

Small time heat kernel asymptotics at the sub-Riemannian cut locus

Davide Barilari

CNRS, CMAP Ecole Polytechnique and equipe INRIA GECO Saclay-Île-de-France, Paris, France
barilari@cmap.polytechnique.fr

Ugo Boscain

CNRS, CMAP Ecole Polytechnique and equipe INRIA GECO Saclay-Île-de-France, Paris, France
boscain@cmap.polytechnique.fr

Robert W. Neel

Department of Mathematics, Lehigh University, Bethlehem, PA, USA
robert.neel@lehigh.edu

April 12, 2012

Abstract

For a sub-Riemannian manifold provided with a smooth volume, we relate the small time asymptotics of the heat kernel at a point y of the cut locus from x with roughly “how much” y is conjugate to x . This is done under the hypothesis that all minimizers connecting x to y are strongly normal, i.e. all pieces of the trajectory are not abnormal. Our result is a refinement of the one of Leandre $4t \log p_t(x, y) \rightarrow -d^2(x, y)$ for $t \rightarrow 0$, in which only the leading exponential term is detected. Our results are obtained by extending an idea of Molchanov from the Riemannian to the sub-Riemannian case, and some details we get appear to be new even in the Riemannian context. These results permit us to obtain properties of the sub-Riemannian distance starting from those of the heat kernel and vice versa. For the Grushin plane endowed with the Euclidean volume we get the expansion $p_t(x, y) \sim t^{-5/4} \exp(-d^2(x, y)/4t)$ where y is reached from a Riemannian point x by a minimizing geodesic which is conjugate at y .

1 Introduction

The heat kernel on sub-Riemannian manifolds has been an object of attention starting from the late 70s [13, 15, 16, 18, 20, 25, 31, 37, 40, 53, 54], as have the geodesics and cut and conjugate loci of such manifolds [2, 3, 5, 8, 9, 24, 28, 42, 51]. In this paper, we provide a general approach to relate the sub-Riemannian distance to the small time asymptotics of the heat kernel at the cut locus, at least in the case when there are no abnormal minimizers to the relevant point in the cut locus.

The problem of relating the sub-Riemannian distance to the heat-kernel is an old problem (see for instance [6, 13, 18, 21, 22, 23, 30, 36, 38, 39, 44, 50, 52]). In the following we recall some of the most relevant results. Let M be a n -dimensional smooth manifold provided with a complete sub-Riemannian structure, inducing a distance d , and also provided with a smooth volume μ . Let

This research has been supported by the European Research Council, ERC StG 2009 “GeCoMethods”, contract number 239748, by the ANR Project GCM, program “Blanche”, project number NT09-504490 and by the DIGITEO project CONGEO.

$p_t(x, y)$ the heat kernel of the sub-Riemannian heat equation $\partial_t \varphi = \Delta \varphi$, where Δ is the sub-Riemannian Laplacian defined as the divergence of the horizontal gradient. In particular Δ could be the sum of the squares of a choice of vector fields defining the sub-Riemannian distance (possibly with a first order term belonging to the distribution).

- **On the diagonal.** For some constant $C > 0$ (depending on the sub-Riemannian structure and x), we have

$$p_t(x, x) = \frac{C + O(\sqrt{t})}{t^{Q/2}}. \quad (1)$$

This result is due to Ben Arous and Leandre [19]. Here Q is the Hausdorff dimension of the sub-Riemannian manifold at x (see also [13]).

- **Off diagonal and off cut locus.** Fix $x \neq y$. If y is not in the cut locus of x and there are no abnormal from x to y , then for some constant $C > 0$ (depending on the sub-Riemannian structure, x , and y), one has

$$p_t(x, y) = \frac{C + O(t)}{t^{n/2}} e^{-d^2(x, y)/4t}.$$

This result is due to Ben Arous [18]. See also Taylor [54].

- **In any point of the space including the cut locus.**

$$\lim_{t \rightarrow 0} 4t \log p_t(x, y) = -d^2(x, y). \quad (2)$$

This result is due to Leandre [38, 39] (see also Taylor [54]). It is very general but is rougher than the one of Ben Arous. Roughly speaking it says that both on and off the cut locus, the leading term for $t \rightarrow 0$ has the form $e^{-d^2(x, y)/4t}$.

These results hold in particular in the Riemannian case. In that case we have $Q = n$ and formula (2) is the celebrated Varadhan formula obtained in [55].

In this paper we give a finer result with respect to the one of Leandre. We show that if y belongs to the cut locus of x and all minimizers connecting x and y are strongly normal (a minimizer is said to be *strongly normal* if every piece of it is not abnormal) then the rate of decay of $p_t(x, y)$ depends, roughly, on “how conjugate” x and y are, along the minimal geodesics connecting them. Intuitively, the more conjugate they are, the slower the decay. These results include Riemannian manifolds as a special case, for which they are completely general, since there are no abnormal minimizers in Riemannian geometry. Some details of the explicit relationship between the heat kernel asymptotics and the conjugacy of the minimal geodesics appears to be new even in the Riemannian context. Our results are also completely general for certain classes of sub-Riemannian geometries for which it is known there are no abnormal, such as contact manifolds and CR-manifolds. For a discussion of the presence of strictly abnormal minimizers in sub-Riemannian geometry one can see [27].

Our main result is Theorem 25 in Section 5, which relates the heat kernel asymptotics of $p_t(x, y)$ with what we call the *hinged energy function*

$$h_{x, y}(z) = \frac{1}{2}(d^2(x, z) + d^2(z, y)), \quad (3)$$

on the set of midpoints of all minimizing geodesics connecting x to y . To avoid overly complicated notation, we state here the following corollary which explains what happens in the case when the first terms of the Taylor expansion of $h_{x,y}$ have a simple expression.

Corollary 1. *Let M be an n -dimensional complete sub-Riemannian manifold provided with a smooth volume μ and let p_t be the heat kernel of the sub-Riemannian heat equation. Given distinct $x, y \in M$ let $h_{x,y}(z)$ be the hinged energy function. Assume that there is only one optimal geodesic joining x to y and that it is strongly normal, and let z_0 be the midpoint of the geodesic.*

Then $h_{x,y}(z)$ is smooth in a neighborhood of z_0 and attains its minimum at z_0 . Moreover, if there exists a coordinate system (z_1, \dots, z_n) around z_0 such that we have the expansion

$$h_{x,y}(z) = \frac{1}{4}d^2(x, y) + z_1^{2m_1} + \dots + z_n^{2m_n} + o(|z_1|^{2m_1} + \dots + |z_n|^{2m_n}),$$

for some integers $1 \leq m_1 \leq m_2 \leq \dots \leq m_n$ then for some constant $C > 0$ (depending on the sub-Riemannian structure, x , and y), one has

$$p_t(x, y) = \frac{C + o(1)}{t^{n - \sum_i \frac{1}{2m_i}}} \exp\left(-\frac{d^2(x, y)}{4t}\right).$$

From an analysis of the relation between the expansion of $h_{x,y}$ and the conjugacy of x and y we get

Corollary 2. *Let M be an n -dimensional complete sub-Riemannian manifold provided with a smooth volume μ and let p_t be the heat kernel of the sub-Riemannian heat equation. Let x and y be distinct and assume that every optimal geodesic joining x to y is strongly normal.*

Then there exist positive constants C_i , and t_0 (depending on M , x , and y) such that

$$(i) \quad \frac{C_1}{t^{n/2}} e^{-d^2(x,y)/4t} \leq p_t(x, y) \leq \frac{C_2}{t^{n-(1/2)}} e^{-d^2(x,y)/4t}, \quad \text{for } 0 < t < t_0.$$

(ii) If x and y are conjugate along at least one minimal geodesic connecting them, then

$$p_t(x, y) \geq \frac{C_3}{t^{(n/2)+(1/4)}} e^{-d^2(x,y)/4t}, \quad \text{for } 0 < t < t_0.$$

(iii) If x and y are not conjugate along any minimal geodesic joining them, then

$$p_t(x, y) = \frac{C_4 + O(t)}{t^{n/2}} e^{-d^2(x,y)/4t}, \quad \text{for } 0 < t < t_0.$$

Note that (iii) shows that the result of Ben Arous (1) holds not only off the cut locus, but also on the cut locus if x and y are not conjugate.

In the corollaries above the concept of sub-Riemannian manifold is quite general. It includes Riemannian manifolds and even sub-Riemannian manifolds which are rank-varying (see Sections 2 and Appendix A for the precise definition). The estimates (i) and (iii) were already known in Riemannian geometry (see [33] and [43] respectively) while (ii) appears to be new even in the Riemannian context.

The sub-Riemannian heat equation is intended with respect to the sub-Riemannian Laplacian which is defined as the divergence of the sub-Riemannian gradient. Here the divergence is computed

with respect to a smooth volume. In the equiregular case (see Definition 4) the most natural volume is Popp's volume, introduced by Montgomery in his book [44]. The hypothesis that the sub-Laplacian is computed with respect to a smooth volume is also essential. For rank-varying sub-Riemannian structures or for sub-Riemannian structure which are not equiregular, one could be tempted to define a sub-Laplacian containing diverging terms with the Popp volume (which is also diverging). This approach is possible. However it provides completely different results with respect to those presented in this paper. See for instance [22] for this approach in the case of the Grushin and Martinet structures.

In addition to these general bounds on the decay of $p_t(x, y)$ our approach provides a technique for computing the heat kernel asymptotics in concrete situations, subject, of course, to one's ability to determine explicit information about the minimal geodesics from x to y and the behavior of $h_{x,y}$ near their midpoints (which is related to the conjugacy of the minimal geodesics).

Conversely, these results allow us to realize the old idea of getting properties of the sub-Riemannian distance from those of the heat kernel (see for instance [6, 44]).

The hypothesis that all optimal geodesics connecting x to y are strongly normal is essential. In the case in which x is reached by y along an abnormal minimizer, it is not clear how to measure how much y is conjugate to x since abnormal extremals are not included in the exponential mapping and are in a sense isolated. The analysis of the heat kernel asymptotics in the presence of abnormal minimizers is an extremely difficult problem (also because of the lack of information about properties of the sub-Riemannian distance) and its study goes beyond the purpose of this paper.

