Introduction

This book is concerned with a fresh development of the eternal idea of the distance as the length of a shortest path between two points. In Euclidean geometry, shortest paths are segments of straight lines that satisfy all the classical axioms. In the Riemannian world, Euclidean geometry is just one of a huge number of possibilities. However, each possibility is well approximated by Euclidean geometry at a very small scale. In other words, Euclidean geometry is treated as the geometry of the initial velocities of paths starting from a fixed point of a Riemannian space rather than the geometry of the space itself.

The Riemannian construction is based on the previous study of smooth surfaces in Euclidean space undertaken by Gauss. The distance between two points on a surface is the length of a shortest path on the surface connecting the points. The initial velocities of smooth curves starting from a fixed point on the surface form a tangent plane to the surface that is a Euclidean plane. The tangent planes at two different points are isometric, but the neighborhoods of the points on the surface are not locally isometric in general, certainly, they are not if the Gaussian curvature of the surface is different at the two points.

Riemann generalized Gauss' construction to higher dimensions and realized that it can be done in an intrinsic way; you do not need an ambient Euclidean space to measure the lengths of curves. Indeed, to measure the length of a curve it is sufficient to know the Euclidean length of its velocities. A Riemannian space is a smooth manifold whose tangent spaces are endowed with Euclidean structures; each tangent space is equipped with its own Euclidean structure, which smoothly depends on the point where the tangent space is attached.

For a habitant sitting at a point of the Riemannian space, the tangent vectors give directions about where to move or, more generally, sending and receiving information. The habitant measures the lengths of vectors, and angles between

vectors attached at the same point, according to the Euclidean rules, and this is essentially all that can be done. It is important that our habitant can, in principle, completely recover the geometry of the space by performing these simple measurements along different curves.

In a sub-Riemannian space we cannot move, receive or send information in all directions. There are restrictions (imposed by God, the moral imperative, the government or simply a physical law). A sub-Riemannian space is a smooth manifold with a fixed admissible subspace in any tangent space where the admissible subspaces are equipped with Euclidean structures. The admissible paths are those curves whose velocities are admissible. The distance between two points is the infimum of the lengths of the admissible paths connecting the points. It is assumed that any pair of points in the same connected component of the manifold can be connected by at least one admissible path. The last assumption might look strange at first glance, but it is not. The admissible subspace depends on the point where it is attached, and our assumption is satisfied for a more or less general smooth dependence on the point; it is perhaps better to say that only for very special families of admissible subspaces is it not satisfied.

Let us describe a simple model. Let our manifold be $\mathbb{R}^3$ with coordinates x, y, z. We consider the differential 1-form $\omega = -dz + \frac{1}{2}(xdy - ydx)$. Then $d\omega = dx \wedge dy$ is the pullback on $\mathbb{R}^3$ of the area form on the xy-plane. In this model the subspace of admissible velocities at the point (x, y, z) is assumed to be the kernel of the form ω. In other words, a curve $t \mapsto (x(t), y(t), z(t))$ is an admissible path if and only if $\dot{z}(t) = \frac{1}{2}\,(x(t)\dot{y}(t) - y(t)\dot{x}(t))$ or, equivalently,

$$z(t) = z(0) + \frac{1}{2}\int_0^t \big(x(s)\dot{y}(s) - y(s)\dot{x}(s)\big)ds.$$

If $x(0) = y(0) = z(0) = 0$ then $z(t)$ is the signed area of the domain bounded by the curve and the segment connecting $(0, 0)$ with $(x(t), y(t))$.

In this geometry, the length of an admissible tangent vector $(\dot{x}, \dot{y}, \dot{z})$ is defined to be $(\dot{x}^2 + \dot{y}^2)^{1/2}$, i.e., the length of the projection of the vector on the (x, y)-plane. By construction, the sub-Riemannian length of the admissible curve in $\mathbb{R}^3$ is equal to the Euclidean length of its projection on the plane.

In this geometry, to compute the shortest paths connecting the origin $(0, 0, 0)$ to a fixed point (x_1, y_1, z_1), we are reduced to solving the classical Dido isoperimetric problem: find a shortest planar curve among those connecting $(0, 0)$ with (x_1, y_1) and such that the signed area of the domain bounded by the curve and the segment joining $(0, 0)$ and (x_1, y_1) is equal to z_1 (see Figure 1).

