$ of space by travelling along a horizontal path $q(t)$?
Using our moving frame X, Y, we see that $q(t)$ is horizontal if and only if

$$\dot{q} = u_1(t)X(q(t)) + u_2(t)Y(q(t)) \tag{1.2}$$

for some smooth functions u_1, u_2. Written in coordinates, this ordinary differential equation (ODE) for horizontality reads

$$\begin{aligned} \dot{x} &= u_1, \\ \dot{y} &= u_2, \\ \dot{z} &= u_1(t)A_1(x, y) + u_2(t)A_2(x, y). \end{aligned} \tag{1.3}$$

Our question asks us to find a *control strategy* $(u_1(t), u_2(t))$ which *steers* us from the origin q_0 to the given point $q_1 = (x_1, y_1, z_1)$ in space in some fixed time, say $t = 1$. As a first pass at a solution, take constant controls $u_1 \equiv x_1, u_2 \equiv y_1$. The resulting path has the form $q(t) = (tx_1, ty_1, z(t))$ and joins the origin to $\hat{q}_1 = (x_1, y_1, Z)$ in time 1. The final height reached is $Z = z(1) = \int_\ell A_1 dx + A_2 dy$, with the integral being over the line segment ℓ connecting $(0, 0)$ to (x_1, y_1) in time 1. We would be dumb lucky if $Z = z_1$. How do we change this final height without changing the final xy endpoint? We fiddle around at (x_1, y_1) by drawing little planar loops $c(t) = (x(t), y(t))$ based at (x_1, y_1) and using the induced controls $(u_1(t), u_2(t)) = (\dot{x}(t), \dot{y}(t))$. The resulting change in z is

$$\Delta z = \int_c A_1 dx + A_2 dy = \int\int_D B(x,y) dx \wedge dy, \tag{1.4}$$

where

$$B(x,y) = -\frac{\partial A_1}{\partial y} + \frac{\partial A_2}{\partial x}, \tag{1.5}$$

where D is the disc bounded by c, and we have used Stokes' theorem. It follows that if $B(x_1, y_1) \neq 0$, then we can increase or decrease the height arbitrarily by using an appropriate loop c, winding around it clockwise or counterclockwise many times if neccessary to go up or down, to get $\Delta z = z_1 - Z$ and ending up at the desired q_1.

The hypothesis $B(x_1, y_1) \neq 0$ is the *opposite* of the hypothesis of the classic integrability theorem of Fröbenius. The hypothesis of Fröbenius is that D is closed under the Lie bracket: the Lie bracket $[X, Y]$ of our frame X, Y for D again lies in D so that $[X, Y] = fX + gY$ for smooth functions f, g. The theorem then asserts that the answer to our question is a strong **no**: the set of all endpoints of horizontal paths leaving the origin forms a smooth two-dimensional surface whose tangent space at any point is the two-plane D at that point. For our frame (see (1.1)) we have

$$[X, Y] = (0, 0, B(x,y))$$

so that the Fröbenius integrability condition is indeed $B \equiv 0$ everywhere. (A variation of the argument in the previous paragraph shows that if $B(x_1, y_1) \neq 0$ *somewhere*, then we can get from the origin to any point of $\mathbb{R}^3$ by a horizontal path.)

To define a sub-Riemannian geometry on $\mathbb{R}^3$ with underlying distribution D, declare the vector fields X and Y to be orthonormal. Then the *length* of a vector $v = u_1 X(q) + u_2 Y(q) \in D(q)$ is $\sqrt{u_1^2 + u_2^2}$, and so the *length* of a horizontal path $q(t) = (x(t), y(t), z(t))$ is the usual Euclidean length $\int_c \sqrt{\dot{x}^2 + \dot{y}^2} dt$ of its horizontal projection $c(t) = (x(t), y(t))$. **The sub-Riemannian geodesic problem is to find, among all horizontal paths connecting two fixed points q_0 to q_1, the one whose length is minimal.**

We can now define a sub-Riemannian geometry. A sub-Riemannian geometry on a manifold Q is a distribution D, meaning a family of k-planes $D(q) \subset T_q Q$ varying smoothly with $q \in Q$, together with a smoothly varying family of inner products on these k-planes. We can then measure the lengths $\ell(\gamma)$ of horizontal paths γ and define a distance function on Q by

$$d_{sR}(q_0, q_1) = \inf\{\ell(\gamma) : \gamma \text{ a horizontal curve joining } q_0 \text{ to } q_1\}. \tag{1.6}$$

1.1. **Magnetic distributions, continued.** Returning to our sub-Riemannian geodesic problem on $\mathbb{R}^3$, we can use the method of Lagrange multipliers to solve it. The horizontal constraint $\dot{q} \in D(q)$ reads $\dot{z}(t) - A_1 \dot{x}(t) - A_2 \dot{y}(t) = 0$, valid for each t. We introduce a continuous family $\lambda(t)$ of Lagrange multipliers to enforce this continuum of constraints. Use the Cauchy–Schwartz inequality to see that minimizing the *energy* $\int \frac{1}{2}(\dot{x}^2 + \dot{y}^2) dt$ is equivalent to minimizing the length $\ell(c) = \int \sqrt{\dot{x}^2 + \dot{y}^2} dt$ *and* parameterizing $c(t)$ by a constant multiple of arclength.

Thus our Lagrange multiplier principle problem becomes to extremize the following:

$$\int \frac{1}{2}(\dot{x}^2 + \dot{y}^2) + \lambda(t)(\dot{z} - A_1(x,y)\dot{x} - A_2(x,y)\dot{y})dt.$$

Since no z occurs explicitly inside the integral, the Euler–Lagrange equations for z read

$$\dot{\lambda} = 0,$$

so that $\lambda(t) = \mathit{const}$. Interpret this constant as *charge* and realize that $\int \dot{z}dt = z_1 - z_0$ is constant once we fix the endpoints, so that we can throw out $\lambda(t)\dot{z}(t)$ from the integrand and arrive at precisely the Lagrangian for a particle of charge λ travelling in the plane under the influence of a *scalar* magnetic field of strength $B(x,y)$ *orthogonal to the plane.* The one-form $\alpha = A_1dx + A_2dy$ with $d\alpha = Bdx \wedge dy$ is the *vector potential* for this magnetic field. The remaining Euler–Lagrange equations read

$$\begin{aligned} \ddot{x} &= \lambda B(x,y)\dot{y}, \\ \ddot{y} &= -\lambda B(x,y)\dot{x}, \\ \dot{z} &= A_1(x,y)\dot{x} + A_2(x,y)\dot{y}. \end{aligned} \tag{1.7}$$

The first two equations are known as the Lorentz equations. The last equation is called the *horizontal lift equation*: given any smooth curve $c(t) = (x(t), y(t))$ in the x, y plane and any initial value of z_0, by solving this equation we obtain the unique horizontal curve $(x(t), y(t), z(t))$ whose projection is $c(t)$ and which passes through $(x(0), y(0), z_0)$ at time $t = 0$.

The most studied case of the Lorentz equations is the case of a constant magnetic field, say $B = 1$. The corresponding sub-Riemannian geometry is called the *Heisenberg group.* The projections of its geodesics are circles of radius $1/|\lambda|$, when $\lambda \neq 0$, and lines when the charge $\lambda = 0$. *Why Heisenberg?* Well, when $B \equiv 1$, then $[X,Y] = Z := (0,0,1)$ and $[X,Z] = [Y,Z] = 0$. These bracket relations on X, Y, Z are those of the Heisenberg algebra, famous from Heisenberg's commutation relations. Take vector potential $\alpha(x,y) = A_1(x,y)dx + A_2(x,y)dy = \frac{1}{2}(xdy - ydx)$ associated to $Bdx \wedge dy = dx \wedge dy$. Then define a group law on $\mathbb{R}^3$ according to

$$(x,y,z)(x',y',z') = (x+x', y+y', z+z') + (0,0,\frac{1}{2}(xy' - x'y)). \tag{1.8}$$

(Note this last component is $\alpha(x,y)(x',y')$.) With this group law $\mathbb{R}^3$ becomes the non-Abelian group known as the Heisenberg group and is denoted here by $\mathbb{H}$. X, Y, Z form a basis for the left-invariant vector fields on $\mathbb{H}$. The distribution D and its inner product are invariant under left translations by $\mathbb{H}$.

