SUBRIEMANNIAN GEOMETRY: TWO OPEN PROBLEMS AND ONE FALLING CAT

RICHARD MONTGOMERY

I describe two open and closely related problems in subRiemannian geometry. I begin introducing this geometry through the problem of the falling cat. I end by reviewing some recent progress on the problems.

1.

A cat, falling from upside down with no spin, will right herself and land on her feet. How does she do it? What is her optimal strategy? We will use these questions to acquaint ourselves with subRiemannian geometry.

Imagine dropping a cat and a brick at the same time from the same height. (Hold them a few meters apart. Drop them from a meter or so above a couch, not from the top of the leaning tower of Pisa.) Drop them with no spin. See figure 1. They fall according to freshman physics, and land at the same time. The brick cannot suddenly stop its descent to hover in mid-air, nor can it start suddenly rotating about its long axis. Neither can the cat.

What the cat can do that the brick cannot is change shape. Through a sequence of mid-air shape changes she achieves an overall 180 degree rotation about the axis of her backbone. Upright, she stops changing shape and continues falling like the brick with no spin and lands on her feet. How do her shape changes lead to an overall rotation?

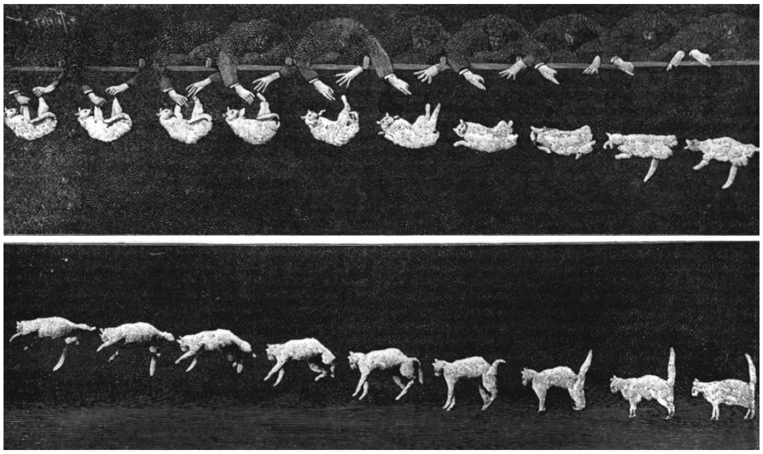

FIGURE 1. One of the earliest stroboscopic photos, taken in 1894 by Étienne-Jules Marey. See `https://publicdomainreview.org/collection/photographs-of-a-falling-cat-1894/` for details.

What is “a shape”? Two objects in space have the “same shape” if a rigid motion takes one object to the other. The set of rigid motions of space forms a Lie group G, the group generated by translations and rotations. So, “a shape” of the cat is a point in the quotient space $\mathcal{S}$ which we call shape space. This quotient space forms the base of a principal G-bundle whose total space is the configuration space of the cat, a space whose points represent shapes of cats, located and oriented in space.

There is nothing at all that the cat can do to affect her overall descent, which is to say the translational part of G. She translates downward just like a brick. So we may as well get rid of the translational part of G and just focus on the rotation group $SO(3)$. We will do so by going into an accelerating reference frame whose origin is at the cat's center of mass and whose axes are parallel to some inertial axes, say the edges made of the walls of the room. Having drug Galileo into our conversation we now drag Einstein in and invoke his principle of equivalence. The physics experienced by a cat in free-fall is the same physics experienced by a cat floating in zero gravity. Now imagine our outer-space cat making her shape changes. At the end of her shape changes she suffers some rigid rotation g about the origin. This rotation is the same as what she would suffer if she performed the same shape changes while in freefall.