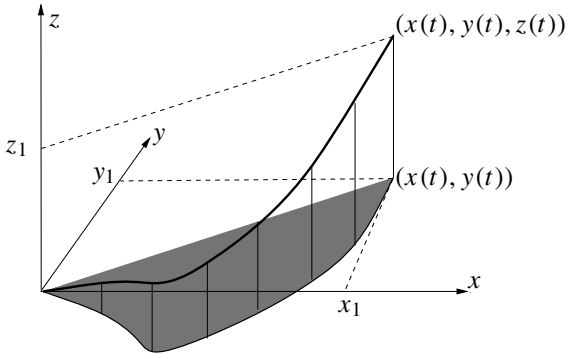

Figure 1 The Dido problem.

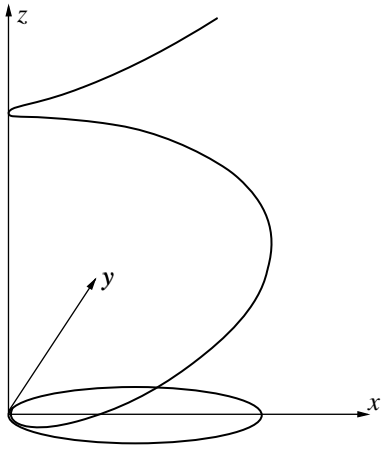

Figure 2 Solutions to the Dido problem.

Solutions of the Dido problem are arcs of circles, and their lifts to $\mathbb{R}^3$ are spirals, where $z(t)$ is the area of the piece of disk cut by the chord connecting $(0,0)$ with $(x(t), y(t))$ (see Figure 2).

A piece of such a spiral is a shortest admissible path between its endpoints while the planar projection of this piece is an arc of a circle. The spiral ceases to be a shortest path when its planar projection starts to run around the circle for a second time, i.e., when the spiral starts its second turn. For this model, sub-Riemannian balls centered at the origin look like apples with singularities at the poles (see Figure 3).

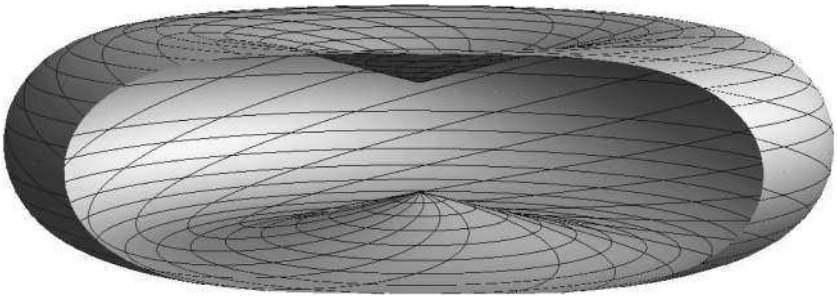

Figure 3 The Heisenberg sub-Riemannian sphere.

Singularities are points on the sphere connected with the center by more than one shortest path. The dilation $(x, y, z) \mapsto (rx, ry, r^2 z)$ transforms a ball of radius 1 into a ball of radius r. In particular, arbitrarily small balls have singularities. This is always the case when the admissible subspaces are proper subspaces.

Another important symmetry connects balls with different centers. Indeed, the product operation

$$(x, y, z) \cdot (x', y', z') \doteq \left(x + x',\ y + y',\ z + z' + \frac{1}{2}(xy' - x'y) \right)$$

turns $\mathbb{R}^3$ into a group, the *Heisenberg group*. The origin in $\mathbb{R}^3$ is the unit element of this group. It is easy to see that left-translations of the group transform admissible curves into admissible curves and preserve the sub-Riemannian length. Hence left-translations transform balls into balls of the same radius. A detailed description of this example and other models of sub-Riemannian spaces is given in Sections 4.4.3, 7.5.1 and 13.2.

Actually, even this simplest model tells us something about life in a sub-Riemannian space. Here we are dealing with planar curves but, in fact, we are operating in a three-dimensional space. Sub-Riemannian spaces always have a kind of hidden extra dimension, as a good and not yet exploited source of mystical speculations and also a source for theoretical physicists who are searching for new crazy formalizations. In mechanics, this is a natural geometry for systems with nonholonomic constraints such as skates, wheels, rolling balls, bearings etc. This kind of geometry could also serve to model social behavior that allows an increase in the level of freedom without violation of a restrictive legal system.