Why "sub" and why "Riemannian"? What does all this have to do with Riemannian geometry? In Riemannian geometry we can move in any direction leaving any point. The length squared of a vector $v \in T_qQ$ is a nondegenerate quadratic form on the tangent space, written in coordinates $ds^2(v) = \Sigma g_{ij}(q)v^iv^j$. We can alternatively diagonalize the metric and write it as a weighted sum of squares of n covectors at a point—one-forms at q. Recall that our distribution D on $\mathbb{R}^3$ was defined by the vanishing of the one-form $\theta = dz - A_1dx - A_2dy$. We can describe a family of Riemannian metrics compatible with our sub-Riemannian one by setting

$$ds^2_\epsilon = dx^2 + dy^2 + \frac{1}{\epsilon^2}\theta^2. \tag{1.9}$$

In this way X, Y, Z are orthogonal to each other, Z's length is $1/\epsilon$, while X and Y remain of length 1. By letting $\epsilon \to 0$, these Riemannian metrics converge to the sub-Riemannian one just described above. A path $q(t)$ moving in such a Riemannian metric gets infinitely penalized for having any vertical component (i.e., $\theta(\dot{q}) \neq 0$) as $\epsilon \to 0$. As a metric space, and as a *length space* [7] the one-parameter family of Riemannian metrics ds^2_ϵ converges to our sub-Riemannian metric.

The limit of (1.9) as $\epsilon \to 0$ is singular. If instead we take the *inverse metric* and let $\epsilon \to 0$, we get something nice. In coordinates the inverse metric, or *cometric*, associated to the metric $ds^2(v) = \Sigma g_{ij}(q)v^i v^j$ is the q-dependent quadratic form $(ds^2_\epsilon)^{-1}(p) = \Sigma g^{ij}(q)p^i p^j$ on the cotangent bundle of Q, where $g^{ij}(q)$ is the inverse matrix to $g_{ij}(q)$ and where the p_i are the coordinates of a covector $p \in T^*_q Q$. Now the basis dx, dy, θ for $T^*\mathbb{R}^3$ used in (1.9) is the dual basis to our basis X, Y, Z for $T\mathbb{R}^3$ from which it follows that

$$(ds^2_\epsilon)^{-1} = X^2 + Y^2 + \epsilon Z^2 \to X^2 + Y^2 \quad \text{as } \epsilon \to 0. \tag{1.10}$$

The limit is now a nice, well-defined gadget: a quadratic form on each $T^*_q Q$ whose rank is everywhere 2.

If we divide the Riemannian cometric by 2 (see (1.10)), we get *kinetic energy*, a Hamiltonian $H(q,p) = \frac{1}{2}\Sigma g^{ij}(q)p^i p^j$, $H : T^*Q \to \mathbb{R}$ whose Hamilton's equations are well known to be a rewriting of the geodesic equations on the Riemannian manifold. In coordinates, Hamilton's equations read $\dot{q}^i = \frac{\partial H}{\partial p_i}$, $\dot{p}_i = -\frac{\partial H}{\partial q^i}$. The projection $q(t)$ of any solution $(q(t), p(t))$ to these ODEs is a Riemannian geodesic, and all Riemannian geodesics arise in this way. It stands to reason that if we view the limiting expression above as a Hamiltonian, and divide it by 2 (not so important), then this *sub-Riemannian kinetic energy* Hamitonian will govern sub-Riemannian geodesics. *We call the resulting Hamiltonian ODEs on T^*Q the normal sub-Riemannian geodesic equations.*

It is standard in differential geometry to write

$$X(x,y,z) = \frac{\partial}{\partial x} + A_1(x,y)\frac{\partial}{\partial z}, \qquad Y(x,y,z) = \frac{\partial}{\partial y} + A_2(x,y)\frac{\partial}{\partial z}, \tag{1.11}$$

instead of using the coordinate notation (see (1.1)) which we used above for our basis for D. The two notations means the same thing. But if we substitute (1.11) into (1.10), what we get seems to be a second-order linear differential operator. (To include Z, use $Z = (0,0,1) = \frac{\partial}{\partial z}$.) But this operator is not what we mean. We want a quadratic form. To get a quadratic form, note that there are at least three distinct ways to think of a vector field on a manifold: as a first-order linear differential operator (see (1.11)), as the right-hand sides of a system of first-order differential equations (see (1.2)), or as a fiber-linear function on the cotangent bundle of the manifold. We are thinking of X and Y in this last way when we view the limit (1.10) as being twice the sub-Riemannian kinetic energy. In order to ensure that we think of vector fields this way, it is helpful to use a different notation. Let

$$P_X = p_x + A_1(x,y)p_z, \qquad P_Y = p_y + A_2(x,y)p_z$$

denote our vector fields X and Y, viewed as linear functions acting on covectors $p = p_x dx + p_y dy + p_z dz \in T^*\mathbb{R}^3$. Our sub-Riemannian kinetic energy Hamiltonian is then

$$H = \frac{1}{2}(P_X^2 + P_Y^2). \tag{1.12}$$

Since no z occurs in H, Hamilton's equations assert that $\dot{p}_z = 0$. Again, interpret this constant p_z as an electric charge, λ from before. Now look in most any classical mechanics or electromagnetism text to see that (1.12) is the Hamiltonian for a particle moving in the plane under the influence of the magnetic field B, i.e., the Hamiltonian generating equations (1.7). As in Riemannian geometry, the projections to $\mathbb{R}^3$ of the solutions to this Hamiltonian differential equation are geodesics. But unlike Riemannian geometry, there may exist other geodesics, not governed by these differential equations. Their existence is why Problem 1 is still open. We return to this point near the end of the review.

When we instead interpret the limit (1.9) as a linear second-order differential operator, we get

$$\Delta = X^2 + Y^2, \tag{1.13}$$

with the vector fields X, Y understood as linear first-order diferential operators as per equation (1.11), then it is the *sub-Laplacian* of our sub-Riemannian structure. The restriction Δ_λ of Δ to functions of the form $\psi(x, y, z) = e^{i\lambda z}\phi(x, y)$ with ϕ square integrable is the Schrödinger operator for a quantum particle of charge λ moving in the xy plane under the influence of our magnetic field $B(x, y)$.

As long as $B(x, y)$ does not vanish to infinite order, then Δ has many properties in common with the usual Laplacian $\frac{\partial}{\partial x}^2 + \frac{\partial}{\partial y}^2 + \frac{\partial}{\partial z}^2$ on $\mathbb{R}^3$. To underline one of these properties, hypoellipticity, consider the partial Laplacian $\Delta_0 = \frac{\partial}{\partial x}^2 + \frac{\partial}{\partial y}^2$ acting on functions on $\mathbb{R}^3$. For f smooth, solutions to $\Delta_0 u = f$ can be as wild as we like. For example, any function $u = u(z)$ of z alone satisfies $\Delta_0 f = 0$. In contradistinction, if $B \neq 0$, then the distributional solution u to $\Delta u = f$ must be smooth, provided f is smooth. Operators enjoying this property are called *hypoelliptic.* The usual Laplacian, being elliptic, is hypoelliptic.

1.2. **Hörmander's condition.** Hörmander [20] showed that second-order linear differential operators of the form

$$\Delta = \sum_{a=1}^{k} X_a^2 \quad \text{on } Q, \qquad \dim(Q) = n > k, \tag{1.14}$$

are hypoelliptic, *provided the collection of vector fields* $\{X_a\}$ are *bracket-generating.* Bracket-generating means that the X_a together with all their Lie brackets $[X_a, X_b]$, $[X_a, [X_b, X_c]], \ldots$ span the tangent space at each point $q \in Q$.

If the X_a are linearly independent, then they define a sub-Riemannian structure on the manifold: the distribution of k-plane fields D is their span, and the inner product is defined by declaring them to be orthonormal. Δ (plus possible first-order terms) is *a sub-Laplacian* for this structure.

1.3. **Sub-Riemannian kinetic energy: Normal geodesics.** The principal symbol of the sub-Laplacian, divided by two, is the sub-Riemannian kinetic energy

$$H = \frac{1}{2}\sum_{a=1}^{k} P_a^2 : T^*Q \to \mathbb{R}, \tag{1.15}$$

where $P_a(q, p) = p(X_a(q))$ is the vector field X_a viewed as fiber-linear functions $P_a : T^*Q \to \mathbb{R}$. This is a direct generalization of the Heisenberg Hamiltonian in (1.12). H **generates sub-Riemannian geodesics:** the projection $q(t)$ to Q of any solution $(q(t), p(t))$ to H's Hamiltonian equations is a sub-Riemannian geodesic. A sub-Riemannian geodesic is defined to be a curve such that all sufficiently short subarcs of the curve are minimizing geodesics between their endpoints. The geodesics which arise from H are called *normal geodesics.* Unlike Riemannian geometry, it can happen that a sub-Riemannian geometry admits geodesics not governed by H, the *abnormal geodesics.* The existence of these abnormal geodesics lies at the heart of the open problem concerning smoothness of geodesics. More on this near the end of this article.