The relevant physics turning a curve in the cat's shape space into an overall rotation is conservation of angular momentum. Angular momentum is a vector quantity which does not change during free-fall. It measures the spin of objects and is zero for non-spinning objects. Since the cat was dropped with no spin, her angular momentum remains zero throughout free-fall. We recall a formula for angular momentum. The total angular momentum of N point masses $m_1, \ldots, m_N$ located instantaneously at positions $q_1, \ldots, q_N \in \mathbb{R}^3$ and moving with velocities $v_1, \ldots v_N$ is given by

$$J(q, v) = \sum_a m_a q_a \times v_a \in \mathbb{R}^3. \tag{1}$$

By taking N large enough we can suppose that the N vectors q_a describe the complete configuration of the cat relative to our origin. Think of them as representative marker points, head, feet, hips, vertebrae, etc. There will be constraints amongst the vectors representing fixed bone lengths, ligaments and muscles holding the cat together to make it a cat. Note that

$$\sum m_a q_a = 0. \tag{2}$$

since the cat's center of mass is the origin. The set of all allowable $q = (q_1, q_2, \ldots, q_N)$ sweep out the cat's configuration space denoted Q. In this model $Q \subset (\mathbb{R}^3)^N$. Alternatively, we could make up some continuum model of the cat. However we model the cat, her configuration space will be a manifold Q on which $SO(3)$ acts freely and the angular momentum will be a vector-valued one-form $J : TQ \to \mathbb{R}^3$ so that $J(q, v)$ is linear in $v \in T_q Q$. Thus

$$D(q) := \{v \in T_q Q : J(q, v) = 0\} \subset T_q Q$$

is a linear subspace. Conservation of angular momentum tells us that in free-fall any motion $q(t) \in Q$ of the cat must satisfy

$$v(t) = \frac{dq}{dt} \in D(q(t)) \tag{3}$$

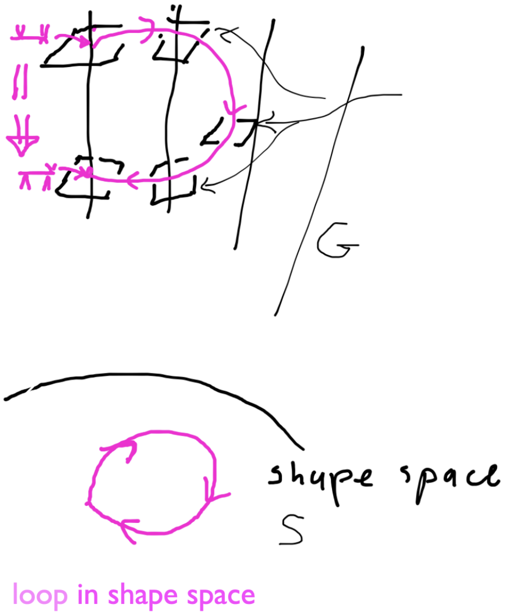

FIGURE 2. The loop in shape space represents a re-orientation strategy for the falling cat. The group G is $SO(3)$. The holonomy, represented by the small top arrow, is the amount by which the cat rotates upon traversing this loop. The total space containing the G-fibers is the configuration space for the cat. The small planes in the configuration space represent the distribution planes.

We reorganize the above facts into a standard geometric set-up. Shape space $\mathcal{S}$ is the quotient space of Q by $SO(3)$ so it forms the base space of a principal $SO(3)$-bundle

$$\pi : Q \xrightarrow{SO(3)} \mathcal{S} \tag{4}$$

The distribution $D = \bigcup_q D(q)$ is the horizontal space for a connection on this principal bundle. This means that $D(q)$ is transverse to the fiber of π. The map π is the quotient map, and so assigns to each cat configuration the shape of that cat.