Remark. Notice that in our approach we start from the sub-Riemannian structure (M, \mathcal{D}, g) , then we define an intrinsic volume, and finally we build the Laplace operator naturally associated with these data. This operator is by construction symmetric, negative and has the form $\Delta = \sum_{i=1}^k X_i^2 + X_0$ where $\{X_i\}_{i=1}^k$ define an orthonormal frame satisfying the Hörmander condition and $X_0 \in \text{span}\{X_i\}_{i=1}^k$.

In the literature one more often finds the reverse procedure [35, 50] (see also [16] and references therein). One starts from a second order differential operator with smooth coefficients \mathcal{L} which is symmetric and negative with respect to a volume μ , and then looks for a distance as a function of which one can give estimates of the fundamental solution of $\partial_t - \mathcal{L}$. This distance is constructed by introducing the so-called *sub-unit* curves for the operator, see for instance ([16, 21]). When \mathcal{L} is of the form $\mathcal{L} = \sum_{i=1}^k X_i^2 + X_0$ where $\{X_i\}_{i=1}^k$ are linearly independent vector fields satisfying the Hörmander condition, the symmetry with respect to μ implies that $X_0 \in \text{span}\{X_i\}_{i=1}^k$. Moreover the distance one gets is the sub-Riemannian distance for which $\{X_i\}_{i=1}^k$ is an orthonormal frame.

Also, let us mention that a wide literature is available about operators of the type $\mathcal{L} = \sum_{i=1}^k X_i^2 + X_0$ where $\{X_i\}_{i=1}^k$ satisfy the Hörmander condition, but $X_0 \notin \text{span}\{X_i\}_{i=1}^k$.

1.1 Structure of the paper

The structure of the paper is as follows. In Section 2 we introduce the concept of sub-Riemannian manifold. To avoid heavy notation, we have decided to restrict ourselves to the case in which the dimension of the distribution does not depend on the point. The rank-varying case is postponed to Appendix A. All the results of the paper holds also in this case.

In Sections 3 we state and prove a result expressing the heat kernel asymptotic as a Laplace integral over a neighborhood of the set of midpoints of minimal geodesics (see Theorem 20). In

Section 4 we discuss the asymptotics of Laplace type integrals and we discuss the relation between the degeneracy of the hinged energy function $h_{x,y}$ around the midpoints the minimal geodesics connecting x and y and the conjugacy of the minimal geodesics connecting them (see Theorem 22). Then in Section 5, we get our main general result, namely the estimates on the heat kernel $p_t(x, y)$ as a consequence of the previous analysis (see Theorem 25).

In Section 6 we apply our general results to some relevant cases. We briefly illustrate our results on the Heisenberg group for which both the optimal synthesis (i.e. the set of all optimal trajectories) from a given point and the heat kernel are known.

The second example is the nilpotent free (3, 6) case. In this case we get an asymptotic expansion on the vertical subspace (see Section 6.3), where all points are conjugate along minimal geodesics, which agrees with the fact that there exists a one parameter family of optimal geodesic reaching these points.

Finally in Section 7 we study the heat kernel in the Grushin plane, with respect to the standard Lebesgue measure. The Grushin structure is the rank-varying sub-Riemannian structure on the plane (x, y) such that $X = \partial_x$ and $Y = x\partial_y$ define an orthonormal frame. The corresponding sub-Laplacian is $\Delta = X^2 + Y^2$. Starting from the Riemannian point $q_0 = (-1, -\pi/4)$ we get, for the asymptotic at the point $q_1 = (1, \pi/4)$, which is reached from q_0 by a minimizing geodesic which is conjugate at q_1 , the expression $p_t(q_0, q_1) \sim t^{-5/4} \exp(-d^2(q_0, q_1)^2/4t)$, computing explicitly the degeneration of the hinged energy function. To our knowledge this is the first time in which an expansion of the type $t^{-\alpha} \exp(-d^2(q_0, q_1)/4t)$, with $\alpha \neq N/2$ for an integer N , is observed in the Riemannian or sub-Riemannian context.

2 Sub-Riemannian geometry

We start by recalling the definition of sub-Riemannian manifold in the case of a distribution of constant rank k smaller than the dimension of the space. For the more general definition of rank-varying sub-Riemannian structure (including as a particular case Riemannian structures) see Appendix A.

Definition 3. A *sub-Riemannian manifold* is a triple (M, \mathcal{D}, g) , where

- (i) M is a connected orientable smooth manifold of dimension $n \geq 3$;
- (ii) \mathcal{D} is a smooth distribution of constant rank $k < n$ satisfying the *Hörmander condition*, i.e. a smooth map that associates to $q \in M$ a k -dimensional subspace \mathcal{D}_q of T_qM such that

$$\text{span}\{[X_1, [\dots [X_{j-1}, X_j]]]_q \mid X_i \in \overline{\mathcal{D}}, j \in \mathbb{N}\} = T_qM, \quad \forall q \in M, \quad (4)$$

where $\overline{\mathcal{D}}$ denotes the set of *horizontal smooth vector fields* on M , i.e.

$$\overline{\mathcal{D}} = \{X \in \text{Vec}(M) \mid X(q) \in \mathcal{D}_q \quad \forall q \in M\}.$$

- (iii) g_q is a Riemannian metric on \mathcal{D}_q which is smooth as function of q . We denote the norm of a vector $v \in \mathcal{D}_q$ by $|v|_g = \sqrt{g_q(v, v)}$.

A Lipschitz continuous curve $\gamma : [0, T] \rightarrow M$ is said to be *horizontal* (or *admissible*) if

$$\dot{\gamma}(t) \in \mathcal{D}_{\gamma(t)} \quad \text{for a.e. } t \in [0, T].$$

Given an horizontal curve $\gamma : [0, T] \rightarrow M$, the *length* of γ is

$$\ell(\gamma) = \int_0^T |\dot{\gamma}(t)|_g dt. \quad (5)$$

Notice that $\ell(\gamma)$ is invariant under time reparametrization of the curve γ . The *distance* induced by the sub-Riemannian structure on M is the function

$$d(q_0, q_1) = \inf\{\ell(\gamma) \mid \gamma(0) = q_0, \gamma(T) = q_1, \gamma \text{ horizontal}\}. \quad (6)$$

The hypothesis of connectedness of M and the Hörmander condition guarantees the finiteness and the continuity of $d(\cdot, \cdot)$ with respect to the topology of M (Chow-Rashevsky theorem, see for instance [11]). The function $d(\cdot, \cdot)$ is called the *Carnot-Caratheodory distance* and gives to M the structure of a metric space (see [11]).

Locally, the pair (\mathcal{D}, g) can be given by assigning a set of k smooth vector fields spanning \mathcal{D} and that are orthonormal for g , i.e.

$$\mathcal{D}_q = \text{span}\{X_1(q), \dots, X_k(q)\}, \quad g_q(X_i(q), X_j(q)) = \delta_{ij}. \quad (7)$$

In this case, the set $\{X_1, \dots, X_k\}$ is called a *local orthonormal frame* for the sub-Riemannian structure.

The sub-Riemannian metric can also be expressed locally in “control form” as follows. We consider the control system,

$$\dot{q} = \sum_{i=1}^m u_i X_i(q), \quad u_i \in \mathbb{R}, \quad (8)$$

and the problem of finding the shortest curve that joins two fixed points $q_0, q_1 \in M$ is naturally formulated as the optimal control problem

$$\int_0^T \sqrt{\sum_{i=1}^m u_i^2(t)} dt \rightarrow \min, \quad q(0) = q_0, \quad q(T) = q_1 \neq q_0. \quad (9)$$

Definition 4. Define $\mathcal{D}^1 := \mathcal{D}$, $\mathcal{D}^{i+1} := \mathcal{D}^i + [\mathcal{D}^i, \mathcal{D}]$, for every $i \geq 1$. A sub-Riemannian manifold is said to be *equiregular* if for each $i \geq 1$, the dimension of \mathcal{D}_q^i does not depend on the point $q \in M$. For an equiregular sub-Riemannian manifold the Hörmander condition guarantees that there exists (a minimal) $m \in \mathbb{N}$, called *step* of the structure, such that $\mathcal{D}_q^m = T_q M$, for all $q \in M$. The sequence

$$\mathcal{G} := (\underbrace{\dim \mathcal{D}}_k, \underbrace{\dim \mathcal{D}^2}, \dots, \underbrace{\dim \mathcal{D}^m}_n),$$

is called the *growth vector* of the sub-Riemannian manifold. The growth vector permits us to compute the Hausdorff dimension of (M, d) as a metric space (see [41])

$$Q = \sum_{i=1}^m i k_i, \quad k_i := \dim \mathcal{D}^i - \dim \mathcal{D}^{i-1}. \quad (10)$$

In particular the Hausdorff dimension is always bigger than the topological dimension of M .

Definition 5. A sub-Riemannian manifold is said to be *nilpotent* if M is a nilpotent Lie group and the sub-Riemannian structure is left-invariant with respect to the group operation.

2.1 Minimizers and geodesics

In this section we briefly recall some facts about sub-Riemannian geodesics. In particular, we define the sub-Riemannian exponential map.

Definition 6. A *geodesic* for a sub-Riemannian manifold (M, \mathcal{D}, g) is an admissible curve $\gamma : [0, T] \rightarrow M$ such that $|\dot{\gamma}(t)|_g$ is constant and, for every sufficiently small interval $[t_1, t_2] \subset [0, T]$, the restriction $\gamma|_{[t_1, t_2]}$ is a minimizer of $\ell(\cdot)$. A geodesic for which $|\dot{\gamma}(t)|_g = 1$ is said to be parameterized by arclength.

A sub-Riemannian manifold is said to be *complete* if (M, d) is complete as a metric space. If the sub-Riemannian metric is the restriction to \mathcal{D} of a complete Riemannian metric, then it is complete.