1.4. **Getting there. Chow and Rashevskii.** Forget about finding a shortest horizontal path. Can we find any horizontal path joining two given points? **Yes,** provided the bracket-generating condition on the X_a holds, where the X_a frame our distribution. Moreover, if the points are close, the path is short so that small sub-Riemannian balls (see (1.6)) are open sets relative to the usual topology on the manifold. This theorem, due to Chow [9] and to Rashevskii [36], can be viewed as *the fundamental theorem for sub-Riemannian geometry.* It preceded Hörmander's theorem by nearly 30 years.

The key idea behind the theorem is summarized by the formula,

$$e^{-tY} e^{-tX} e^{tY} e^{tX} = e^{t^2[X,Y]} + O(t^3),$$

for the commutators of the flows e^{tX}, e^{tY} of two vector fields. The formula says that we can move in the *hard* direction $Z = [X, Y]$ by moving along a small square in the X and Y directions, and that the amount moved in the hard direction is roughly the area of this small square. Iterated appropriately, this idea leads to a proof of the Chow–Rashevskii theorem.

1.5. **Growth vector.** Take two random elements g_1, g_2 of the symmetric group on n letters, and with high probability they will generate the group. Take two random vector fields X, Y on $\mathbb{R}^n$, and generically they will bracket-generate $\mathbb{R}^n$. Thus, leaving the origin and travelling only along linear combinations of the X and Y directions, we can reach any point. How fast? How do the open sets which X and Y generate grow with increasing path length? How does this growth depend on the brackets? The answers are encoded by the growth of the span of the Lie brackets. From

$$X, Y, [X, Y], [X, [X, Y]], [Y, [X, Y]], [X, [X, [X, Y]]], \ldots,$$

compute the dimension d_k of the vector space spanned by all k-fold brackets and n_k of the space spanned by these and all brackets involving k or fewer brackets. Then $n_k = n_{k-1} + d_k$. We start out with $n_1 = 2 = d_1$ for the span of X and Y. There is only one 2-fold bracket, $[X, Y]$, and typically it is linearly independent of X and Y, so $d_2 = 1$ and $n_2 = 3$. Continuing, for our generic X, Y we will get the list of dimensions

$$\vec{d} = (2, 1, 2, 3, 6, 9, 18, 30, 56, \ldots)$$

and

$$\vec{n} = (2, 3, 5, 8, 14, 23, 41, 71, 127, \ldots).$$

The first list of numbers for our generic case is the graded dimension vector of the *free Lie algebra* on two generators ([17]). The second list is called the *growth vector* of the distribution spanned by X and Y. We truncate the lists when we run out of room due to the fact that our ambient space has finite dimension n. For example, if $n = 100$, then the last component of the growth vector must be $n_9 = 100$ rather than $n_9 = 127$, and correspondingly $d_9 = 29 = 100 - 71$ rather than $d_9 = 56$, with the earlier d_i and n_i staying the same.

We just wrote down the maximal possible growth vector for a rank 2 distribution—this being the growth vector of a generic distribution of rank 2 above. There are distributions having slower growth. At the other extreme of slow growth, there is a distribution whose growth vector is $(2, 3, 4, 5, \ldots, n-1, n)$. The length of the growth vector is called the *step* of the distribution at the point in question: it equals the total number of Lie brackets required to generate the tangent space at the point in question. The growth vector may depend on the point. For example, if $A_1 = 0$ and $A_2 = y^2$ above, then along the plane $y = 0$ the growth vector is $(2, 2, 3)$, while off this plane the growth is $(2, 3)$. A point at which the growth vector is locally constant is called a *regular* point for the distribution.

The growth vector game works for distributions of any rank k. Next, we form the graded vector space of total dimension n and graded dimensions $\vec{d}$, given by

$$\mathfrak{g}(q_0) = V_1 \oplus V_2 \oplus \cdots \oplus V_s, d_i = n_i - n_{i-1} = \dim(V_i) \tag{1.16}$$

with $V_1 = D(q_0), V_2 = D^2(q_0)/D(q_0), \ldots, V_i = D^i(q_0)/D^{i-1}(q_0)$, where $D^i(q_0)$ is the vector space obtained by evaluating all iterated Lie brackets up to step i at the point q_0. If q_0 is a regular point for D, then $\mathfrak{g}(q_0)$ inherits a Lie bracket from the Lie bracket of vector fields. This Lie bracket respects the grading,

$$[V_i, V_j] \subset V_{i+j} \tag{1.17}$$

and is nilpotent: all $(s+1)$-fold brackets are zero ($V_k = 0$ for $k \geq s+1$). Let $\mathbb{G}(q_0)$ be the simply connected Lie group having Lie algebra $\mathfrak{g}(q_0)$. We call this group the nilpotent approximation, or Carnot group, at q_0.

Definition 1.1. A Carnot group is a simply connected finite-dimensional Lie group G whose Lie algebra $\mathfrak{g}$ is graded nilpotent and which is Lie generated by V_1.

Since $\mathfrak{g}$ is nilpotent, the exponential map $\mathfrak{g} \to G$ is a global diffeomorphism, and in these exponential coordinates, group multiplication becomes a polynomial map $\mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$. Thus we can view the Carnot group G as being the Lie algebra $\mathfrak{g}$, endowed with a non-Abelian polynomial multiplication law. Bellaïche [2], Folland, and others have suggested thinking of Carnot groups as *non-Abelian vector spaces.*

Example 1.2. When D has growth vector $(2, 3)$, then its Carnot group is isomorphic to the Heisenberg group $\mathbb{H}$. Such a distribution is called a three-dimensional *contact distribution.*

Scalar multiplication $v_1 \to \lambda v_1$ on V_1 induces, by Lie bracket, the scaling $v_i \mapsto \lambda^i v_i$ of V_i. Put together, these linear maps form a single linear map $\mathfrak{g} \to \mathfrak{g}$ which is a one-parameter family of Lie algebra automorphisms, $\delta_\lambda : \mathfrak{g} \to \mathfrak{g}$. This automorphism exponentiates to a one-parameter family of automorphisms denoted by the same name, $\delta_\lambda : \mathbb{G} \to \mathbb{G}$.

Left-translating V_1 defines a left-invariant distribution $\hat{D} \subset T\mathbb{G}$. A *norm* on V_1 can be similarly translated, giving $\hat{D}$ a fiber norm and so a notion of length for horizontal paths, and thus, by taking infimums of lengths connecting points, a left-invariant metric $d_{\mathbb{G}} : \mathbb{G} \times \mathbb{G} \to \mathbb{R}$ on $\mathbb{G}$. If $\mathbb{G} = \mathbb{G}(q_0)$ as above, it makes sense to use the given Euclidean norm of $V_1 = D(q_0)$, in which case this structure is sub-Riemannian. When the norm is not neccessarily a Euclidean norm, then this structure is called a *sub-Finsler structure* or sometimes a *Carnot–Carathéodory metric*. Regardless of norm used, δ_λ is a metric dilation,

$$d_{\mathbb{G}}(\delta_\lambda g_1, \delta_\lambda g_2) = \lambda d_{\mathbb{G}}(g_1, g_2).$$

There is a converse to the above construction, summarized by the following satisfying theorem in synthetic metric geometry by Le Donne.

Theorem 1.3 ([22]). *A homogeneous metric space, which is locally compact, geodesic, and which admits a metric dilation δ_λ, $\lambda > 0$ (a single one will do), is a Carnot group endowed with a left-invariant sub-Finsler structure.*

(A metric space is *geodesic* when the distance between its points equals the length of the shortest curve joining them.)

A homogeneous metric space has a well-defined Hausdorff dimension N (possibly infinite). Lie groups have Haar measures μ. In the case of a Carnot group these ideas are linked through $\mu(B(r)) = Cr^N$, where $B(r)$ denotes the ball of radius r about the identity and $C = \mu(B(1))$. We have,

$$N = \sum id_i, \quad d_i = \dim(V_i).$$

This formula can be derived by observing that when using exponential coordinates, the usual Lebesgue measure $d^n x$ on $\mathfrak{g}$ is a Haar measure, and by computing $\delta_\lambda^* d^n x = \lambda^N d^n x$. Since $n = \sum d_i$, we have that the Hausdorff dimension of $\mathbb{G}$ is greater than its topological dimension. For example, the Hausdorff dimension of the Heisenberg group is 4.