As the cat changes her shape this shape traces out a curve $s(t)$ in $\mathcal{S}$. The actual cat configuration, the curve $q(t)$ of "located oriented cats" sitting in space, is the horizontal lift of this curve, which means that $\pi(q(t)) = s(t)$ and that the velocity constraint (3) holds. High-speed photography shows that the shape curve $s(t)$ is nearly closed: $s(0) = s(1)$. The cat's first problem, 'how do I change my shape so as to land on my feet?" has become the problem of finding a loop $s(t), 0 \leq t \leq 1$ in her shape space whose holonomy is the element $g \in SO(3)$ which is a 180 degree rotation about her ventral axis: $q(1) = gq(0)$. This problem has a function space's worth of solutions $s(t)$.

The second problem we had posed for our falling cat was to find an optimal re-orientation strategy $s(t)$. The answer of course depends on the choice of what to optimize. What the actual cat optimizes, if anything, is more up to evolution, chance, timing, and survival of the fittest, than it is up to mathematics or physics. But , being mathematicians, we use the natural metric lying around from mechanics. Return to our N landmark points $q_1, q_2, \ldots q_N$ and their velocities $v_1, \ldots, v_N$. The kinetic energy K of a motion is given by

$$K(v) := \frac{1}{2} \sum m_a |v_a|^2 \tag{5}$$

which we write as $\|v\|^2 = 2K(v)$ where $\|v\|^2 = \langle v, v \rangle$ and where $\langle \cdot, \cdot \rangle$ is an $SO(3)$-invariant Euclidean inner product on $(\mathbb{R}^3)^N$ and hence, by restriction, a Riemannian metric on Q. The second problem then becomes: minimize $\int_0^1 K(v(t))dt$ subject to the constraint that the loop $s(t)$ in $\mathcal{S}$ solves the first problem: namely its horizontal lift $q(t)$ leads to an upright cat $q(1)$, which is to say the correct holonomy g.

This is a constrained minimization problem. The constraints are that $v(t) \in D(q(t))$ and that the endpoints generated by the associated horizontal curve $q(t), 0 \leq t \leq 1$ with $\dot{q}(t) = v(t)$ are related by the required holonomy: $q(1) = gq(0)$ where g is the 180 degree rotation about the ventral axis of the located cat $q(0)$.

2. What is subRiemannian geometry?

Forget the Lie group action on Q. Keep the linear sub-bundle

$$D \subset TQ$$

of its tangent bundle TQ and the fiber inner product $\langle \cdot, \cdot \rangle$ on D. The data $(D, \langle \cdot, \cdot \rangle)$ defines a subRiemannian geometry on the manifold Q.

Call an absolutely continuous path 'horizontal' if it is tangent to D. (Recall the derivative of an absolutely continuous path $q : I \to Q$ exists at almost all points t of the interval I of its domain.) Use the inner product to measure the length of such a path, as in Riemannian geometry:

$$\ell(q) = \int_I \sqrt{\langle \dot{q}, \dot{q} \rangle} dt. \tag{6}$$

Given two points $A, B \in Q$ we look for the shortest horizontal path that joins them. Such a path is called *a subRiemannian geodesic*. The optimal cat re-orientation problem as we formulated it is a problem of finding a subRiemannian geodesic.

Open Problem 1. *Must a subRiemannian geodesic be smooth?*

3. Geting from A to B

Before delving into the subRiemannian geodesic problem we ought to know whether or not there is any horizontal path at all joining A to B. If D is involutive then the answer is typically 'no'. Recall that to say that D is involutive means $[D, D] \subset D$. The brackets denote Lie brackets of vector field and we have abused notation by letting D also mean the sheaf of all smooth vector fields on Q which take values in $D \subset TQ$. Thus $[D, D]$ denotes the sheaf generated by the Lie brackets of pairs of vector fields taken from the sheaf D. The Frobenius theorem asserts that if D is involutive then the set of all points B joined to A by horizontal paths forms the leaf of a foliation which is tangent to D. These leaves are immersed submanifolds of Q whose dimension equals the rank of D. A typical B does not lie on the leaf through A, so one cannot go from A to B.