Under the assumption that the manifold is complete, a version of the Hopf-Rinow theorem (see [26, Chapter 2]) implies that the manifold is geodesically complete (i.e. all geodesics are defined for every $t \geq 0$) and that for every two points there exists a minimizing geodesic connecting them.

Trajectories minimizing the distance between two points are solutions of first order necessary conditions for optimality, which in the case of sub-Riemannian geometry are given by a weak version of the Pontryagin Maximum Principle ([48]).

Theorem 7. Let $q(\cdot) : t \in [0, T] \mapsto q(t) \in M$ be a solution of the minimization problem (8),(9) such that $|\dot{q}(t)|_g$ is constant and $u(\cdot)$ be the corresponding control. Then there exists a Lipschitz map $p(\cdot) : t \in [0, T] \mapsto p(t) \in T_{q(t)}^*M \setminus \{0\}$ such that one and only one of the following conditions holds:

$$(i) \quad \dot{q} = \frac{\partial H}{\partial p}, \quad \dot{p} = -\frac{\partial H}{\partial q}, \quad u_i(t) = \langle p(t), X_i(q(t)) \rangle,$$

$$\text{where } H(q, p) = \frac{1}{2} \sum_{i=1}^k \langle p, X_i(q) \rangle^2.$$

$$(ii) \quad \dot{q} = \frac{\partial \mathcal{H}}{\partial p}, \quad \dot{p} = -\frac{\partial \mathcal{H}}{\partial q}, \quad 0 = \langle p(t), X_i(q(t)) \rangle,$$

$$\text{where } \mathcal{H}(t, q, p) = \sum_{i=1}^k u_i(t) \langle p, X_i(q) \rangle.$$

For an elementary proof of Theorem 7 see [3].

Remark 8. If $(q(\cdot), p(\cdot))$ is a solution of (i) (resp. (ii)) then it is called a *normal extremal* (resp. *abnormal extremal*). It is well known that if $(q(\cdot), p(\cdot))$ is a normal extremal then $q(\cdot)$ is a geodesic (see [3, 11]). This does not hold in general for abnormal extremals. An admissible trajectory $q(\cdot)$ can be at the same time normal and abnormal (corresponding to different covectors). If an admissible trajectory $q(\cdot)$ is normal but not abnormal, we say that it is *strictly normal*.

Abnormal extremals are very difficult to treat and many questions are still open. For instance it is not known if abnormal minimizers are smooth (see [44]).

Definition 9. A minimizer $\gamma : [0, T] \rightarrow M$ is said to be *strongly normal* if for every $[t_1, t_2] \subset [0, T]$, $\gamma|_{[t_1, t_2]}$ is not an abnormal minimizer.

In the following we denote by $(q(t), p(t)) = e^{t\bar{H}}(q_0, p_0)$ the solution of (i) with initial condition $(q(0), p(0)) = (q_0, p_0)$. Moreover we denote by $\pi : T^*M \rightarrow M$ the canonical projection.

Normal extremals (starting from q_0) parametrized by arclength correspond to initial covectors $p_0 \in \Lambda_{q_0} := \{p_0 \in T_{q_0}^*M \mid H(q_0, p_0) = 1/2\}$.

Definition 10. Let (M, \mathcal{D}, g) be a complete sub-Riemannian manifold and $q_0 \in M$. We define the *exponential map* starting from q_0 as

$$\mathcal{E}_{q_0} : \Lambda_{q_0} \times \mathbb{R}^+ \rightarrow M, \quad \mathcal{E}_{q_0}(p_0, t) = \pi(e^{t\tilde{H}}(q_0, p_0)). \quad (11)$$

Next, we recall the definition of cut and conjugate time.

Definition 11. Let $q_0 \in M$ and $\gamma(t)$ an arclength geodesic starting from q_0 . The *cut time* for γ is $t_{cut}(\gamma) = \sup\{t > 0, \gamma|_{[0,t]}$ is optimal $\}$. The *cut locus* from q_0 is the set $\text{Cut}(q_0) = \{\gamma(t_{cut}(\gamma)), \gamma$ arclength geodesic from $q_0\}$.

Definition 12. Let $q_0 \in M$ and $\gamma(t)$ a normal arclength geodesic starting from q_0 with initial covector p_0 . Assume that γ is not abnormal. The *first conjugate time* of γ is $t_{con}(\gamma) = \min\{t > 0, (p_0, t)$ is a critical point of $\mathcal{E}_{q_0}\}$. The (first) *conjugate locus* from q_0 is the set $\text{Con}(q_0) = \{\gamma(t_{con}(\gamma)), \gamma$ arclength geodesic from $q_0\}$.

It is well known that, for a geodesic γ which is not abnormal, the cut time $t_* = t_{cut}(\gamma)$ is either equal to the conjugate time or there exists another geodesic $\tilde{\gamma}$ such that $\gamma(t_*) = \tilde{\gamma}(t_*)$ (see for instance [9]).

Remark 13. In sub-Riemannian geometry, the exponential map starting from q_0 is never a local diffeomorphism in a neighborhood of the point q_0 itself. As a consequence the sub-Riemannian balls are never smooth and both the cut and the conjugate loci from q_0 are adjacent to the point q_0 itself (see [1]).

2.2 The sub-Laplacian

In this section we define the sub-Riemannian Laplacian on a sub-Riemannian manifold (M, \mathcal{D}, g) , provided with a smooth volume μ .

The sub-Laplacian is the natural generalization of the Laplace-Beltrami operator defined on a Riemannian manifold, defined as the divergence of the gradient.

The sub-Riemannian gradient can be defined with no difficulty. On a sub-Riemannian manifold (M, \mathcal{D}, g) , the gradient is the unique operator $\nabla : \mathcal{C}^\infty(M) \rightarrow \overline{\mathcal{D}}$ defined by

$$g_q(\nabla\varphi(q), v) = d\varphi_q(v), \quad \forall \varphi \in \mathcal{C}^\infty(M), q \in M, v \in \mathcal{D}_q.$$

By definition, the gradient is a horizontal vector field. If X_1, \dots, X_k is a local orthonormal frame, it is easy to see that it is written as follows $\nabla\varphi = \sum_{i=1}^k X_i(\varphi)X_i$, where $X_i(\varphi)$ denotes the Lie derivative of φ in the direction of X_i .

The divergence of a vector field X with respect to a volume μ is the function $\text{div } X$ defined by the identity $L_X\mu = (\text{div } X)\mu$, where L_X stands for the Lie derivative with respect to X .

The sub-Laplacian associated with the sub-Riemannian structure, i.e. $\Delta\varphi = \text{div}(\nabla\varphi)$, is written in a local orthonormal frame X_1, \dots, X_k as follows

$$\Delta = \sum_{i=1}^k X_i^2 + (\text{div } X_i)X_i. \quad (12)$$

Notice that Δ is always expressed as the sum of squares of the element of the orthonormal frame plus a first order term that belongs to the distribution and depends on the choice of the volume μ .

The existence of a smooth heat kernel for the operator (12), in the case of a complete sub-Riemannian manifold, is stated in [53].

2.2.1 Popp's volume and the intrinsic sub-Laplacian

In this section we recall how to construct an intrinsic Laplacian (i.e. that depends only on the sub-Riemannian structure) in the case of an equiregular sub-Riemannian manifold.

On a Riemannian manifold the Euclidean structure defined on the tangent space defines in a standard way a canonical volume: the Riemannian volume.

In the case of an equiregular sub-Riemannian manifold (M, \mathcal{D}, g) , even if there is no global scalar product defined in $T_q M$, it is possible to define an intrinsic volume, namely the Popp volume [44]. This is a smooth volume on M that is defined from the properties of the Lie algebra generated by the family of the horizontal vector fields. In the Riemannian case this coincide with the Riemannian volume.

On an equiregular manifold of dimension 3 the Popp volume is easily defined as $\nu_1 \wedge \nu_2 \wedge \nu_3$, where ν_1, ν_2, ν_3 is the dual basis to X_1, X_2 and $[X_1, X_2]$, where $\{X_1, X_2\}$ is any local orthonormal frame for the structure. This definition happens to be independent on the choice of X_1, X_2 . For the general definition see, e.g [6, 4].

Notice that the Popp volume is not the unique intrinsic volume that one can build from the geometric structure of (M, \mathcal{D}, g) . Since a sub-Riemannian manifold is a metric space (with the Carnot-Carathéodory distance), one can define the Q -dimensional Hausdorff measure on M , where Q is defined in (10). In contrast with the Riemannian case, starting from dimension 5, the Q -dimensional Hausdorff measure does not coincide in general with Popp's (see [4, 14] for details about these results).

The intrinsic sub-Laplacian is defined as the sub-Laplacian where the divergence is computed with respect to the Popp volume.

Remark 14. In the case of a left-invariant structure on a Lie group (and in particular for a nilpotent structure), the Popp volume is left-invariant, hence proportional to the left Haar measure.

For unimodular Lie groups, and in particular for nilpotent groups, one gets for the intrinsic sub-Laplacian the “sum of squares” form (see [6])

$$\Delta = \sum_{i=1}^k X_i^2.$$

Remark 15. To define the Popp volume the equiregularity assumption is crucial. In the non-equiregular case or in the rank-varying case the Popp volume diverges approaching the non-regular points, as do the coefficients of the intrinsic sub-Laplacian [6, 22].

3 General expression as a Laplace integral

From now on by a sub-Riemannian manifold we mean a structure in the sense of Section 2 or Appendix A, which include as a particular case Riemannian structures.

In the following we denote by $\Sigma \subset M \times M$ the set of pairs (x, y) with $x \neq y$ such that there exists a unique minimizing geodesic from x to y and such that this geodesic is strictly normal and not conjugate. Notice that Σ is an open set in $M \times M$ (see [10, 49] and [3, Chapter: Regularity of SR distance]).

Recall that the heat kernel is the the fundamental solution of the heat equation $\partial_t \varphi = \Delta \varphi$, where Δ is the sub-Riemannian Laplacian defined with respect to some smooth volume μ on a

sub-Riemannian manifold M . We begin by recalling the asymptotic expansion of the heat kernel away from the cut locus, due to Ben Arous [18] (see Theorem 3.1, and adjust for the fact that our heat kernel is for “ Δ ” rather than “ $\Delta/2$ ”).