1.6. **Summary of section.** We introduced sub-Riemannian geometry by way of the *magnetic examples* on $\mathbb{R}^3$ (see (1.1) and (1.2)) and wrote out their geodesic equations in (1.7). We then went to general sub-Riemannian geometries, introduced the key bracket-generating condition and its consequent theorems, including the fundamental theorem of sub-Riemannian geometry, the getting-there theorem of Chow and Rashevskii. We ended with the notion of the growth vector of a distribution at a point q and how a non-Abelian vector space, the Carnot group $\mathbb{G}(q)$, endowed with its own sub-Riemannian geometry, is attached to each regular point of a sub-Riemannian geometry.

2. Sources of sub-Riemannian geometry

I describe three sources of inspiration for sub-Riemannian geometry: The first (section 2.1, **I**) is the focus of the book under review. The third (section 2.3, **III**) is the focus of the book [8]. The second source (section 2.2, **II**) is touched on in both of these books and my book [30], but it is probably best understood by going back to the original sources.

2.1. **I. Control theory, geodesics, and Hamiltonian dynamics.** Control theory has ubiquitous applications in modern daily life, exemplified, sometimes disastrously, by drones, a 757 taking off on autopilot, and the use of anti-lock brakes. The basics of control theory, as it fits into sub-Riemannian geometry, are perhaps best understood on the first pass by considering the driving and parallel parking of a car. You have two controls: the angle of your steering wheel and your foot's pressure on the accelerator or brake. (See the beautiful section on Lie brackets in Edward Nelson's book [32].) Sub-Riemannian geometries can be viewed as *under-actuated systems*: they have fewer controls than states. The early prophet of control theoretic thinking in sub-Riemannian geometry was Roger Brockett [5] who espoused a view of Heisenberg geometry as a kind of platonic ideal for all motors. In Heisenberg geometry, moving around circles in the plane leads to averaged linear motion upward along the z axis, thus converting cyclic motion to rectilinear motion.

The focus of this source is to understand sub-Riemannian geodesics, normal geodesic flow, conjugate points, cut points, and the structure of sub-Riemannian balls. The methods are Hamiltonian dynamics and optimal control, combined with some differential topology, functional analysis, and singularity theory. Many of the masters of this area—the authors of the book under review included—come from a control theory background, not a Riemannian geometry background.

A continuous time control system on a manifold Q^n is defined by a parameterized family of ODEs which generalizes equation (1.2),

$$\dot{q} = f(q,u); f : Q \times K \to TQ; K \subset \mathbb{R}^k, \tag{2.1}$$

and where we insist that $f(q,u) \in T_q Q$. Continuing with the language following equation (1.2), we speak of a control strategy $u : [0,1] \to K$ steering us from q_0 to q_1 in some given time interval. In sub-Riemannian geometry the control system is linear in the controls,

$$f(q,u) = \sum_{a=1}^{k} u^a(t) X_a(q(t)), \tag{2.2}$$

where the X_a are an orthonormal frame for D.

For optimal control, we introduce a running cost function $L(q,u)$ and ask to minimize $\int L(q(t), u(t))dt$ among all controls that steer from q_0 to q_1. We might fix the time interval $b - a$, or not. For the sub-Riemannian geodesic problem, we take $L = \frac{1}{2}\sum(u^a)^2$. Hamiltonian dynamics enters by way of the maximum principle which provides a neccessary condition for optimality in terms of Hamiltonian systems on T^*Q. The sub-Riemannian kinetic energy Hamiltonian with its consequent normal geodesics is one outcome of this principle.

Control theorists were used to nonsmooth paths: K might be a box or just the vertices of the box. The optimal control may switch from vertex to vertex, jumping discontinuously. This familiarity with discontinuous path phenomenon appearing in optimal processes gave control theorists advantages over Riemannian geometers when it came to understanding sub-Riemannian geodesics. Riemannian geometers kept searching for canonical connections and tangent bundle formulations of the geodesic equations. There are none. And they largely missed the meaning and importance of the abnormal extremals. Most of the big advances in the 1990s and 2000s in sub-Riemannian path geometry came from control theorists using their Maximum Principle. (See the last section of this review).

2.2. **II. Metric limits and Carnot groups in the wild.**

This source we've divided in three. First we present some fundamental internal facts to sub-Riemannian geometry and nilpotent Lie groups. We follow with descriptions of two problems in which Carnot geometries arose organically as limits and ultimately solved the original problems.

2.2.1. *IIc. Tangent cones. Asymptotic cones.* Gromov defined limits of sequences of pointed metric spaces; see [7]. Using this notion he defined the metric tangent cone of a metric space Q at a point q_0 as the limit of the pointed metric spaces

$(\lambda Q, q_0)$ as $\lambda \to \infty$. Here λQ means Q is endowed with the scaled metric λd. He defined the asymptotic cone to the metric space by going in the other direction $\lambda \to 0$. These limits need not exist. If they exist, they need not be unique. Nevertheless:

Theorem 2.1.

(1) *The metric tangent cone to a sub-Riemannian manifold at a regular point q_0 is its nilpotent approximation, the Carnot group $G(q_0)$, equipped with the induced sub-Riemannian structure; see* [25],

(2) *If q is an irregular point of a sub-Riemannian manifold, then there is a Carnot group $\mathbb{G}$ of topological dimension greater than n and a $\mathbb{G}$-homogeneous space $\mathbb{G}/H$ such that the metric tangent cone at q is $\mathbb{G}/H$; see* [2].

(3) *If G is any simply connected nilpotent Lie group endowed with a left-invariant Riemannian metric, then its asymptotic cone is a Carnot group with sub-Riemannian structure induced from one on $V_1 = G/[G,G]$; see* [34].

The first item of the theorem, known as Mitchell's theorem, when applied to Riemannian geometry ($k = n$) asserts that the the metric tangent cone to a Riemannian manifold at a point is its usual tangent space, endowed with the Euclidean metric given to it by the Riemannian metric. Mitchell's theorem asserts that *Carnot geometry is to sub-Riemannian geometry as Euclidean geometry is to Riemannian.*

2.2.2. *IIa. Geometric group theory. Groups of polynomial growth.* Think of the lattice $\mathbb{Z}^2 \subset \mathbb{R}^2$. Connect the dots in the usual way, to form an infinite sheet of graph paper, a tiling composed of unit squares. Shrink the edges by ϵ forming $\epsilon\mathbb{Z}^2$. Then $\epsilon\mathbb{Z}^2 \to \mathbb{R}^2$ as a metric space, if we give $\mathbb{R}^2$ the metric coming from the L_1-norm $\|(x,y)\|_1 = |x| + |y|$. In the language of the preceding paragraph, the asymptotic cone for $\mathbb{Z}^2$ is $\mathbb{R}^2$.

We can repeat this construction with any finitely generated infinite discrete group Γ. The word metric on Γ is defined by selecting some finite number ℓ of generators for Γ, forming the resulting *Cayley graph* and letting each edge have length 1. The vertex set of this graph is Γ. Leaving each vertex are 2ℓ edges labelled by the generators and their inverses, with vertex x joined to vertex y by the edge labelled γ if and only if $y = x\gamma$. Look at the *volume* $V(R)$ of balls of radius R about the identity, i.e., the number of elements of Γ in that ball. If $V(R)$ is bounded by a polynomial in R, then we say the group has polynomial growth, with growth rate being the degree of that polynomial. The problem here is to characterize the groups of polynomial growth.

$\mathbb{Z}^2$ has polynomial growth of degree 2. An Abelian group of rank r has polynomial growth of rate r. Abelian groups are special examples of *nilpotent groups*, groups whose descending series of commutatant subgroups ends in the trivial group. Joe Wolf [40] and Hyman Bass [1] proved that every finitely generated nilpotent group has polynomial growth. Growth rate is unchanged by passing to finite index subgroups. So, if we call Γ *virtually nilpotent* when it admits a finite index normal subgroup which is nilpotent, then these virtually nilpotent groups have polynomial growth.

Question 2. If a group has polynomial growth, is it virtually nilpotent?

Gromov [16] answered yes by showing that if a group Γ has polynomial growth, then $\epsilon\Gamma \to \mathbb{G}$ as $\epsilon \to 0$, where $\mathbb{G}$ is some Carnot group. He needed polynomial growth at a crucial stage to infer that the limiting object ($\mathbb{G}$) had finite Hausdorff dimension. He finished the proof, roughly speaking, by showing that the original Γ *almost* embeds as a lattice in $\mathbb{G}$ and hence is virtually nilpotent.

For me, this is a mind-blowing proof, with its miraculous construction of a continuous Lie group *out of thin air.* By itself, this is enough to show that Carnot groups warrant study.