Anyway, in this book we are performing a purely mathematical study of sub-Riemannian spaces in order to provide an appropriate formalization ready for all potential applications. Riemannian spaces appear as a very special case. Of course, we are not the first to study the sub-Riemannian stuff. There is a broad literature, even if there are not so many experts who could claim that sub-Riemannian geometry is their main field of expertise. Important motivations come from CR geometry, hyperbolic geometry, the analysis of hypoelliptic operators and some other domains. Our first motivation is control theory: length-minimizing is a nice class of optimal control problems.

Indeed, one can find a control theory spirit in our treatment of the subject. First of all, we include admissible paths in admissible flows, which are flows generated by vector fields whose values at all points belong to admissible subspaces. The passage from admissible subspaces attached at different points of the manifold to a globally defined space of admissible vector fields makes the structure more flexible and well-adapted to algebraic manipulations. We pick generators $f_1, \dots, f_k$ of the space of admissible fields, and this allows us to describe all admissible paths as solutions to time-varying ordinary differential equations of the form $\dot{q}(t) = \sum_{i=1}^{k} u_i(t) f_i(q(t))$. Different admissible paths correspond to the choice of different *control functions* $u_i(\cdot)$ and initial points $q(0)$, while the vector fields f_i are fixed at the outset.

We also use a Hamiltonian approach supported by the Pontryagin maximum principle to characterize shortest paths. A few words about the Hamiltonian approach: sub-Riemannian geodesics are admissible paths whose sufficiently small pieces are length-minimizers, i.e., the length of such a piece is equal to the distance between its endpoints. In the Riemannian setting, any geodesic is uniquely determined by its velocity at the initial point q. In the general sub-Riemannian situation we have many more geodesics based at the point q than admissible velocities at q. Indeed, every point in a neighborhood of q can be connected with q by a length-minimizer, while the dimension of the admissible-velocities subspace at q is usually smaller than the dimension of the manifold.

What is a natural parametrization of the space of geodesics? To understand this question, we adapt a classical "trajectory–wave front" duality. Given a length-parametrized geodesic $t \mapsto \gamma(t)$, we expect that the values at a fixed time t of geodesics starting at $\gamma(0)$ and close to γ fill a piece of a smooth hypersurface (see Figure 4). For small t this hypersurface is a piece of a sphere of radius t, while in general it is only a piece of the "wave front".

Moreover, we expect that $\dot{\gamma}(t)$ is transversal to this hypersurface. This is not always the case but it is true for a generic geodesic.

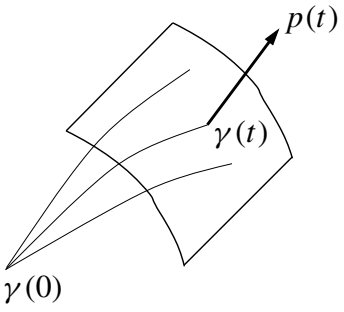

Figure 4 The "wave front" and the "momentum".

The "momentum" $p(t) \in T^*_{\gamma(t)}M$ is the covector orthogonal to the "wave front" and is normalized by the condition $\langle p(t), \dot\gamma(t)\rangle = 1$. The curve $t \mapsto (p(t), \gamma(t))$ in the cotangent bundle T^*M satisfies a Hamiltonian system. This is exactly what happens in rational mechanics or geometric optics.

The sub-Riemannian Hamiltonian $H : T^*M \to \mathbb{R}$ is defined by the formula $H(p,q) = \frac{1}{2}\langle p, v\rangle^2$, where $p \in T^*_qM$ and $v \in T_qM$ is an admissible velocity of length 1 that maximizes $\langle p, w\rangle$ among all admissible velocities w of length 1 at $q \in M$.

Any smooth function on the cotangent bundle defines a Hamiltonian vector field and such a field generates a Hamiltonian flow. The Hamiltonian flow on T^*M associated with H is the *sub-Riemannian geodesic flow*. The Riemannian geodesic flow is just a special case.