Theorem 16. *Let M be an n -dimensional complete sub-Riemannian manifold in the sense of Section 2 or Appendix A, with a smooth volume μ and associated heat kernel p_t and let $(x, y) \in \Sigma$. Then for every non-negative integer m , we have the following asymptotic expansion as $t \searrow 0$:*

$$p_t(x, y) = \frac{1}{t^{n/2}} \exp\left(-\frac{d^2(x, y)}{4t}\right) \left(\sum_{j=0}^m c_j(x, y)t^j + O(t^{m+1})\right).$$

Here the c_i are smooth functions on Σ with $c_0(x, y) > 0$. Further, if $K \subset \Sigma$ is a compact set, then the expansion is uniform over K .

We will also need some preliminary control of the heat kernel at the cut locus, which is provided by a well-known result of Leandre [40]. In particular, Theorem 1 of [38] and Theorem 2.3 of [39] give (again taking into account our normalization of the heat kernel)

Theorem 17. *Let M be a complete sub-Riemannian manifold with a smooth volume μ and associated heat kernel p_t . For any compact subset K of $M \times M$, the following holds uniformly for $(x, y) \in K$:*

$$\lim_{t \searrow 0} 4t \log p_t(x, y) = -d^2(x, y).$$

Remark 18. Theorem 16 and 17 were originally stated in \mathbb{R}^n for sub-Riemannian metrics whose orthonormal frame consists of vector fields which are bounded with bounded derivatives. However, it is not hard to see that these results hold in the more general context of complete sub-Riemannian structures (where closed balls are compact).

Notation. In what follows we use sometimes the abbreviation $E(x, y) = d^2(x, y)/2$ for the energy function. For any two distinct points x and y , we let Γ be the set of midpoints of minimal geodesics from x to y . Further, we let $N(\Gamma)$ be a neighborhood of Γ , which we will feel free to make small enough to satisfy various assumptions. Finally, we let $h_{x,y}(z) = E(x, z) + E(z, y)$ be the hinged energy function. It’s clear from the definition that $h_{x,y}(z)$ is continuous.

Lemma 19. *The function $h_{x,y}$ attains its minimum exactly on Γ and $\min h_{x,y} = d^2(x, y)/4$.*

Proof. Let us consider a geodesic joining x and y and denote its midpoint by z_0 . We want to prove that $h_{x,y}(z_0) \leq h_{x,y}(z)$ for every z and that we have equality if and only if z is a midpoint of a geodesic joining x and y .

Let $a = d(x, z_0) = d(z_0, y)$, $b = d(x, z)$, and $c = d(y, z)$. By the triangle inequality, we have $2a \leq b + c$. Moreover we can assume that both b and c are less than or equal to $2a$, since otherwise the statement is trivial. Let $\varepsilon \geq 0$ be such that $2a + \varepsilon = b + c$ and compute

$$\begin{aligned} h_{x,y}(z) &= \frac{1}{2}(b^2 + c^2) = \frac{1}{2}((2a + \varepsilon - c)^2 + c^2) \\ &\geq a^2 + (a - c)^2 + \frac{\varepsilon^2}{2} + \varepsilon(2a - c) \geq a^2 = h_{x,y}(z_0) \end{aligned}$$

Moreover we have equality in the two inequalities if and only if $\varepsilon = 0$ and $a = c$, which is precisely the case where z is the midpoint of a geodesic joining x and y . Finally $h_{x,y}(z_0) = d^2(x, y)/4$. \square

We will need some basic assumptions about, and properties of, $N(\Gamma)$. First, basic properties of the distance function on M imply that Γ is compact. Next, all of our work will take place under the condition that we are “away from” any abnormal geodesics. In particular, assume that x and y are distinct and that every minimizer from x to y is a strongly normal geodesic. While we certainly allow y to be in the cut locus of x (which is a symmetric arrangement), the midpoint of every minimal geodesic from x to y will be a positive distance from the cut loci of both x and y .

More precisely, let $\lambda \in T_x^*M$ be a covector such that $\mathcal{E}_x(t\lambda)$ for $t \in [0, d(x, y)]$ parametrizes a minimal geodesic from x to y . (Here we adopt the convention that $\mathcal{E}_x(\lambda) = \pi \circ e^{\vec{H}}(x, \lambda)$.) Call this geodesic γ , and let z_0 be its midpoint. Since the cut time along γ is at least $d(x, y)$ and since the cut time is continuous as a function on T_x^*M near λ , it follows that \mathcal{E}_x is a diffeomorphism from a neighborhood U of $(d(x, y)/2)\lambda$ to a neighborhood U' of $z_0 = \mathcal{E}_x((d(x, y)/2)\lambda)$. Further, assuming U small enough, there is a unique minimal geodesic from x to each point $\mathcal{E}_x(\xi)$ in U' given by $\mathcal{E}_x(s\xi)$ for $s \in [0, 1]$, and this geodesic is not conjugate. In the case when $\xi = (d(x, y)/2)\lambda$, we have the “first half” of γ , from x to z_0 . Because γ is strongly normal, this piece of γ is strictly normal. Because the property of being strictly normal is an open condition on geodesics (see [10, 49] and [3, Chapter: Regularity of SR distance]), after possibly shrinking U and U' we have that all of the minimal geodesics from x to points in U' are also strictly normal.

One consequence is that by choosing U (and thus U') small enough, the distance function from x is smooth on U' (which we recall is some neighborhood of z_0 , the midpoint of a minimal geodesic γ from x to y). Another is that (after possibly further shrinking U and U') the Ben Arous expansion holds for $p_t(x, z)$ uniformly for $z \in U'$.

Note that the discussion in the previous paragraph also holds if we reverse the roles of x and y . Then, since Γ is compact, we see that for sufficiently small neighborhood $N(\Gamma)$ the distance functions from both x and y are smooth on $N(\Gamma)$ and the Ben Arous expansion holds for both $p_t(x, z)$ and $p_t(y, z)$ uniformly for $z \in N(\Gamma)$. It follows that $h_{x,y}$ is also smooth on $N(\Gamma)$. These are the key consequences of assuming that every minimizer from x to y is a strongly normal geodesic. We will also occasionally take advantage of the structure of the exponential map based at either x or y in a neighborhood of any point $z \in N(\Gamma)$. From now on, we will assume that, for such x and y , $N(\Gamma)$ is chosen in this way.

We now describe the main idea for determining the expansion on the cut locus. The intuition benefits from recalling that the heat kernel is also the transition density of Brownian motion on M . By the semi-group property (or the Markov property, from a stochastic point of view), a particle that travels from a point x to a point $y \neq x$ in time t first goes to some “halfway” point at time $t/2$, and then continues the rest of the way to y . For small t , a particle traveling from x to y is most likely to do so via a path which is approximately a geodesic (traversed at uniform speed). This is the usual intuition from large deviation theory. Thus, at time $t/2$, such a particle is likely to be near the midpoint of some minimal geodesic from x to y . The key insight, originally due to Molchanov [43] in the Riemannian case, is that, even in the case $y \in \text{Cut}(x)$, we can choose $N(\Gamma)$ as just discussed so that the expansion of Ben Arous can be applied to both the first and second halves of the particle’s journey from x to y (at least with high probability). The expansion at the cut locus is thus obtained by “gluing together” two copies of the Ben Arous expansion along the midpoints of the minimal geodesics from x to y . Making this argument precise, using only geometric analysis (we use stochastic notions only to bolster our intuition in the present paper), provides the proof of the next theorem. (The same “gluing idea” was employed to compute asymptotics of logarithmic derivatives in the Riemannian case in [46, 47].)

Theorem 20. *Let M be an n -dimensional complete sub-Riemannian manifold in the sense of Section 2 or Appendix A, and let x and y be distinct points such that all minimal geodesics from x to y are strongly normal. Then for $N(\Gamma)$ and c_0 as above there exists $\delta > 0$ such that*

$$p_t(x, y) = \int_{N(\Gamma)} \frac{2^n}{t^n} e^{-h_{x,y}(z)/t} (c_0(x, z)c_0(z, y) + O(t)) \mu(dz) + o\left(\exp\left[\frac{-E(x, y)/2 - \delta}{t}\right]\right).$$

Here the $O(t)$ term in the integral is uniform over $N(\Gamma)$.

Proof: By the semi-group property (or Chapman-Kolmogorov equation, for probabilists), we have

$$p_t(x, y) = \int_M p_{t/2}(x, z)p_{t/2}(z, y) \mu(dz).$$

We first divide M into two regions, $N(\Gamma)$ and $M \setminus N(\Gamma)$. As just discussed, both $p_t(x, \cdot)$ and $p_t(\cdot, y)$ are uniformly approximated by the Ben Arous expansion on $N(\Gamma)$ (since we assume that $\varepsilon > 0$ is sufficiently small). Using just the first term, we see that

$$p_t(x, y) = \int_{N(\Gamma)} \left(\frac{2}{t}\right)^n e^{-h_{x,y}(z)/t} (c_0(x, z)c_0(z, y) + O(t)) \mu(dz) + \int_{M \setminus N(\Gamma)} p_{t/2}(x, z)p_{t/2}(z, y) \mu(dz),$$

where the $O(t)$ terms are uniform over $N(\Gamma)$ by the uniformity of the Ben Arous expansion there.