2.2.3. *IIb. Mostow rigidity and the visual boundary of rank* 1 *symmetric spaces.* First, let us introduce our "Carnot group heros". They will enter this play near the end. We can write the Heisenberg multiplication law (1.8) as

$$(z,t)(w,s) = (z+w, t+s+\frac{1}{2}\mathrm{Im}(\bar{z}w))$$

upon identifying $\mathbb{R}^2$ with $\mathbb{C}$, so that $z = x+iy,\ w = x'+iy' \in \mathbb{C},\ t,s \in \mathbb{R} \cong \mathrm{Im}(\mathbb{C})$. Now, replace z and w with r-vectors $z = (z_1,\dots,z_r), w = (w_1,\dots,w_r)$ and replace $\mathrm{Im}(\bar{z}w)$ by $\mathrm{Im}(\sum \bar{z}_i w_i)$, and we get the standard Heisenberg group structure on $V_1 \oplus V_2 = \mathbb{C}^r \oplus \mathbb{R}$. We write this as $\mathbb{H}^r_{\mathbb{C}}$. Replacing $\mathbb{C}$ by the quaternions $\mathbb{K}$, the same formula holds and yields a Carnot group structue $\mathbb{H}^r_{\mathbb{K}} = \mathbb{K}^r \oplus \mathrm{Im}(\mathbb{K}) = \mathbb{R}^{4r} \oplus \mathbb{R}^3$. There is also an octonionic Heisenberg group of the form $\mathbb{R}^8 \oplus \mathbb{R}^7$, where $\mathbb{R}^8$ has the structure of the octonions and $\mathbb{R}^7$ are the imaginary octonions. We will refer to any of these extended Heisenberg groups as $\mathbb{H}^r_{\mathbb{K}}$ below, letting $\mathbb{K}$ run through the four real division algebras $\mathbb{R}$, $\mathbb{C}$, the quaternions, and the octonions, with the case of $\mathbb{K} = \mathbb{R}$ being just $\mathbb{H}^n_{\mathbb{R}} = \mathbb{R}^n$ since $\mathrm{Im}(\mathbb{R}) = 0$.

Now the play starts. The fundamental group Γ of a compact manifold M^n is a finitely generated discrete group. If M admits a Riemannian metric of (variable) negative curvature, then Γ is infinite and has exponential, rather than polynomial growth. Mostow rigidity asserts that if the metric on M has *constant* negative curvature and if $n > 2$, then the group Γ determines the manifold M^n up to an isometry.

The Cartan–Hadamard theorem asserts that the universal cover of any compact manifold M with negative curvature, whether constant or variable, is the unit open ball B^n in n-space. Γ acts on B^n by deck transformations and the metric lifts to a Γ-invariant metric on B^n. If M has constant negative curvature, then this lifted metric is that of the standard hyperbolic n-space, denoted $\mathbb{R}H^n$, with its rich group of isometries, denoted $G_{\mathbb{R}}$, and Γ embeds in $G_{\mathbb{R}}$ as a lattice. Mostow's proof proceeds by promoting an abstract isomorphism between two such lattices, say Γ and Γ', to a conjugacy: $\Gamma' = g\Gamma g^{-1}$ by some element $g \in G_{\mathbb{R}}$. This g yields the desired isometry.

Standard hyperbolic space $\mathbb{R}H^n$ forms one of four families of negatively curved symmetric spaces—Riemannian metrics on the ball having large interesting isometry groups. We denote the others by $\mathbb{K}H^n$, where $\mathbb{K}$ runs over the other three division algebras. We call them complex, quaternionic, and octonionic hyperbolic spaces. (The octonionic one only exists for $n = 2$.) We will write $G_{\mathbb{K}}$ for their isometry groups. Mostow extended his rigidity theorem to lattices sitting within all the $G_{\mathbb{K}}$'s [31]. Developing and understanding the essence of Mostow's methods demanded new fundamental ideas in sub-Riemannian geometry, ideas developed by Pansu in [35] and Mostow and Margulis in [24].

The metric defining $\mathbb{K}H^n$ lives on the open unit ball $B^{dn} \subset \mathbb{K}^n$, where $d = \dim_{\mathbb{R}} \mathbb{K}$. Every isometry $g \in G_{\mathbb{K}}$ extends to a smooth map of S^{dn-1}, but this extended map is typically *not* an isometry of the bounding sphere. Indeed, the metric on B^{dn} is complete, so it must blow up as we approach the bounding sphere $S^{dn-1} = \partial B^{dn}$, and there is no obvious metric to be preserved. But there is something preserved by the $G_{\mathbb{K}}$ action on the sphere. This something is a conformal sub-Riemannian structure.

In the case of $\mathbb{R}$ this structure preserved by $G_{\mathbb{R}}$ is the standard conformal structure of the sphere, and so $G_{\mathbb{R}}$ identifies with the Möbius group—the group of all conformal automorphisms of the sphere. Now the sphere is conformally flat: stereographic projection maps $\mathbb{R}^n$ onto $S^n \setminus \{pt\}$ and takes the usual flat metric on $\mathbb{R}^n$ to a positive function times the round metric on the sphere, so that we identify S^n with the conformal compactification of $\mathbb{R}^n$. A key fact is that every local conformal transformation of the sphere at infinity arises from an isometry inside, at least when $n > 2$.

Mostow's proof proceeded in the original real case by first promoting the assumed abstract isomorphism $\Gamma \to \Gamma'$ to an equivariant map $B^n \to B^n$ with a weak type of regularity called being a *quasi-isometry.* Next, he had to show that this quasi-isometry induced a map of the bounding sphere of a type called a *quasi-conformal map* of the boundary. The next and perhaps hardest step was to use the lattices to gain extra regularity for his quasi-conformal map and to show that it is actually conformal. Then, by the key fact described in the end of the preceding paragraph, he is done.

For the other $\mathbb{K}$'s the basic structure of this proof holds, with the difference being that the geometry inherited by their sphere at infinity is no longer conformal but conformal sub-Riemannian. The conformal sub-Riemannian structure at infinity is similarly a conformal compactifiction of a *flat* structure, but now that flat structure is that of our model 2-step Heisenberg type Carnot groups $\mathbb{H}^r_{\mathbb{K}}$. There is a group-theoretically induced *stereographic projection map*, $\mathbb{H}^r_{\mathbb{K}} \to S^{dr-1} \setminus \{pt\}$ which maps the sub-Riemannian structure described above on this Carnot group to the sub-Riemannian structure on the $\mathbb{K}$-sphere, up to multiplication of the inner product on the D's by overall postive function. The analogous key fact holds: isometries of $\mathbb{K}H^n$ induce sub-Riemannian conformal automorphisms of the corresponding sphere at infinity, and every such sub-Riemannian conformal automorphism arises in this way.

The quaternionic case and octonionic cases are particularly interesting. Any smooth map preserving these distributions automatically is an element of $G_{\mathbb{K}}$—no conformality condition is needed! This is an old theorem of E. Cartan, going back to his work on classifying groups of finite and infinite Lie type. The hard work comes because the quasi-subconformal map induced through the Mostow procedure need not be smooth. Somehow, Pansu had to manufacture more regularity for these maps. He did so through the notion of the Pansu differential and by developing analysis on Carnot groups.

Pansu's hard work came in obtaining regularity for his quasi-conformal maps. He modified the usual definition of the derivative of a map between real vector spaces by replacing scalar multiplication of vector spaces with the Carnot dilation and arrived at a new type of derivative for maps between Carnot groups, called the Pansu derivative. With the aid of the nilpotent approximation, Pansu's derivative can be promoted to a derivative for maps between sub-Riemannian manifolds, at least at equiregular points.

Theorem 2.2 (Pansu [35]; Mostow and Margulis [24])**.** *A Lipshitz map between sub-Riemannian manifolds is Pansu differentiable almost everywhere. At each point of differentiablitly, the derivative is a homomorphism of Carnot groups.*

Corollary 2.3. *There is no Lipshitz embedding of a sub-Riemannian manifold into a Riemannian one.*

Proof of Corollary. There are no injective homomorphisms from a non-Abelian Carnot group to $\mathbb{R}^m$. □

2.3. **III. Subelliptic PDE and geometric measure theory.**

2.3.1. *Linear PDE.* The fundamental solution $K(x,y) = \Delta^{-1}(x,y)$ for the usual Laplacian $\Delta = \Delta_{\mathbb{R}^n}$ on $\mathbb{R}^n$, is

$$\Delta^{-1}_{\mathbb{R}^N}(x,y) = \frac{C}{\|x-y\|^{n-2}}, \quad n > 2,$$

where $C = C(n)$ is the reciprocal of the $(n-1)$-dimensional *surface area* of the sphere in $\mathbb{R}^n$. Being the *fundamental solution* means $u(x) = K(x,y)$ satisfies $(\Delta_{\mathbb{R}^n} u)(x) = \delta_y(x)$, and so by convolution, inverts the Laplacian. Folland [14] explicitly computed the fundamental solution for the sub-Laplacian (1.14) on the Heisenberg group,

$$\Delta^{-1}(q,q') := \frac{1/8\pi}{\rho^2}, \qquad (x,y,z) = q^{-1}q', \qquad \rho = \{(x^2+y^2)^2 + 16z^2\}^{1/4}.$$