At the other extreme of being involutive are the bracket-generating distributions. Suppose that $[D, D] \neq D$. Set $D^2 = [D, D]$ and keep taking Lie brackets with D, setting $D^3 = [D, D^2]$ and so on. If for some j we have that $D^j = TQ$ then we say that D is bracket-generating. Equivalently, D is bracket-generating if every vector in the tangent space to Q at any point can be written as a linear combination of iterated Lie brackets of vector fields tangent to D at that point. A theorem due to Rashevskii in 1938 and independently to Chow in 1939 asserts that for a bracket-generating distribution on a connected manifold any two points can be connected by a horizontal path.

I pause the narrative for a biographical "station break". Chow (1911-1995) is the same Chow whose name is central in algebraic geometry. His remarkable story of leaving China, beginning college education in Kentucky, continuing in Chicago, Gottingen, and Hamburg, marrying, leaving Nazi Germany, returning to China to find himself in Nanking during the war with Japan, and then, with some help from Chern, moving again to the U.S. where he had a position at the Institute of Advanced Studies can be found at [3]. Rashevskii (1907-1983) was born, graduated, and died in Moscow. He was the head of the differential geometry department at Moscow State (Lomonosov) University from 1964 until his death. Agrachev, a central figure in the research around the problems I describe here, wrote to me about Rashevskii's influence on him when he was an undergraduate, "I more liked history, philosophy and things like that but it was the Soviet Union and I realised that any activity of this kind is under total ideological control of the authorities. Trying to find something ideologically neutral, I took the book by Hilbert 'Foundations of Geometry' (Russian translation). This is an extremely boring book but there was a long introductory article (something like 50 pages or more) about the history of the 5th Euclidean postulate written by Rashevskii. It was so exciting and I decided to try to enter the Math. department."

4. Control theory

We can view the Chow-Rashevksii theorem and subRiemannian geodesics from a control-theoretic perspective. For simplicity, assume D is rank 2 and is framed by two smooth vector fields $X_1, X_2 : Q \to TQ$. Then the derivative of any horizontal path $q(t)$ can be expressed as

$$\dot{q}(t) = u_1(t)X_1(q(t)) + u_2(t)X_2(q(t)) \tag{7}$$

The $u_1(t), u_2(t)$ are viewed as "controls"- internal torques for changing the shape of the cat, if you will. Take the controls to be square integrable functions of t with $t \in I = [0, 1]$. Consider (7) as an initial value problem. Specifying the initial condition $q(0) = A$ and the controls yields a unique absolutely continuous path $q(t)$, $t \in I$ by solving (7). Write Ω_A for the space of horizontal absolutely continuous horizontal paths starting at A and parameterized by I. Then (7) together with $q(0) = A$ defines a global chart $\phi : L_2(I, \mathbb{R}^2) \to \Omega_A$. The endpoint map is the map $q(\cdot) \mapsto q(1)$ which sends a path to its endpoint. We write it as

$$end_A : \Omega_A \to Q; \qquad end_A(q) = q(1) \tag{8}$$

The Chow-Rashevskii theorem asserts that if D is bracket-generating then end_A is an open mapping. It is not hard to show that the image of end_A is also closed, so as a corollary we get the original version of Chow-Rashevskii: every point B can be connected to A by a horizontal path when D is bracket-generating and Q is connected.

We move on to the problem of subRiemannian geodesics. We may assume that X_1, X_2 are orthonormal relative to our inner product on D. Then the length (6) of our horizontal path as specified by our chart is

$$\ell(q) = \ell(u) = \int_I \sqrt{u_1(t)^2 + u_2(t)^2} dt.$$

The subRiemannian geodesic problem is thus a constrained minimization problem: minimize $\ell(u)$ subject to the constraint $end_A(u) = B$.

As we explain in the next section, a non-smooth subRiemannian geodesic $q \in \Omega_A$ must be a critical point of the endpoint map. Consequently, the following problem becomes important in answering our first problem, problem 1.