Next, we estimate the integral over $M \setminus N(\Gamma)$. First, assume that M is compact. By Theorem 17, we have that, on M ,

$$p_t(u, v) = \exp\left[\frac{-d^2(u, v)/2 + r(t, u, v)}{2t}\right],$$

where $r(t, u, v)$ goes to zero uniformly with t on all of M . (In the remainder of the proof, we will use r to denote a function with this property, the exact definition of which may change from line to line.) We see that

$$p_{t/2}(x, z)p_{t/2}(z, y) = \exp\left[\frac{-h_{x,y}(z) + r(t, z)}{t}\right].$$

Further, the minimum of $h_{x,y}(z)$ on $M \setminus N(\Gamma)$ is strictly greater than $h_{x,y}(\Gamma) = E(x, y)/2$. Because $M \setminus N(\Gamma)$ has finite volume (by compactness), we see that there exists $\delta > 0$ such that

$$\int_{M \setminus N(\Gamma)} p_{t/2}(x, z)p_{t/2}(z, y) \mu(dz) = o\left(\exp\left[\frac{-E(x, y)/2 - \delta}{t}\right]\right). \quad (13)$$

Next, consider the case when M is not compact. Then for large enough R , we see that x , y , and $N(\Gamma)$ are all inside of $B_x(R)$ (the ball of radius R centered around x). We split the integral over $M \setminus N(\Gamma)$ into an integral over $B_R(x) \setminus N(\Gamma)$ and an integral over $M \setminus B_R(x)$. The previous argument can be applied to the integral over $B_R(x) \setminus N(\Gamma)$. Further, for large enough R , the integral over $M \setminus B_R(x)$ is also $o(\exp[(-E(x, y)/2 - \delta)/t])$. Thus Equation (13) holds in the case when M is non-compact as well. Combining these estimates completes the proof. \square

This theorem, in principle, gives the small-time asymptotics of the heat kernel in great generality. To get more concrete information, one needs to be able to determine the small-time asymptotics of the integral over $N(\Gamma)$. Fortunately, this is a well-studied type of integral, called a Laplace integral, as we shall discuss shortly.

Finally, we have stopped with the first term of the Ben Arous expansion only for convenience. As much of that expansion can be kept as desired, in which case the $c_0(x, z)c_0(z, y) + O(t)$ in the integrand is replaced by a more general product of Taylor series. However, it is unclear how much additional information this really provides. It seems that relatively little is known about the functions c_0 , and higher-order coefficients in the Ben Arous expansion are even less well-understood. Further, including such terms means that we would also want to determine higher-order terms in the asymptotic behavior of the Laplace integral over $N(\Gamma)$, which doesn't seem practical in general. For these reasons, we content ourselves with the leading term.

4 Understanding the Laplace integral

We wish to determine the asymptotics of the integral that appears in Theorem 20. To this end, we first review this type of integral, from which we see that the behavior of $h_{x,y}$ near Γ is the key factor. Then we discuss the geometric meaning of the behavior of $h_{x,y}$ in terms of the conjugacy of minimal geodesics from x to y .

4.1 A brief discussion of Laplace asymptotics

We now discuss techniques for determining the small t asymptotics of integrals of the type

$$\int_D f(x)e^{-g(x)/t} dx. \tag{14}$$

Here D is a compact set of \mathbb{R}^n having the origin in its interior, f is smooth in a neighborhood of D , and g is a function which is smooth in a neighborhood of D , is zero at the origin, and is strictly positive on D minus the origin. (We also assume the integral is with respect to Lebesgue measure; to treat any other measure with a smooth density we can simply incorporate the density into f .) Our assumption that g has zero as its minimum is no loss of generality; for any $a \in \mathbb{R}$ we have that

$$\int_D f(x)e^{-(g+a)/t} dx = e^{-a/t} \int_D f(x)e^{-g(x)/t} dx.$$

We start with the one-dimensional case. Further, we assume that g can be written as x^{2m} for some integer $m \geq 1$. Again, if this can be accomplished by first performing a smooth change of coordinates (and possibly shrinking D), we can just absorb the Jacobian into f . While it is not always possible to find such a change of coordinates, this is the most important case. Then (see, for example, [29])

$$\int_D f(x)e^{-x^{2m}/t} dx = f(0)\frac{\Gamma(1/(2m))}{m}t^{1/(2m)} + O\left(t^{3/(2m)}\right), \quad \text{as } t \searrow 0,$$

(here “ Γ ” is the usual Gamma function, not the set of midpoints of minimal geodesics). We note that higher terms in this expansion are known, but in the present context we continue to focus only on the leading term.

The higher dimensional situation is more complicated. If we assume that g can be written as

$$g(x) = \sum_{i=1}^n x_i^{2m_i},$$

for some integers $1 \leq m_1 \leq m_2 \leq \dots \leq m_n$, then the expansion essentially decomposes as a product of one-dimensional integrals. This immediately gives

$$\int_D f(x) e^{-g(x)/t} dx = t^{\frac{1}{2m_1} + \dots + \frac{1}{2m_n}} \left[f(0) \prod_{i=1}^n \frac{\Gamma(1/2m_i)}{m_i} + O\left(t^{1/m_n}\right) \right]. \quad (15)$$

In particular, if the Hessian of g is non-degenerate at the origin, the Morse lemma guarantees that we can always find coordinates near the origin in which g is a sum of squares, and thus the above expansion holds in these coordinates with $m_i = 1$ for all i . However, if the Hessian is degenerate, it will not necessarily be true that g can be put into the form of Equation (15) by a smooth change of coordinates.

Nonetheless, we recall that the “splitting lemma” for smooth functions (which can be found in [32]) allows us to split off non-degenerate directions and thus partially diagonalize g . In that spirit, the following result guarantees that, around an isolated degenerate critical point of corank 1, there always exists a coordinate set in which g is diagonal. It is a generalization of the classical Morse lemma for nondegenerate critical points and is a particular case of the splitting lemma just mentioned.

Lemma 21. *Let g be a smooth function on a neighborhood of the origin in \mathbb{R}^n , such that the origin is a local minimum of g and the only critical point of g . Assume that $g(0) = dg(0) = 0$ and that $\dim \ker d^2g(0) = 1$. Then there exists a diffeomorphism φ from a neighborhood of the origin to a neighborhood of the origin and a smooth function $\psi : \mathbb{R} \rightarrow \mathbb{R}$ such that*

$$g(\varphi(u)) = \sum_{i=1}^{n-1} u_i^2 + \psi(u_n), \quad \text{where} \quad \psi(u_n) = O(u_n^4).$$

More generally, suppose that g is equal to its Taylor series near the origin. Even this doesn't cover all possible cases (in particular, if g is smooth but not real-analytic), but it seems to be the most general case for which there is a satisfactory theory. In the case where g is equal to its Taylor series near the origin, Arnold and his collaborators (see [12] and the references therein) have given a powerful analysis of the resulting asymptotics. Briefly, if g is real-analytic with a unique minimum of zero at the origin then the leading term in the expansion (assuming $f(0) \neq 0$) is of the form $cf(0)t^\alpha |\log t|^m$ where c is a positive constant, α is a positive rational, and m is an integer between 0 and $n - 1$ inclusive. Estimates on α and m can be given in terms of combinatorial information derived from which monomials in the Taylor series of g have non-zero coefficients (more precisely, one looks at various features of the Newton diagram of g). Moreover, generically (in a sense which can be made precise) α and m are determined by this combinatorial information.

The above assumes that g has an isolated minimum at the origin. Suppose, instead, that g assumes its minimum along some smooth submanifold. In this case, one can choose coordinates for the minimum set and then extend them to coordinates near the minimum set by adding coordinates for the normal bundle. Then at each point of the minimum set, one can try to apply the above analysis to the corresponding fiber of the normal bundle, and then attempt to integrate the result

over the minimum set. The simplest such case is when g is a Morse-Bott function, in which case the asymptotics on each fiber will be just those corresponding to a non-degenerate Hessian of the appropriate dimension (this is what we see, for example, for the Heisenberg group in Section 6), although in general the situation can be more complicated.

In the case when this is not possible (for example, if the minimum set has a more complicated structure than a submanifold), a somewhat more general statement can be made. If g is real-analytic, one can use a resolution of singularities to reduce the situation to that of a sum of integrals of the form given in Equation (14), where in each term of the sum g is a monomial in the new coordinates and f is a smooth function times the absolute value of a monomial. (Essentially, the resolution of singularities amounts to a type of generalized change of coordinates under which g has this more restricted form.) The small-time asymptotics in such a case are again given by a rational power of t times an integer power of $|\log t|$. In contrast to the above case of an isolated minimum, here there does not seem to be a way of understanding the powers of t and $|\log t|$ without determining the resolution of singularities and computing the asymptotics of each of the resulting integrals.

The interested reader is referred to the references above for complete details, or to Sections 3.5 and 3.6 of [46] which contain a more detailed summary of these results (and which seems too much of a digression to repeat here).

4.2 Conjugacy and the behavior of $h_{x,y}$

We now discuss how the behavior of $h_{x,y}$ near its minima relates to the structure of the minimal geodesics from x to y , specifically, to the conjugacy of these geodesics. Suppose we have distinct points x and y such that every minimizer from x to y is strongly normal. We begin by introducing notation.

Consider any point $z_0 \in \Gamma$, which corresponds to some minimal geodesic γ from x to y . Then there is a unique covector $\lambda \in T_x^*M$ such that $\mathcal{E}_x(2\lambda) = y$ and that $\mathcal{E}_x(2\lambda, t)$ for $t \in [0, 1]$ parametrizes γ . (Recall that $\mathcal{E}_x(\lambda) = \pi \circ e^{\bar{H}}(x, \lambda)$.)

Let $\lambda(s)$ be a smooth curve of covectors $\lambda : (-\varepsilon, \varepsilon) \rightarrow T_x^*M$ (for some small $\varepsilon > 0$) such that $\lambda(0) = \lambda$ and the derivative never vanishes. Thus $\lambda(s)$ is a one-parameter family of perturbations of λ which realizes the first-order perturbation $\lambda'(0) \in T_\lambda(T_x^*M)$. Also, we let $z(s) = \mathcal{E}_x(\lambda(s))$, so that $z(0) = z_0$. Because \mathcal{E}_x is a diffeomorphism from a neighborhood of λ to a neighborhood of z_0 , we see that the derivative of $z(s)$ also never vanishes. Thus $z(s)$ is a curve which realizes the vector $z'(0) \in T_{z_0}M$. Further, we've established an isomorphism of the vector spaces $T_\lambda(T_x^*M)$ and $T_{z_0}M$ by mapping $\lambda'(0)$ to $z'(0)$, except that we've excluded the origin by insisting that both vectors are non-zero.