(Folland uses a different framing than ours. To get from his X, Y to ours, set his t equal to our $4z$.) The function ρ is called the Koranyi gauge (see, e.g., [8, p. 16]), and it provides an alternative to the sub-Riemannian distance function $d_{\mathbb{H}}$ that we have been using. This Koranyi gauge is a *norm* on $\mathbb{H}$,

$$\rho(\delta_\lambda g) = \lambda\rho(g), \qquad \rho(gh) \le \rho(g) + \rho(h),$$

and so $d_K(g,g') = \rho(g^{-1}g')$ provides an alternative to the usual horizontal path-based sub-Riemannian distance $d_{\mathbb{H}}$. Any two norms on a Carnot group yield Lipshitz equivalent metrics, so that Folland's fundamental solution satisfies $\Delta^{-1}(q,q') \sim C/d_{\mathbb{H}}(q,q')^{N-2}$, where $N = 4$ is the Hausdorff dimension of $\mathbb{H}$. This estimate generalizes to any Carnot group $\mathbb{G}$, yielding

$$\Delta^{-1}(x,y) \sim \frac{C}{d_{\mathbb{G}}(x,y)^{N-2}},$$

where N is the Hausdorff dimensionn of $\mathbb{G}$ and $d_{\mathbb{G}}$ is the sub-Riemannian distance (see [8, p. 109], [15]).

Folland, Rothschild, Stein, and others developed the nilpotent approximation using Carnot groups well before Mitchell's thesis result, Theorem 2.1(1). They called their groups *stratified groups* and developed them as a tool for creating parametrices (inverses up to a compact operator) for general sub-Laplacians (1.14). Unlike Bellaïche, in Theorem 2.1(2) they did not concern themselves with niceties of taking the *smallest* approximation to a variable sub-Riemannian structure at irregular points, but rather they went all the way and attached the free s-step nilpotent group to each point of the sub-Riemannian manifold, where $s = \max_q s(q)$ is the *global step* of the distribution—the minimal number of brackets needed to Lie generate the tangent space at any point of the manifold. See [15] for a wonderful overview.

The interplay between classical and quantum mechanics has been a thriving sideshow in Riemannian geometry. Geodesic flow represents classical mechanics, and the spectral properties of the Laplacian represents the quantum. In the 1910s Weyl established a relation between the growth rate of the number of eigenvalues and the volume of phase space. Starting in the 1950s with Selberg [38], came a series of *trace formulas* by Colin de Verdiére in [11], Gullemin and Duistermaat in [13], Guzwiller in [18], and others, which established tight relationships between certain sums of lengths of closed geodesics and sums of eigenvalues of the Laplacian. In 1974 Schnirelman [37] kicked off the field of quantum ergodicity, which continues to boom, the effort being centered around finding quantum signals for ergodic geodesic flow. Recently in sub-Riemannian geometry there has been a flurry of effort following these classical-quantum lines [12], with the normal sub-Riemannian geodesic flow representing classical mechanics and the spectral properties of the sub-Laplacian representing quantum mechanics. Lying like a ghost in the background, are the abnormal geodesics. What is their quantum trace? (The article [29] provides strong evidence that there is one.)

2.3.2. *Nonlinear PDE.* Minimal and constant mean curvature surfaces, or *soap bubbles* in Euclidean space, are central subjects for geometric analysis. These surfaces locally minimize area or minimize it subject to the constraint of bounding a domain of fixed volume. The PDE characterizing these surfaces are nonlinear and have garnered an immense body of work.

Analogous surfaces can be defined in sub-Riemannian geometry. Surface area is problematic, even in the Heisenberg group $\mathbb{H}$, since the Hausdorff dimension of a surface in $\mathbb{H}$ is 3 at most points, and it becomes difficult to define near characteristics—points where the distribution is tangent to the surface. Geometric measure theory enters big time. One way to define surface area is to use the notion of Minkowski content obtained by thickening the surface by an amount ϵ, then differentiating the Haar measure of the resulting domain with respect to ϵ.

Perhaps the most basic such problem, the isoperimetric problem, remains unsolved in $\mathbb{H}$. An isoperimetric surface is a compact surface having the property that among all surfaces bounding a fixed volume, it has the least surface area. In $\mathbb{R}^3$ an isoperimetric surface is a sphere. In 1982 Pansu [33] formulated the Heisenberg isoperimetric problem and provided a conjectural answer, now called a *bubble set.* As in Euclidean spaces, the isoperimetric surface leads to the optimal constants in various Sobolev embedding theorems.

Open Problem 2. Solve Pansu's conjecture. Show that the bubble sets solve the $\mathbb{H}$-isoperimetric problem.

To form a bubble set, fix two points on the z-axis, say $(0,0,-z_1)$ to $(0,0,+z_1)$, and form the surface swept out by all minimizing $\mathbb{H}$-geodesics which join them. We have placed the two points in the same manner as the two points we used way back when the magnetic field B was introduced; see equations (1.4) and (1.5). Each minimizing geodesic projects to a circle in the xy plane passing through the origin and enclosing area $\Delta z = 2z_1$. By rotating any one of them about the z-axis, we generate all the others, hence the bubble set is a surface of revolution. It is not a sphere relative to either our original sub-Riemannian metric or any one of the

Koranyi-type distance functions; see, e.g., [8] for a parameterization. The bubble set is C^2-everywhere and fails to be C^3 at the *poles*, that is, the two points we started with.

The book [8] is inspired by this problem and is a good reference. It reports significant progress to date. For example, the conjecture is true if competing surfaces are restricted to be surfaces of revolution, or *convex* in the Euclidean sense.

2.4. **Summary of section.** I described three sources of inspiration for research in sub-Riemannian geometry: optimal control and motion planning, metric geometry and geometric group theory, and subelliptic PDE. A source I skipped arises in trying to understand an area of the visual cortex known as V1. There is overwhelming evidence that V1 is inhabited by cells whose purpose is to detect orientations of *edges*. There is a natural sub-Riemannian structure on the space of contact elements on the plane—meaning pairs (point, line) where the line and point are incident. Problems of finding optimal (perhaps minimal) surfaces which fill in curves of this three-dimensional sub-Riemannian geometry have proved useful to image processing and to the understanding of some optical illusions; see [10].

3. Abnormal geodesics

In equations (1.2) and (2.2) we represented horizontal paths as solutions to

$$\dot{q}(t) = \sum_{a=1}^{k} u^a(t) X_a(q(t)). \tag{3.1}$$

We view the u^a's as control strategies, taking them to be L^2 functions of t. Fixing $q(0) = q_0$ turns equation (3.1) into an initial value problem, thus it coordinatizes the space of all horizontal paths leaving q_0 by the Hilbert space $L^2 = L^2([0,1], \mathbb{R}^k)$ of control strategies u. The *endpoint map*

$$G : L^2 \to Q; G(u) = q(1) \tag{3.2}$$

sends such a path, or control strategy, to its endpoint. G is smooth. G is an open mapping provided D is bracket-generating. Using the endpoint map, the sub-Riemannian geodesic problem becomes a constrained optimization problem on L_2: among all controls u satisfying the constraint $G(u) = q_1$, find the one(s) minimizing the squared L_2-norm $F(u) = \frac{1}{2}\|u\|_2^2$.

We proceed to solve for the optimal u, using the method of Lagrange multipliers. Form the differential, $\lambda_0 dF(u) + \lambda dG(u)$, insisting that $(\lambda_0, \lambda) \neq (0,0)$, set it to zero, and solve for u. Note that $dG(u) : L^2 \to T_{q(1)}Q$ so that we have $\lambda \in T^*_{q(1)}Q$ and $\lambda dG(u) := \lambda \circ dG(u) : L^2 \to \mathbb{R}$. We say that the horizontal path, or its control strategy u, is *regular* for the endpoint map if $dG(u)$ is onto. Otherwise we call the path *singular* for the endpoint map. If a minimizer u is regular for the endpoint map, then the implicit function theorem kicks in. Near u constraint space $\{u : G(u) = q_1\}$ forms a smooth Hilbert manifold whose tangent space is the kernel of $dG(u)$. With a bit of work, we can then show that the corresponding horizontal path $q(t)$ must be the projection of a solution $(q(t), p(t))$ to the normal sub-Riemannian geodesic as defined above by the sub-Riemannian kinetic energy (1.15).