Open Problem 2. *Does Sard's theorem hold for the endpoint map? In other words, is the set of critical values of the endpoint map (8) a set of measure zero in Q?*

A 'yes' answer to this question would imply that for almost every end point B every subRiemannian geodesic connecting A to B is smooth.

We recall that the standard Sard theorem asserts that if G is a smooth map between finite-dimensional smooth manifolds then the set of critical values of G has measure zero within the range manifold. This theorem is one of the basic workhorse theorems of differential topology. However the Sard theorem is false in general when the domain manifold is infinite-dimensional. See for example [5] where Kupka wrote out a scalar cubic polynomial $P : \ell_2 \to \mathbb{R}$ whose set of critical values is the unit interval.

We reserve a special name for the horizontal curves which are critical points for the endpoint map.

Definition 4.1. *A singular curve for a distribution D is a horizontal curve which is a critical point for the endpoint map (8), where the base point A of the map is the curve's starting point.*

5. Singular Geodesics.

The subRiemannian geodesic problem is a constrained minimization problem: minimize F subject to a constraint $G = k$. Here F and G are smooth functions and k is a constant. We teach multi-variable calculus students to approach such problems by introducing a Lagrange multiplier λ for the constraint function. We tell them to form $F + \lambda G$, set its differential to zero and solve, remembering to impose $G = const$. This prescription works great if k is a regular value for the constraint function G. But if k is a singular value for G then the method may fail to detect minimizers which lie on the singular locus of G. As a simple example suppose that $G(x, y) = y^3 - x^2$ and $k = 0$. Then the cusp $(x, y) = (0, 0)$ of $G = 0$ is a local minimum for *any* linear function $F(x, y) = ax + by$ for which $b > 0$. The prescription we teach our students fails to catch the singular point on the zero locus.

To fix up the Lagrange multiplier method we must also include a multiplier λ_0 for the function F to be minimized. We insist that $(\lambda_0, \lambda) \neq (0, 0)$. It is a theorem that any constrained minimizer must lie among the critical points of the pencil of functions $\lambda_0 F + \lambda G$. The cusp point of our above example is caught by allowing $\lambda_0 = 0$, $\lambda \neq 0$. Minimizers for which $\lambda_0 = 0$ are called "singular" minimizers in that they correspond to critical points of the constraint function and so possible singular points of the variety $G = k$. The extremals for which $\lambda_0 \neq 0$ are called 'normal' or "regular".

The regular minimizers of the subRiemannian geodesic problem are characterized as being the projections to Q of solutions to a smooth ODE, in particular of a smooth Hamiltonian vector field. For several decades it was believed that all subRiemannian geodesics were regular minimizers. Several false proofs could be found in reputable journals. In 1990 I established the existence of singular minimizers: subRiemannian metrics supporting geodesics which are singular curves which are not regular minimizers. My examples could not be perturbed away: their singular geodesics persist when the distribution and metric on it are perturbed. The existence of these geodesics turned the previously apparently closed problem (1) back into an open one.

6. Magnetic fields and first examples

Simplify the cat problem by replacing the group $SO(3)$ by the additive group $\mathbb{R}$ and the principal bundle (4) by the bundle $\mathbb{R} \to \mathbb{R}^3 \to \mathbb{R}^2$ where $\pi(x, y, z) = (x, y)$. Continue to take the distribution D to be the horizontal distribution associated to a connection on this bundle. Such a distribution can be expressed as the vanishing of a connection one-form

$$\theta = dz - A_1(x,y)dx - A_2(x,y)dy \tag{9}$$

$$= dz - \alpha, \tag{10}$$

so that a curve $c(t) = (x(t), y(t), z(t))$ is horizontal if and only if

$$\dot{z}(t) = A_1(x(t), y(t))\dot{x}(t) + A_2(x(t), y(t))\dot{y}(t)$$

where the dots denote time derivatives. The horizontal curve $c(t)$ is uniquely determined by its planar projection $(x(t), y(t))$ and the value of z at a single value of t.