We say that γ is conjugate in the direction $\lambda'(0)$ (or with respect to the perturbation $\lambda'(0)$) if $\frac{d}{ds}\mathcal{E}(2\lambda(s))|_{s=0} = 0$. Note that this only depends on $\lambda'(0)$. We say that the Hessian of $h_{x,y}$ at z_0 is degenerate in the direction $z'(0)$ if $\frac{d^2}{ds^2}h_{x,y}(z(s))|_{s=0} = 0$. This last equality is equivalent to writing the Hessian of $h_{x,y}$ as a matrix in some smooth local coordinates, applying it as a quadratic form to $z'(0)$ expressed in these coordinates, and getting zero. This equivalence, as well the fact that whether the result is zero or not depends only on $z'(0)$, follows from the fact that z_0 is a critical point of $h_{x,y}$.

The point of the the next theorem is that conjugacy in the direction $\lambda'(0)$ is equivalent to degeneracy in the direction $z'(0)$. Thus the Hessian of $h_{x,y}$ encodes information about the conjugacy

of γ , and it is a more geometric object than it might seem at first.

Theorem 22. *Let M be a complete sub-Riemannian manifold in the sense of Section 2 or Appendix A, and let x and y be distinct points such that every minimal geodesic from x to y is strongly normal. Define Γ , $z_0 \in \Gamma$, $h_{x,y}$ and the curves $\lambda(s), z(s)$ as above. Then*

- (i) γ is conjugate if and only if the Hessian of $h_{x,y}$ at z_0 is degenerate.
- (ii) In particular γ is conjugate in the direction $\lambda'(0)$ if and only if the Hessian of $h_{x,y}$ at z_0 is degenerate in the corresponding direction $z'(0)$.
- (iii) The dimension of the space of perturbations for which γ is conjugate is equal to the dimension of the kernel of the Hessian of $h_{x,y}$ at z_0 .

Proof. We know that there is a unique shortest geodesic from y to $z(s)$ for all $s \in (-\varepsilon, \varepsilon)$, assuming ε small enough. Let $\eta(s)$ be the corresponding smooth curve of covectors in T_y^*M (that is, $\mathcal{E}_y(t\eta(s))$ for $t \in [0, 1]$ parametrizes the minimal geodesic from y to $z(s)$). Let $\tilde{\lambda}(s)$ and $\tilde{\eta}(s)$ be the images of λ and η , respectively, under the corresponding Hamiltonian flow on the cotangent bundle. We see that $\tilde{\lambda}(0) + \tilde{\eta}(0) = 0$.

Observe that $d(E(x, \cdot))|_{z(s)} = \tilde{\lambda}(s)$ and $d(E(y, \cdot))|_{z(s)} = \tilde{\eta}(s)$. (Here d stands for the differential.) It follows that

$$dh_{x,y}|_{z(s)} = \tilde{\lambda}(s) + \tilde{\eta}(s) \quad \left(\in T_{z(s)}^*M \right).$$

Next note that γ is conjugate in the direction $\lambda'(0)$ if and only if $\tilde{\lambda}(s) + \tilde{\eta}(s) = O(s^2)$, as follows directly from consideration of the exponential map. Thus γ is conjugate in the direction $\lambda'(0)$ if and only if $dh_{x,y}|_{z(s)} = O(s^2)$.

We claim that $dh_{x,y}|_{z(s)}$ is $O(s^2)$ if and only if $h_{x,y}(z(s))$ is $O(s^3)$. Equivalently, the derivative of $dh_{x,y}$ (as a one-form) in the $z'(0)$ direction is zero if and only if its pairing with $z'(0)$ is zero. This relationship is most easily expressed in local coordinates. Let H be the $n \times n$ matrix for the Hessian of $h_{x,y}$ at z_0 in some local coordinates, and let v be $z'(0)$ expressed in these coordinates. Then the derivative of $dh_{x,y}$ in the $z'(0)$ direction is Hv , which we think of an operator on vectors u by $\langle Hv, u \rangle$ where is the standard Euclidean inner product for these coordinates, or equivalently by $u^T Hv$ where u and v are written as column-vectors.

The claim now follows from the following simple fact from linear algebra: for any symmetric and positive semi-definite $n \times n$ real matrix A , we have that, for any $x \in \mathbb{R}^n$, $\langle Ax, x \rangle = 0$ if and only if $Ax = 0 \in \mathbb{R}^n$, where $\langle \cdot, \cdot \rangle$ is the standard Euclidean inner product. Because A is symmetric, we can find an orthonormal basis v_i , $i = 1, \dots, n$ for \mathbb{R}^n consisting of eigenvectors of A with corresponding eigenvalues $\lambda_i \geq 0$. Then, writing $x = \sum_{i=1}^n x_i v_i$, the above fact follows from the identities $Ax = \sum_{i=1}^n \lambda_i x_i v_i$ and $\langle Ax, x \rangle = \sum_{i=1}^n \lambda_i x_i^2$.

Since the Hessian of $h_{x,y}$ at z_0 is symmetric and positive semi-definite, the claim follows. Thus we have proven statement (ii) in the theorem (and (i) a fortiori), namely that γ is conjugate in the direction $\lambda'(0)$ if and only if the Hessian of $h_{x,y}$ at z_0 is degenerate in the direction $z'(0)$.

Statement (iii) is an immediate consequence of (ii) plus the fact that the correspondence between $\lambda'(0)$ and $z'(0)$ gives an isomorphism of vector spaces between $T_\lambda(T_x^*M)$ and $T_{z_0}M$, as discussed just before the theorem. \square

We now briefly discuss the situation of higher-order derivatives of the exponential map and higher-order derivatives of $h_{x,y}$. This situation is more complicated than what we just saw for lower-order derivatives.

Recall that $\mathcal{E}_x(2\lambda) = y$. Further, consider

$$\frac{d^m}{ds^m} \mathcal{E}_x(2\lambda(s))|_{s=0}.$$

For $m = 1$, this is zero if and only if y is conjugate to x along the geodesic through z_0 in the direction of $\lambda'(0)$. If this first derivative is zero, then the number of higher order derivatives which vanish describes, in a sense, how conjugate y is to x with respect to the perturbation $2\lambda(s)$. (Of course, it's possible for all derivatives to vanish; for example, if $2\lambda(s)$ describes a one-parameter family of minimal geodesics from x to y , as occurs for the Heisenberg group.) We can compare the vanishing of these derivatives to the vanishing of the derivatives $\frac{d^k}{ds^k} h_{x,y}(z(s))$.

Suppose that, for some positive integer m , we have that $\mathcal{E}_x(2\lambda(s)) = ws^m + O(s^{m+1})$ for some non-zero $w \in \mathbb{R}^n$ in some (smooth) system of coordinates around y (so that y is at the origin of these coordinates). If this holds in one such system, it holds in any other such system with w re-expressed in the new coordinates. Because the exponential map is a diffeomorphism from a neighborhood of each $\tilde{\eta}(s) \in T_{z(s)}^*M$ to a neighborhood of y , we see that this expansion for $\mathcal{E}_x(2\lambda(s))$ is equivalent to having

$$\tilde{\lambda}(s) + \tilde{\eta}(s) = dh_{x,y}|_{z(s)} = vs^m + O(s^{m+1}) \in T_{z(s)}^*M$$

for some non-zero one-form v written with respect to some (smooth) system of coordinates around z_0 . Again, if this holds for one such system, it holds for any other such system with v re-expressed relative to the new coordinates.

Thus, the one-form $dh_{x,y}|_{z(s)}$ vanishes to the same order as the derivatives of $\mathcal{E}_x(2\lambda(s))$. However, when we look at $h_{x,y}(z(s))$, we see that

$$\begin{aligned} h_{x,y}(z(s)) - h_{x,y}(z_0) &= \int_0^s dh_{x,y}|_{z(t)}(z'(t)) dt = \\ &= \int_0^s (v(z'(t))t^m + O(t^{m+1})) dt = \frac{1}{m+1}v(z'(0))s^{m+1} + O(s^{m+2}). \end{aligned}$$

So we have that $h_{x,y}(z(s)) - h_{x,y}(z_0) = cs^{m+1} + O(s^{m+2})$ for non-zero c if and only if $v(z'(0)) \neq 0$. For $m = 1$, it is always the case that $v(z'(0)) \neq 0$, as we saw in the previous theorem. However, for $m > 1$, this need not be true. In such a case we can only conclude that $h_{x,y}(z(s)) - h_{x,y}(z_0) = O(s^{m+2})$, and the exact order of vanishing of the derivatives of $h_{x,y}$ is unclear in general.

Remark 23. In the special case when M is two-dimensional and z_0 is an isolated minimum such that $h_{x,y}$ vanishes to finite order at z_0 , the relationship is simpler. Namely, because the Hessian of $h_{x,y}$ is clearly non-degenerate along the direction of γ , we can apply Lemma 21 to write $h_{x,y} = u_1^2 + g(u_2)$ for some coordinates u_i around z_0 and some smooth function g . Then if $z(s)$ corresponds to the curve $(u_1, u_2) = (0, s)$, we see by direct computation that $v(z'(0)) \neq 0$. In this way, the degree of degeneracy of the Hessian and the degree of conjugacy correspond precisely in this case.

We also note that, at the opposite extreme, there is again a nice correspondence between the behavior of the exponential map and of $h_{x,y}$. Namely, $\mathcal{E}_x(2\lambda(s)) = y$ for all $s \in (-\varepsilon, \varepsilon)$ if and only if $h_{x,y}(z(s)) = h_{x,y}(z_0)$ for all $s \in (-\varepsilon, \varepsilon)$, as follows directly from Lemma 19.

All of this seems to indicate that, loosely speaking, the more conjugate the geodesic through z_0 is, the more degenerate $h_{x,y}$ is at z_0 . However, it also seems that looking at curves through z_0 corresponding to one-parameter perturbations of the geodesic is too naive in general, and that a more sophisticated approach is needed to describe the exact relationship between the higher order

derivatives of the exponential map and higher order terms in the Taylor series of $h_{x,y}$. As we do not need anything beyond the results of Theorem 22 in what follows (except perhaps to give geometric intuition to $h_{x,y}$), we do not pursue this direction any further. (In light of the above, it seems that the claims about higher-order derivatives in Lemma 3.1 of [46] are over-simplified. Fortunately, in that paper, as in the present, only the content of Theorem 22 is used in subsequent arguments. The higher-order relationship serves only to provide a more geometric meaning to the behavior of $h_{x,y}$.)