If a minimizer is regular for the endpoint map, then $\lambda_0 \neq 0$. The solutions to the Lagrange multiplier equations, for which $\lambda_0 = 0$, are precisely the singular curves for the endpoint map. Indeed, for such a u we have $\lambda dG(u) = 0$, and the image of $dG(u)$ is contained in the hyperplane $\{\lambda = 0\} \subset T_{q(1)}Q$.

Minimizing geodesics must either be normal or be singular for the endpoint map. (They might be both.) For awhile, it was believed that every sub-Riemannian geodesic was normal. The first example of a minimizer which was not normal arose in a sub-Riemannian geometry of *magnetic type* as described in the beginning of this review. Suppose the magnetic field $B(x, y)$ (see (1.5)) has a nondegenerate zero locus, so that there is a point (x_0, y_0) with $B(x_0, y_0) = 0$ and $dB(x_0, y_0) \neq 0$. Then the locus $\{B = 0\}$ forms a smooth embedded curve passing through (x_0, y_0).

Theorem 3.1 ([28]). *Any sufficiently short subarc of a horizontal lift of the nondegenerate zero locus of a magnetic field is the unique minimizing sub-Riemannian geodesic between its endpoints. If the planar curvature of this arc is nonzero at* (x_0, y_0) *(i.e., if the arc is not contained in a line), then the horizontal lift is not a normal sub-Riemannian geodesic.*

The growth vector for the distribution in the theorem is $(2, 2, 3)$ along the abnormal minimizer. Abnormal minimizers do not exist for two-step distributions—ones having growth vector (k, n)—since for these distributions one can show that every *nonconstant* path is regular for the endpoint map. Any rank 2 distribution in dimension $n \geq 4$ has step 3 or greater. The abnormal minimizer of the theorem above was soon shown to be part of a generic phenomenon. Rank 2 distributions in dimensions $n \geq 4$ admit abnormal minimizers passing through *every* point. Distributions of type $(2, 3, 5, \ldots)$ have abnormal minimizers passing through every point in every horizontal direction; see [23] and [6].

To date, all known abnormal minimizers found have been smooth. But singular curves for the endpoint map, without being subject to minimality, can be as nonsmooth as we wish. Indeed, take a rank 2 D on $\mathbb{R}^3$ as above whose magnetic field $B(x, y)$ has as zero locus a curve as rough as can be, and horizontally lift this curve to get a nonsmooth singular curve. *The existence of abnormal minimizers is why the problem of whether or not sub-Riemannian geodesics are always smooth remains open.*

Perhaps there is a nonsmooth minimizing geodesic joining q_0 to nearby q_1 relative to some sub-Riemannian geometry. We do not know. We would at least like to say that *most* of the geodesics leaving q_0 are smooth. This would be true provided that the answer to the next question was yes.

Open Problem 3. Does Sard's theorem hold for the endpoint map?

In other words, is the set of endpoints of singular paths leaving q_0 a set of measure zero? For the special case of three-manifolds endowed with analytic distributions, this question was very recently answered affirmatively; see [3].

4. The book

I am glad to have this book.

One of the main tools in optimal control is the Maximum Principle, due to Boltjanskiĭ, Gamkrelidze, and Pontrjagin in [4]. This principle is a computationally effective Hamiltonian reformulation of the Lagrange multiplier method. (My favorite reference on the principle is [41].) A direct line connects the book's senior author, Agrachev, to the Maximum Principle: Pontrjagin advised Gamkrelidze who advised Agrachev. Agrachev has built a prolific school of geometric control theory which dominates mathematical control theory in parts of Europe and Russia. The other two authors, Barilari and Boscain, are members of this school.

For decades Agrachev and his school have been developing powerful tools for sub-Riemannian geometry based on methods and thinking coming primarily from geometric control theory, but also having significant input from symplectic geometry, differential topology, and functional analysis. Agrachev has over a hundred papers listed on MathSci Net. He is not the most patient or friendly of writers. Many of his papers I could not penetrate. So I was overjoyed to read this book and find the writing clear. It is wonderful to have a wide swath of the work of this school explained clearly and set down in one place. I am understanding some of the concepts described for the first time. I am grateful to the three authors for their efforts in putting this book together.

They describe and demonstrate several of powerful tools originating from their work that I could not find in any other books. One such tool is Agrachev's *chronological calculus* which is a nonlinear version of what physicists call the path-ordered exponential for expressing the flow of time-dependent vector fields, such as equation (3.1). This calculus is a flexible tool, good for computing variations such as those needed in understanding the derivatives of the endpoint map.

Another set of tools is built around understanding second derivatives in horizontal path space. It is well known that the second derivative of a function on a manifold has no intrinsic meaning except at a critical point. How do you then make sense of the second derivative of a map between manifolds, or a map like the endpoint map, between a function space and a manifold? *Agrachev and his school are masters of unexpected incarnations of the second derivative.* In Riemannian geometry, the second derivative of arclength leads to Jacobi fields along a geodesic, a basic tool. What are the Jacobi fields in sub-Riemannian geometry? How does one compute the second variation in sub-Riemannian geometry? Along an abnormal sub-Riemannian geodesic? How to you make sense of the second derivative of the endpoint map? To answer these questions, the authors use *intrinsic second derivatives*, calculus on the Lagrange Grassmannian, and a topological study of vector-valued quadratic maps. A Jacobi field in their hands is a curve on the Lagrange Grassmannian. In a pleasant and surprising reversal, the authors *define* the Riemann curvature tensor as the derivative of their Jacobi field, and then show that the result of their definition can be interpreted as the usual tensorial Riemann curvature.

The book also showcases gems developed by other researchers. One such gem is the most significant progress to date on the regularity problem, Open Question 1 above. This result is the *no-corners* theorem of Hakavuori and Le Donne [19] which implies that any piecewise C^1 minimizer must in fact be C^1.

The book is based on a decade of lectures in various places by the three authors. It is structured something like a year-long course aimed at a strong master's student who will be working with one of these researchers. The book begins with classical geometry of surfaces in $\mathbb{R}^3$, looked at through a control theorist's lens. Chapter 2 concerns the general theory of vector fields on manifolds, both autonomous and nonautonomous, their flows, Lie brackets, etc. Sub-Riemannian geometries and the problem of understanding their geodesics arrives on page 71. In the next chapter they describe the Maximum Principle with symplectic geometry put to the fore. The chronological calculus enters in a chapter soon after. We get a break to look at various incarnations of integrable sub-Riemannian geodesic flows on Lie groups mixed in. Carnot groups arrive on page 331. Abnormal minimizers take center stage a bit more than half-way through the book, on page 402. The Lagrange Grassmannians takes center stage on page 513, and here I will stop my description with over 200 pages in the book left to go.

At 724 pages the book feels at times more like an encyclopaedia than a *Comprehensive introduction to sub-Riemannian geometry.* The index is lacking, but this lack is partly made up by a quite detailed table of contents. A real strength of the book are the wonderful Bibliographic Notes ending each chapter. I was, however, surprised by a particular missing reference in the Bibliographic Notes after the titled *Left-invariant Hamiltonian systems on Lie groups*, since it left out the work of Gershkovich and Vershik [39] on this subject. The preceding chapter, Curvature in 3D contact sub-Riemannian geometry, uses their curves-in-Lagrange-Grassmanians approach to curvature to derive the same two curvature invariants found in the 1990s by Keener Hughen [21] in his masterful thesis. I prefer Hughen's method, which is based on Cartan's method of equivalence and moving frames, but my preference might be due to familiarity.

For the reader who wants to get more of a feel for sub-Riemannian geometry beyond this review, but who does not want to commit to reading over 700 pages, you might want to browse my old book [30]. But for students and researchers who are in this field already or are certain they want to be in this field, this is probably the right book to buy. It has begun to replace my book as the book of choice in bibilographies of current research papers on sub-Riemannian geometry.

Acknowledgments

I would like to thank Persi Diaconis, Jon Wolfson, Junecue Sue, Pierre Pansu, and Enrico LeDonne for critical readings of drafts of this review. Figure 1.1 courtesy of wiki: `https://en.wikipedia.org/wiki/File:Standard_contact_structure.svg`.

References

[1] H. Bass, *The degree of polynomial growth of finitely generated nilpotent groups*, Proc. London Math. Soc. (3) **25** (1972), 603–614, DOI 10.1112/plms/s3-25.4.603. MR379672

[2] A. Bellaïche, *The tangent space in sub-Riemannian geometry*, J. Math. Sci. (New York) **83** (1997), no. 4, 461–476, DOI 10.1007/BF02589761. Dynamical systems, 3. MR1442527

[3] A. Belotto da Silva, A. Figalli, P. A., and L. L Rifford, *Strong sard conjecture and regularity of singular minimizing geodesics for analytic sub-riemannian structures in dimension 3*, `arXiv:1810.03347` (2018).