Write

$$d\alpha = \beta(x,y)dx \wedge dy$$

For reasons to be described momentarily we think of β as a magnetic field, or more accurately, the z-component of a magnetic field orthogonal to the xy plane.

Theorem 6.1. *A non-constant horizontal curve is singular if and only if its projection lies in the zero locus $\beta = 0$ of the magnetic field. If zero is a regular value of β then any sufficiently short arc of such a singular horizontal curve is a minimizing geodesic between its endpoints.*

This fact regarding the geodesic nature of the zero locus is independent of the inner product on D used to define the subRiemannian structure. The lengths of the "sufficiently short" arcs which minimize will depend on this inner product. In order to explain why we used the words "magnetic", take for inner product the one for which the length of a horizontal curve is the Euclidean length of its planar projection. With respect to this inner product the horizontal vector fields $X_1 = \frac{\partial}{\partial x} + A_1(x,y)\frac{\partial}{\partial z}$ and $X_2 = \frac{\partial}{\partial y} + A_2(x,y)\frac{\partial}{\partial z}$ form an orthonormal frame for D.

Endowed with this subRiemannian structure the ODEs characterizing the regular geodesics are:

$$\ddot{x} = \lambda\beta(x, y)\dot{y}$$

$$\ddot{y} = -\lambda\beta(x, y)\dot{x}$$

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

$$\dot{z} = A_1(x, y)\dot{x} + A_x(x, y)\dot{y}$$

The constant λ plays the role is the Lagrange multiplier of $F + \lambda G$. *These are precisely the ODEs which describe the motion of a non-relativistic charged particle travelling in the plane under the influence of the planar magnetic field β.* The constant λ is $\lambda = e/m$ where e is the charge of the particle and m its mass.

A simple computation shows that the curve $\beta = 0$ (assumed a smooth curve) satisfies the regular geodesic equations if and only if the curve is a straight line. Consequently as long as our zero locus does not contain a straight line segment no arc of any of the geodesics of the theorem are regular geodesics.

7. Progress

We describe three recent results which represent progress on these two open problems.

7.1. **No Corners.**

Theorem 6.1 covered the case where zero is a regular value of β and consequently the zero locus $\beta^{-1}(0)$ is a smooth curve. However we can take 0 to be a critical value and the zero locus continues to capture all of the singular curves. That is, any singular non-constant horizontal curve lies within the zero locus. As an example, take $\beta = xy$. Its zero locus is the union of the x and y axes. Travel down along the y axis to the origin then turn along the x axis to form a right angle. Parameterize it and take its horizontal lift. Could this be a geodesic? No!

Theorem 7.1. [4] *SubRiemannian geodesics cannot have corners. More precisely, if at some point along a geodesic both its right and left derivatives exist then these two derivatives must be equal.*

This theorem, published in 2016 by Hakavouri and LeDonne in [4], represents the most serious progress made to date on problem (1).

At first glance it might seem this theorem solves problem 1, or at least shows that the geodesics must be continuously differentiable (i.e. C^1). However a rectifiable path (parameterized by arclength) can have singularities much less tractable than corners. It can have spirals leading to a point on the curve where neither the right nor the left-handed derivative exist. Worse, its set of non-differentiable points may form a measure zero Cantor set in which case we cannot separate the bad points from each other. Showing that the set of non-differentiable points of a singular geodesic forms a finite set would be huge progress. So the problem remains open and **we do not even know if geodesics need be C^1.**

7.2. **Singular minimizers.** The extent of our ignorance around problem 2 is large. We cannot even exclude the possibility that the set of endpoints of singular curves starting at A forms a neighborhood of A. However, if we restrict ourselves to singular curves which are also minimizers the situation is markedly better.