5 General consequences of Laplace asymptotics

We are now in a position to see what the theory of Laplace asymptotics summarized in the previous section gives when applied to the integral in Theorem 20. The ideas expand upon those of Section 5.3 of [33], where inequality (18) of Theorem 25 is given in the Riemannian case. We note that the results we give in this section, most interestingly those which depend on whether or not x and y are conjugate along a given geodesic, are also valid in the Riemannian case.

We begin with a basic lemma. For this lemma, we say that the u_i are “coordinates around z_0 ” if they are coordinates on some neighborhood of z_0 such that $u_i(z_0) = 0$ for all i . Inequalities for such coordinates are understood to hold on some such neighborhood.

Lemma 24. *Under the assumptions of Theorem 22, let z_0 be any point of Γ . Then there exist smooth coordinates u_1, \dots, u_n around z_0 such that*

$$h_{x,y}(u_1, \dots, u_n) \geq \frac{1}{4}d^2(x, y) + u_1^2. \quad (16)$$

Also, there exist smooth coordinates v_1, \dots, v_n around z_0 such that

$$h_{x,y}(v_1, \dots, v_n) \leq \frac{1}{4}d^2(x, y) + v_1^2 + \dots + v_n^2.$$

Finally, if the geodesic from x to y passing through z_0 is conjugate, then the v_i can be chosen so that

$$h_{x,y}(v_1, \dots, v_n) \leq \frac{1}{4}d^2(x, y) + v_1^2 + \dots + v_{n-1}^2 + v_n^4.$$

Proof. For z in some neighborhood of z_0 , let $u_1(z) = u_1 = d(x, z) - (d(x, y)/2)$. If the neighborhood is small enough, this is a smooth function with non-vanishing derivative (since $d(x, z)$ has these properties as a function of z) and $u_1(z_0) = 0$. Thus it is a valid coordinate, and we can complete this to a full set of coordinates around z_0 . Further, the triangle inequality gives

$$d(y, z) \geq d(x, y) - d(x, z) = \frac{1}{2}d(x, y) - u_1.$$

Thus we compute

$$\begin{aligned} h_{x,y}(z) &= \frac{1}{2} [d(x, z)^2 + d(y, z)^2] \\ &\geq \frac{1}{2} \left[\left(u_1 + \frac{1}{2}d(x, y) \right)^2 + \left(\frac{1}{2}d(x, y) - u_1 \right)^2 \right] = u_1^2 + \frac{1}{4}d(x, y)^2. \end{aligned}$$

which gives the estimate (16).

For the second estimate, recall that $h_{x,y}$ is smooth and assumes its minimum at z_0 , and thus the derivative of $h_{x,y}$ vanishes at z_0 . It follows that for any system of coordinates w_i around z_0 , there is a small enough neighborhood of z_0 and positive constant C such that

$$h_{x,y}(v_1, \dots, v_n) \leq \frac{1}{2}E(x, y) + C(w_1^2 + \dots + w_n^2)$$

on this neighborhood. Thus, we can simply rescale the w_i to get coordinates v_i as required.

The proof of the final inequality is based on Lemma 21.

Because the geodesic through z_0 is conjugate, the Hessian of $h_{x,y}$ at z_0 cannot have full rank, as we see from Theorem 22. First assume that the rank is exactly $n - 1$. Then Lemma 21 shows that there are coordinates v_i around z_0 such that

$$h_{x,y}(v_1, \dots, v_n) = \frac{1}{2}E(x, y) + v_1^2 + \dots + v_{n-1}^2 + O(v_n^4)$$

Then, after possibly rescaling v_n , we see that the desired inequality holds.

Next assume the Hessian of $h_{x,y}$ has rank less than $n - 1$. Then let φ be a smooth function on a neighborhood of z_0 which is non-negative, zero at z_0 (hence with vanishing derivative at z_0), and such that $h_{x,y} + \varphi$ has Hessian of rank $n - 1$ at z_0 (φ looks like a sum of squares of an appropriate number of coordinates, for example). Applying the previous result to $h_{x,y} + \varphi$ shows that there are coordinates v_i around z_0 such that

$$(h_{x,y} + \varphi)(v_1, \dots, v_n) \leq \frac{1}{2}E(x, y) + v_1^2 + \dots + v_{n-1}^2 + v_n^4.$$

Since φ is non-negative, the desired estimate for $h_{x,y}$ follows. \square

These estimates allow us to say more about the integral appearing in Theorem 20.

Theorem 25. *With the same assumptions and notation as Theorem 20, we have that for any sufficiently small neighborhood $N(\Gamma)$*

$$p_t(x, y) = \int_{N(\Gamma)} \left(\frac{2}{t}\right)^n e^{-h_{x,y}(z)/t} (c_0(x, z)c_0(z, y) + O(t)) \mu(dz). \quad (17)$$

Again, the “ $O(t)$ ” term in the integral is uniform over $N(\Gamma)$. Also, there exist positive constants C_i , and t_0 (depending on M , x , and y) such that

$$\frac{C_1}{t^{n/2}} e^{-d^2(x,y)/4t} \leq p_t(x, y) \leq \frac{C_2}{t^{n-(1/2)}} e^{-d^2(x,y)/4t} \quad (18)$$

for $0 < t < t_0$. Further, if x and y are conjugate along any minimal geodesic connecting them, then (perhaps after changing t_0), we have

$$p_t(x, y) \geq \frac{C_3}{t^{(n/2)+(1/4)}} e^{-d^2(x,y)/4t} \quad (19)$$

for $0 < t < t_0$. Finally, if x and y are not conjugate along any minimal geodesic joining them, then

$$p_t(x, y) = \frac{C_4 + O(t)}{t^{n/2}} e^{-d^2(x,y)/4t}. \quad (20)$$

Remark 26. Notice that Corollary 1 in the Introduction is a direct consequence of formulas (17) and (15). Corollary 2 contains the estimates (18), (19) and (20).

Proof. We begin with the general bounds on $p_t(x, y)$. Choose any $z_0 \in \Gamma$ and let $V \subset N(\Gamma)$ be some neighborhood on which there are coordinates $v = (v_1, \dots, v_n)$ as in the previous lemma (that is, $h_{x,y}$ is estimated by the sum of squares of the v_i). Since the integrand in Theorem 20 is positive for sufficiently small t , we have that

$$\begin{aligned} \int_{N(\Gamma)} \left(\frac{2}{t}\right)^n e^{-h_{x,y}(z)/t} (c_0(x, z)c_0(z, y) + O(t)) \mu(dz) \\ \geq \left(\frac{2}{t}\right)^n e^{-E(x,y)/2t} \int_V e^{-(v_1^2 + \dots + v_n^2)/t} (c_0(x, v)c_0(v, y) + O(t)) \mu(dv) \end{aligned}$$

for all sufficiently small, positive t .

Because μ is a smooth volume and v a smooth coordinate system, we know that there is a smooth, positive function F such that $\mu(dv) = F(v)dv_1 \cdots dv_n$. Then the results of the previous section, namely Equation (15), show that

$$\begin{aligned} \int_V e^{-(v_1^2 + \dots + v_n^2)/t} (c_0(x, v)c_0(v, y) + O(t)) F(v) dv_1 \cdots dv_n \\ = t^{n/2} \left[F(0) (c_0(x, z_0)c_0(z_0, y) + O(t)) \pi^{n/2} + O(t) \right] \end{aligned}$$

(where we've used that $\Gamma(1/2) = \sqrt{\pi}$). Note that there's no difficulty handling the $O(t)$ in the integrand since we simply estimate it by $|O(t)| \leq Ct$ for some positive C and factor the t out of the integral. Putting this together with the fact that $F(0)c_0(x, z_0)c_0(z_0, y)$ is positive, we see that there exist positive C_1 and t_0 such that

$$\int_{N(\Gamma)} \left(\frac{2}{t}\right)^n e^{-h_{x,y}(z)/t} (c_0(x, z)c_0(z, y) + O(t)) \mu(dz) \geq \frac{C_1}{t^{n/2}} e^{-E(x,y)/2t}$$

for $0 < t < t_0$. Comparing this to Theorem 20, we note that the $o\left(\exp\left[\frac{-E(x,y)/2-\delta}{t}\right]\right)$ term is dominated by the right-hand side of the above inequality. Thus, after possibly adjusting C_1 and t_0 , we see that the relevant inequality in the theorem holds.

For the other side of the first inequality, note that we can find coordinates u_i as in the previous lemma around every point of Γ , and each of these systems of coordinates is defined on some open neighborhood. Because Γ is compact, there is a finite set of such neighborhoods which cover Γ ; denote them by U_1, \dots, U_m and the corresponding systems of coordinates by $u_j = (u_{j,1}, \dots, u_{j,n})$ for $j = 1, \dots, m$. Now choose $N(\Gamma)$ small enough so that $N(\Gamma) \subset \cup_{j=1}^m U_j$. Then we have

$$\begin{aligned} \int_{N(\Gamma)} \left(\frac{2}{t}\right)^n e^{-h_{x,y}(z)/t} (c_0(x, z)c_0(z, y) + O(t)) \mu(dz) \\ \leq \sum_{j=1}^m \left(\frac{2}{t}\right)^n e^{-E(x,y)/2t} \int_{U_j} e^{-u_{j,1}^2/t} (c_0(x, u_j)c_0(u_j, y) + O(t)) \mu(du_j) \end{aligned}$$

for all sufficiently small, positive t . As above, μ is a smooth volume, and Equation (15) gives that, for each j , there is a positive constant K_j such that

$$\int_{U_j} e^{-u_{j,1}^2/t} (c_0(x, u_j)c_0(u_j, y) + O(t)) \mu(du_j) = \sqrt{t} (K_j + O(t)).$$

Summing j from 1 to m allows us to conclude that there exists positive C_2 such that, after possibly making t_0 smaller,

$$\int_{N(\Gamma)} \left(\frac{2}{t}\right)^n e^{-h_{x,y}(z)/t} (c_0(x, z)c_0(z, y) + O(t)) \mu(dz) \leq \frac{C_2}{t^{n-(1/2)}} e^{-E(x,y)/2t}$$

for $0 < t < t_0$. Again, comparing this to Theorem 20, we see that the other side of the first inequality in the theorem holds, after possibly adjusting C_2 and t_0 .