[4] V. G. Boltjanskiĭ, *The maximum principle in the theory of optimal processes* (Russian), Dokl. Akad. Nauk SSSR **119** (1958), 1070–1073. MR0120108

[5] R. W. Brockett, *Nonlinear control theory and differential geometry*, Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Warsaw, 1983), PWN, Warsaw, 1984, pp. 1357–1368. MR804784

[6] R. L. Bryant and L. Hsu, *Rigidity of integral curves of rank* 2 *distributions*, Invent. Math. **114** (1993), no. 2, 435–461, DOI 10.1007/BF01232676. MR1240644

[7] D. Burago, Y. Burago, and S. Ivanov, *A course in metric geometry*, Graduate Studies in Mathematics, vol. 33, American Mathematical Society, Providence, RI, 2001. MR1835418

[8] L. Capogna, D. Danielli, S. D. Pauls, and J. T. Tyson, *An introduction to the Heisenberg group and the sub-Riemannian isoperimetric problem*, Progress in Mathematics, vol. 259, Birkhäuser Verlag, Basel, 2007. MR2312336

[9] W.-L. Chow, *Über Systeme von linearen partiellen Differentialgleichungen erster Ordnung* (German), Math. Ann. **117** (1939), 98–105, DOI 10.1007/BF01450011. MR1880

[10] G. Citti and A. Sarti, *A cortical based model of perceptual completion in the roto-translation space*, J. Math. Imaging Vision **24** (2006), no. 3, 307–326, DOI 10.1007/s10851-005-3630-2. MR2235475

[11] Y. Colin de Verdière, *Spectre du laplacien et longueurs des géodésiques périodiques* (French), C. R. Acad. Sci. Paris Sér. A-B **275** (1972), A805–A808. MR313968

[12] Y. Colin de Verdière, L. Hillairet, and E. Trélat, *Spectral asymptotics for sub-Riemannian Laplacians, I: Quantum ergodicity and quantum limits in the 3-dimensional contact case*, Duke Math. J. **167** (2018), no. 1, 109–174, DOI 10.1215/00127094-2017-0037. MR3743700

[13] J. J. Duistermaat and V. W. Guillemin, *The spectrum of positive elliptic operators and periodic geodesics*, Differential geometry (Proc. Sympos. Pure Math., Vol. XXVII, Part 2, Stanford Univ., Stanford, Calif., 1973), Amer. Math. Soc., Providence, R. I., 1975, pp. 205–209. MR0423438

[14] G. B. Folland, *A fundamental solution for a subelliptic operator*, Bull. Amer. Math. Soc. **79** (1973), 373–376, DOI 10.1090/S0002-9904-1973-13171-4. MR315267

[15] G. B. Folland, *Applications of analysis on nilpotent groups to partial differential equations*, Bull. Amer. Math. Soc. **83** (1977), no. 5, 912–930, DOI 10.1090/S0002-9904-1977-14326-7. MR457928

[16] M. Gromov, *Groups of polynomial growth and expanding maps*, Inst. Hautes Études Sci. Publ. Math. **53** (1981), 53–73. MR623534

[17] A. S. 'Groupprops', *Formula for dimension of graded component of free lie algebra*, `https://groupprops.subwiki.org/wiki/Formula_for_dimension_of_graded_component_of_free_Lie_algebra`, 2020 (accessed April 20, 2020).

[18] M. C. Gutzwiller, *The quantization of a classically ergodic system*, Classical quantum models and arithmetic problems, Lecture Notes in Pure and Appl. Math., vol. 92, Dekker, New York, 1984, pp. 287–351. MR756248

[19] E. Hakavuori and E. Le Donne, *Non-minimality of corners in subriemannian geometry*, Invent. Math. **206** (2016), no. 3, 693–704, DOI 10.1007/s00222-016-0661-9. MR3573971

[20] L. Hörmander, *Hypoelliptic second order differential equations*, Acta Math. **119** (1967), 147–171, DOI 10.1007/BF02392081. MR222474

[21] W. K. Hughen, *The sub-Riemannian geometry of three-manifolds*, ProQuest LLC, Ann Arbor, MI, 1995. Thesis (Ph.D.)–Duke University. MR2692648

[22] E. Le Donne, *A metric characterization of Carnot groups*, Proc. Amer. Math. Soc. **143** (2015), no. 2, 845–849, DOI 10.1090/S0002-9939-2014-12244-1. MR3283670

[23] W. Liu and H. J. Sussman, *Shortest paths for sub-Riemannian metrics on rank-two distributions*, Mem. Amer. Math. Soc. **118** (1995), no. 564, x+104, DOI 10.1090/memo/0564. MR1303093

[24] G. A. Margulis and G. D. Mostow, *The differential of a quasi-conformal mapping of a Carnot-Carathéodory space*, Geom. Funct. Anal. **5** (1995), no. 2, 402–433, DOI 10.1007/BF01895673. MR1334873

[25] J. Mitchell, *On Carnot-Carathéodory metrics*, J. Differential Geom. **21** (1985), no. 1, 35–45. MR806700

[26] R. Montgomery, *Isoholonomic problems and some applications*, Comm. Math. Phys. **128** (1990), no. 3, 565–592. MR1045885

[27] R. Montgomery, *Gauge theory of the falling cat*, Dynamics and control of mechanical systems (Waterloo, ON, 1992), Fields Inst. Commun., vol. 1, Amer. Math. Soc., Providence, RI, 1993, pp. 193–218. MR1232916

[28] R. Montgomery, *Abnormal minimizers*, SIAM J. Control Optim. **32** (1994), no. 6, 1605–1620, DOI 10.1137/S0363012993244945. MR1297101

[29] R. Montgomery, *Hearing the zero locus of a magnetic field*, Comm. Math. Phys. **168** (1995), no. 3, 651–675. MR1328258

[30] R. Montgomery, *A tour of subriemannian geometries, their geodesics and applications*, Mathematical Surveys and Monographs, vol. 91, American Mathematical Society, Providence, RI, 2002. MR1867362

[31] G. D. Mostow, *Strong rigidity of locally symmetric spaces*, Princeton University Press, Princeton, N.J.; University of Tokyo Press, Tokyo, 1973. Annals of Mathematics Studies, No. 78. MR0385004

[32] E. Nelson, *Tensor Analysis*, Annals of Mathematics Studies, No. 78, 1967, Princeton University Press, Princeton, N.J.

[33] P. Pansu, *Une inégalité isopérimétrique sur le groupe de Heisenberg* (French, with English summary), C. R. Acad. Sci. Paris Sér. I Math. **295** (1982), no. 2, 127–130. MR676380

[34] P. Pansu, *Croissance des boules et des géodésiques fermées dans les nilvariétés* (French, with English summary), Ergodic Theory Dynam. Systems **3** (1983), no. 3, 415–445, DOI 10.1017/S0143385700002054. MR741395

[35] P. Pansu, *Métriques de Carnot-Carathéodory et quasiisométries des espaces symétriques de rang un* (French, with English summary), Ann. of Math. (2) **129** (1989), no. 1, 1–60, DOI 10.2307/1971484. MR979599

[36] P. Rashevskii, *About connecting two points of complete nonholonomic space by admissible curve*, Uch. Zapiski ped. inst. Libknexta, **2** (1938), 83–94.

[37] A. I. Šnirelʹman, *The asymptotic multiplicity of the spectrum of the Laplace operator* (Russian), Uspehi Mat. Nauk **30** (1975), no. 4 (184), 265–266. MR0413209

[38] A. Selberg, *Harmonic analysis and discontinuous groups in weakly symmetric Riemannian spaces with applications to Dirichlet series*, J. Indian Math. Soc. (N.S.) **20** (1956), 47–87. MR88511

[39] A. M. Vershik and V. Ya. Gershkovich, *The geometry of the nonholonomic sphere for three-dimensional Lie group*, Global analysis—studies and applications, III, Lecture Notes in Math., vol. 1334, Springer, Berlin, 1988, pp. 309–331, DOI 10.1007/BFb0080435. MR964707

[40] J. A. Wolf, *Growth of finitely generated solvable groups and curvature of Riemannian manifolds*, J. Differential Geometry **2** (1968), 421–446. MR248688

[41] L. C. Young, *Lectures on the calculus of variations and optimal control theory*, Foreword by Wendell H. Fleming, W. B. Saunders Co., Philadelphia-London-Toronto, Ont., 1969. MR0259704

Richard Montgomery
Mathematics Department
University of California, Santa Cruz
Santa Cruz California 95064
Email address: rmont@ucsc.edu