As mentioned earlier, in general the construction described above cannot be applied to all geodesics: the so-called abnormal geodesics are missed out. An abnormal geodesic $\gamma(t)$ also possesses a "momentum" $p(t) \in T^*_{\gamma(t)}M$ but this momentum belongs to the orthogonal complement to the subspace of admissible velocities and does not satisfy the above Hamiltonian system. Geodesics that are trajectories of the geodesic flow are called *normal*. Actually, abnormal geodesics belong to the closure of the space of normal geodesics, and elementary symplectic geometry provides a uniform characterization of the momenta for both classes of geodesics. Such a characterization is, in fact, a very special case of the Pontryagin maximum principle.

Recall that all velocities are admissible in the Riemannian case and that the Euclidean structure on the tangent bundle induces the identification of tangent vectors and covectors, i.e., of the velocities and momenta. We should remember, however, that this identification depends on the metric. One can think of a sub-Riemannian metric as the limit of a family of Riemannian

metrics when the lengths of forbidden velocities tend to infinity while the lengths of admissible velocities remain untouched. It is easy to see that on the one hand the Riemannian Hamiltonians defined by such a family converge with all derivatives to the sub-Riemannian Hamiltonian; hence, Riemannian geodesics with a prescribed initial momentum converge to the sub-Riemannian geodesic with the same initial momentum. On the other hand, we cannot expect any reasonable convergence for a family of Riemannian geodesics with a prescribed initial velocity: those with forbidden initial velocities disappear at the limit while the number of geodesics with admissible initial velocities jumps to infinity.

Outline of the Book

We start in Chapter 1 by considering surfaces in $\mathbb{R}^3$, which are the beginning of everything in differential geometry and also form a starting point for the story told in this book. There are not yet Hamiltonians here, but a control flavor is already present. The presentation is elementary and self-contained. A student in applied mathematics or analysis who missed the geometry of surfaces at university, or simply is not satisfied by his or her understanding of these classical ideas, might find it useful to read just this chapter even if he or she does not plan to study the rest of the book.

In Chapter 2 we recall some basic properties of vector fields and vector bundles. Sub-Riemannian structures are defined in Chapter 3, where we also study three fundamental facts: the finiteness and continuity of the sub-Riemannian distance, the existence of length-minimizers and the infinitesimal characterization of geodesics. The first constitutes the classical Rashevskii–Chow theorem and the other two are simplified versions of the Filippov existence theorem and of the Pontryagin maximum principle.

In Chapter 4, we introduce symplectic language. We define the geodesic Hamiltonian flow, we consider some interesting two- and three-dimensional problems and we prove a general sufficient condition for the length-minimality of normal trajectories. Chapter 5 is devoted to integrable Hamiltonian systems. We explain the construction of the action-angle coordinates and describe classical examples of integrable geodesic flows, such as the geodesic flow on ellipsoids.

Chapters 1–5 form the first part of the book, in which we do not use any tool from functional analysis. In fact, even the knowledge of Lebesgue integration and elementary real analysis is not essential, with the unique exception of the existence theorem in Section 3.3. In all other parts of the text the reader will

nevertheless understand the content by just replacing the terms "Lipschitz" and "absolutely continuous" with "piecewise C^1" and the term "measurable" with "piecewise continuous".

We start to use some basic functional analysis in Chapter 6. In this chapter we give elements of an operator calculus that simplifies and clarifies calculations with nonstationary flows, their variations and compositions. In Chapter 7 we give a brief introduction to Lie group theory. Lie groups are introduced as subgroups of the groups of diffeomorphisms of the manifold M induced by a family of vector fields whose Lie algebra is finite dimensional. Then we study left-invariant sub-Riemannian structures and their geodesics.

In Chapter 8, we interpret "momenta" as Lagrange multipliers for constrained optimization problems and apply this point of view to the sub-Riemannian case. We also introduce the sub-Riemannian exponential map and study cut and conjugate points.

In Chapter 9, we consider two-dimensional sub-Riemannian metrics; such a metric coincides with a Riemannian metric on an open and dense subset. We describe in detail the model space of this geometry, known as the Grushin plane, and we discuss several properties in the generic case, including a Gauss–Bonnet-like theorem.