Theorem 7.2. *The set of endpoints of singular minimizers forms a set whose complement is open and dense.*

Recall that a subset of a manifold whose complement is open and dense may still have postive measure. Examples are the fat Cantor subsets of the real line. So this theorem allows for the possibility that the endpoints of singular minimizers form a set of positive measure. See chapter 11 of [1] for a proof of the theorem and further information around it.

7.3. **Strong Sard.** This next result is specific to rank 2 distributions in 3 dimensions. The magnetic field case of (10) is such a case. We can define such a distribution locally by the vanishing of a one-form θ as we did in equation (10). Choose a volume $dx \wedge dy \wedge dz$ for $Q = Q^3$ and write

$$\theta \wedge d\theta = f dx \wedge dy \wedge dz \tag{11}$$

In our magnetic example $f = \beta$ is the magnetic field. The distribution is said to be "contact" if $f \neq 0$. The study of contact distributions and their invariants has become an active area in the last 30 years, the area called "contact topology". But contact distributions admit no non-constant singular curves. Consequently, the Sard problem becomes vacuous near contact points of our distribution, i.e. near points A with $f(A) \neq 0$.

Suppose that θ and the volume form are analytic so that f is analytic. It is zero if and only if the distribution is involutive. So suppose that f is a nonconstant analytic function with a nontrivial zero locus

$$\{q : f(q) = 0\} = M \subset Q.$$

This zero locus M, known as the Martinet surface, is an analytic surface which may have singularities, and which contains all singular curves. Since surfaces of a 3-manifold have measure zero, the Sard problem, Open Problem 2, is automatically answered in the affirmative.

Subsequent analysis suggested the following strengthening of the Sard problem.

Problem 7.1 (Strong Sard)**.** *Suppose $A \in M$ and that $M \subset Q$ with M the Martinet (=non-contact) hypersurface for an analytic rank 2 distribution as described above. Do the set of critical values B of the end point map have measure zero within M?*

In 2022 A. Belotto da Silva, A. Figalli, A. Parusínski and L. Rifford [2] answered this question affirmatively.

8. Further Reading

I can recommend two books on the subject, [1] and [6]. For a short and lively introduction making connections between subRiemannian geometry and the geometries associated to other 2nd order linear PDE see the article "Realms of Mathematics: Elliptic, Hyperbolic, Parabolic, Sub-Elliptic" [7]. SubRiemannian geometry is the geometry of the "sub-elliptic" realm, being represented by the subLaplacian for a subRiemannian geometry.

References

[1] A. Agrachev, D. Barilari, and U. Boscain, (2020), **A Comprehensive Introduction to subRiemannian Geometry**, Cambridge U. Press,.

[2] A. Belotto da Silva, A. Figalli, A. Parusínski and L. Rifford (2022), *Strong Sard Conjecture and regularity of singular minimizing geodesics for analytic sub-Riemannian structures in dimension 3,* Invent. Math., **229**, 395-448

[3] `https://mathshistory.st-andrews.ac.uk/Biographies/Chow/`

[4] E. Hakavuori and E. Le Donne, (2016), *Non-minimality of corners in subriemannian geometry*, Invent. Math., **206**, 693 - 704.

[5] I. Kupka, (1965), *Counterexample to the Morse-Sard theorem in the case of infinite-dimensional manifolds*, Proc. Amer. Math. Soc., **16**, 954-957.

[6] R. Montgomery, (2002), **A tour of sub-Riemannian geometries, their geodesics and applications**, Mathematical Surveys and Monographs, **91**, American Mathematical Society.

[7] R. Strichartz, (1987) *Realms of Mathematics: Elliptic, Hyperbolic, Parabolic, Sub-Elliptic*, The Math. Intelligencer, **9**, no. 3.

Mathematics Department, University of California, Santa Cruz, Santa Cruz CA 95064

Email address: `rmont@ucsc.edu`