In Chapter 10 we construct the nonholonomic tangent space at a point q of the manifold: it is a first quasi-homogeneous approximation to the space if you observe and exploit it from q by means of admissible paths. In general, such a tangent space is a homogeneous space of a nilpotent Lie group equipped with an invariant vector distribution; its structure may depend on the point where the tangent space is attached. At generic points this is a nilpotent Lie group endowed with a left-invariant vector distribution. The construction of the nonholonomic tangent space does not need a metric; if we take into account the metric then we obtain the Gromov–Hausdorff tangent to the sub-Riemannian metric space. Useful "ball–box" estimates of small balls follow automatically.

In Chapter 11 we study the general analytic properties of the sub-Riemannian distance as a function of points of the manifold. It is shown that the distance is smooth on an open dense subset and is Lipschitz outside the points connected by abnormal length-minimizers. Moreover, if these bad points are absent then almost every sphere is a Lipschitz submanifold.

In Chapter 12 we turn to abnormal geodesics, which provide the deepest singularities of the distance. Abnormal geodesics are critical points of the endpoint map defined on the space of admissible paths, and the main tool for

their study is the Hessian of the endpoint map. This study permits us to prove also that the cut locus from a point is adjacent to the point itself as long as the structure is not Riemannian.

Chapter 13 is devoted to the explicit calculation of the sub-Riemannian optimal synthesis for model spaces. After a discussion on Carnot groups, we describe a technique based on the Hadamard theorem that permits one, under certain assumptions, to compute the cut locus explicitly. We then apply this technique to several relevant examples.

This is the end of the second part of the book; the next few chapters are devoted to curvature and its applications. Let $\Phi^t : T^*M \to T^*M$, for $t \in \mathbb{R}$, be a sub-Riemannian geodesic flow. The submanifolds $\Phi^t(T^*_qM)$, $q \in M$, form a fibration of T^*M. Given $\lambda \in T^*M$, let $J_\lambda(t) \subset T_\lambda(T^*M)$ be the tangent space to the leaf of this fibration.

Recall that Φ^t is a Hamiltonian flow and T^*_qM are Lagrangian submanifolds; hence the leaves of our fibrations are Lagrangian submanifolds and $J_\lambda(t)$ is a Lagrangian subspace of the symplectic space $T_\lambda(T^*M)$.

In other words, $J_\lambda(t)$ belongs to the Lagrangian Grassmannian of $T_\lambda(T^*M)$, and $t \mapsto J_\lambda(t)$ is a curve in the Lagrangian Grassmannian, i.e., a *Jacobi curve* of the sub-Riemannian structure. The curvature of the sub-Riemannian space at λ is simply the "curvature" of this curve in the Lagrangian Grassmannian.

Chapter 14 is devoted to the elementary differential geometry of curves in the Lagrangian Grassmannian. In Chapter 15 we apply this geometry to Jacobi curves, which are curves in the Lagrange Grassmannian representing Jacobi fields.

The language of Jacobi curves is translated to the traditional language in the Riemannian case in Chapter 16. We recover the Levi-Civita connection and the Riemannian curvature and demonstrate their symplectic meaning. In Chapter 17 we compute explicitly the sub-Riemannian curvature for contact three-dimensional spaces, and we show how the curvature invariants appear in the classification of sub-Riemannian left-invariant structures on three-dimensional Lie groups. In Chapter 18, after a brief introduction to Poisson manifolds, we prove the integrability of the sub-Riemannian geodesic flow on 3D Lie groups. As a byproduct, we obtain a classification of coadjoint orbits on 3D Lie algebras. In Chapter 19 we study the small-distance asymptotics of the exponential map for the 3D contact case and see how the structure of the conjugate locus is encoded in the curvature.

In Chapter 20 we address the problem of defining a canonical volume in sub-Riemannian geometry. First, we introduce the Popp volume, which is a canonical volume that is smooth for equiregular sub-Riemannian manifolds,

and we study its basic properties. Then we define the Hausdorff volume and study its density with respect to the Popp volume.

In the final chapter, Chapter 21, we define the sub-Riemannian Laplace operator and study its properties (hypoellipticity, self-adjointness etc.). We conclude with a discussion of the sub-Riemannian heat equation and an explicit formula for the heat kernel in the 3D Heisenberg case.

The book is completed by an appendix on the canonical frames for a wide class of curves in the Lagrangian Grassmannians, written by Igor Zelenko. This is necessary background for a deeper systematic study of the curvature-type sub-Riemannian invariants, which is beyond the scope of this book. Notes at the ends of the chapters contain references and suggestions for further reading.