The two-sided inequality we've just proved now shows that the term $o\left(\exp\left[\frac{-E(x,y)/2-\delta}{t}\right]\right)$ in Theorem 20 is unnecessary; it can be "included" in the $O(t)$ term in the integral (as we've already taken advantage of above). This establishes the first claim in the theorem.

Now we consider the case when x and y are conjugate along some minimal geodesic. Suppose that z_0 is the midpoint of this geodesic. Then we can find coordinates v_i around z_0 , defined on some neighborhood $V \subset N(\Gamma)$, such that

$$h_{x,y}(v_1, \dots, v_n) \leq \frac{1}{2}E(x, y) + v_1^2 + \dots + v_{n-1}^2 + v_n^4.$$

Analogous to the previous lower bound, we have that

$$\begin{aligned} \int_{N(\Gamma)} \left(\frac{2}{t}\right)^n e^{-h_{x,y}(z)/t} (c_0(x, z)c_0(z, y) + O(t)) \mu(dz) \\ \geq \left(\frac{2}{t}\right)^n e^{-\frac{E(x,y)}{2t}} \int_V e^{-(v_1^2 + \dots + v_{n-1}^2 + v_n^4)/t} (c_0(x, v)c_0(v, y) + O(t)) \mu(dv), \end{aligned}$$

for all sufficiently small, positive t . Equation (15) (along with smoothness of μ and positivity of the c_0) then shows that, for some positive constants C_3 and t_0 (possibly different from before),

$$\begin{aligned} \int_V e^{-(v_1^2 + \dots + v_{n-1}^2 + v_n^4)/t} (c_0(x, v)c_0(v, y) + O(t)) \mu(dv) \\ \geq t^{(n/2)-(1/4)} \left(C_3 + O\left(\sqrt{t}\right)\right), \end{aligned}$$

for $0 < t < t_0$. Combining these estimates and the first claim in the theorem, we see that, after possibly adjusting C_3 and t_0 ,

$$p_t(x, y) \geq \frac{C_3}{t^{(n/2)+(1/4)}} e^{-E(x,y)/2t},$$

for $0 < t < t_0$.

Finally, we suppose that x and y are not conjugate along any minimal geodesic joining them. Then for any $z_0 \in \Gamma$, Theorem 22 and the Morse lemma imply that z_0 is isolated. Since Γ is compact, we see that in fact Γ consists of finitely many points, say z_1, \dots, z_m (so there are only

finitely many minimal geodesics from x to y). Further, we can find coordinates $u_{j,1}, \dots, u_{j,n}$ around each z_j , on some neighborhood U_j , such that

$$h_{x,y}(u_{j,1}, \dots, u_{j,n}) = \frac{1}{2}E(x, y) + u_{j,1}^2 + \dots + u_{j,n}^2 \quad \text{on } U_j,$$

and $N(\Gamma)$ is the disjoint union of the U_j (for small enough U_j). Thus, using the first claim in the theorem,

$$\begin{aligned} p_t(x, y) &= \left(\frac{2}{t}\right)^n e^{-E(x,y)/2t} \sum_{j=1}^m \int_{U_j} e^{-(u_{j,1}^2 + \dots + u_{j,n}^2)/t} \times \\ &\quad \times c_0(x, u_j) c_0(u_j, y) + O(t) \mu(du_j). \end{aligned} \quad (21)$$

We have that $\mu(du_j) = F_j(u_j) du_{j,1} \cdots du_{j,n}$ for smooth, positive F_j . As above, we compute

$$\begin{aligned} \int_{U_j} e^{-(u_{j,1}^2 + \dots + u_{j,n}^2)/t} (c_0(x, u_j) c_0(u_j, y) + O(t)) F_j(u_j) du_{j,1} \cdots du_{j,n} \\ = t^{n/2} \left[F_j(0) (c_0(x, z_j) c_0(z_j, y) + O(t)) \pi^{n/2} + O(t) \right]. \end{aligned}$$

Summing over j , we have

$$p_t(x, y) = \frac{C_4 + O(t)}{t^{n/2}} e^{-E(x,y)/2t},$$

where $C_4 = (4\pi)^{n/2} \sum_{j=1}^m F_j(0) c_0(x, z_j) c_0(z_j, y)$, which is clearly positive. \square

One consequence of this result is that the exponent of $1/t$ in the small-time expansion of $p_t(x, y)$ “sees” whether or not x and y are conjugate along any minimal geodesic. Said differently, the exponent of t detects the part of the cut locus of x which comes from conjugacy (assuming that the necessary geodesics are strictly normal, of course). That naturally leads to the question of what happens at cut points which are not conjugate.

We first note that, if y is not in the cut locus of x , then the results of this analysis fit nicely with the expansion of Ben Arous, which applies in a neighborhood of y . In this case, there is a single minimal geodesic from x to y and it is not conjugate. Let z_1 be the midpoint. Then the same analysis as in the last part of the previous proof (just with $m = 1$) shows that

$$p_t(x, y) = (F(z_1) c_0(x, z_1) c_0(z_1, y) + O(t)) \frac{(4\pi)^{n/2}}{t^{n/2}} e^{-E(x,y)/2t},$$

where $F(z_1)$ is the density of μ with respect to coordinates which make the Hessian of $h_{x,y}$ at z_1 the identity matrix. Since the Ben Arous expansion applies to $p_t(x, y)$, we also have

$$p_t(x, y) = (c_0(x, y) + O(t)) \frac{1}{t^{n/2}} e^{-E(x,y)/2t}.$$

So in this case, Theorem 25 provides a relationship between $c_0(x, y)$ on the one hand, and $c_0(x, z_1)$, $c_0(z_1, y)$, and second-order behavior of $h_{x,y}$ at z_1 (which is encoded by $F(z_1)$) on the other.

Now suppose that y is in the cut locus of x , but that none of the minimal geodesics from x to y are conjugate (and the assumptions of Theorem 25 hold, of course). Let $\gamma_1(s)$ be one such

geodesic, parametrized by arc-length so that $\gamma_1(0) = x$ and $\gamma_1(d(x, y)) = y$. Then we claim that $\lim_{s \nearrow d(x, y)} c_0(x, \gamma_1(s))$ exists and is positive, and we denote it α_1 . This follows from the relationship between $c_0(x, \gamma_1(s))$ and $c_0(x, \gamma_1(s/2))$, $c_0(\gamma_1(s/2), y)$, and $F(\gamma_1(s/2))$ just discussed, and that fact that these last three quantities are continuous in s and remain positive. (Indeed, we've already seen in the proof of Theorem 25 that $\alpha_1 = (4\pi)^{n/2} F(z_1) c_0(x, z_1) c_0(z_1, y)$ where $z_1 = \gamma_1(d(x, y)/2)$.) Alternatively, one can think of lifting a neighborhood of $\gamma_1([0, s])$ to a "local" universal cover and then applying the Ben Arous expansion.

Continuing, we let $\gamma_2, \dots, \gamma_m$ be the other minimal geodesics from x to y , where we know that there can only be finitely many and that m must be at least 2. We let $\alpha_2, \dots, \alpha_m$ be the associated limits of $c_0(x, \cdot)$ along these geodesics, analogous to α_1 . Then the final part of the proof of Theorem 25 shows that

$$p_t(x, y) = \left[\sum_{j=1}^m \alpha_j + O(t) \right] \frac{1}{t^{n/2}} e^{-E(x, y)/2t}.$$

The point of relating the coefficient of $t^{-n/2} e^{-d^2(x, y)/4t}$ in the above to the c_0 along the γ_j is that we see that this coefficient is discontinuous at y . That is, for any γ_j , we know that $c_0(x, \gamma_j(s))$ is continuous in a neighborhood of $\gamma_j(s)$ as long as $0 < s < d(x, y)$. However, when s increases to $d(x, y)$, the value of this coefficient "jumps up" to the sum of the α_j . Thus, points which are not in the cut locus of x and points that are but are not conjugate to x along any minimal geodesics both have small-time heat kernel expansions that look like a constant times $t^{-n/2} e^{-d^2(x, y)/4t}$. These two types of points can be distinguished by whether or not the coefficient (the constant) is continuous at the point in question. However, if one has that much information about the small-time heat kernel asymptotics in a neighborhood of a point y , then presumably one already understands $d(x, \cdot)$ near y , from which one should be able to understand the local structure of the cut locus. Thus looking at this coefficient, from the perspective of locating the cut locus, seems unlikely to be of much help.

This potentially stands in contrast to the case when y is conjugate to x along a minimal geodesic, in which case only the power of t appearing in the expansion at the point y needs to be determined (in order to conclude that y is conjugate to x along a minimal geodesic).

6 Examples

In this section we discuss our results in some examples of 2-step sub-Riemannian structures. In these cases, an integral expression of the heat kernel (which can be explicitly written in some cases) has been found in [17].

In the first example, namely the Heisenberg group, we briefly compute the Hessian of the hinged energy function $h_{x, y}$ when x is the origin and y is a point on the cut locus. In this case, being that both the optimal synthesis and the heat kernel known explicitly, we verify the results of Theorem 25.

The second example is the free nilpotent sub-Riemannian structure with growth vector (3,6). Here we use a "reverse" argument, starting from the formula for the heat kernel to find the asymptotics for points belonging to the vertical subspace, where all points are both cut and conjugate. This asymptotic agrees with the fact there exists a one parameter family of optimal geodesics that reach this point (for a detailed discussion about the optimal synthesis see [45]).

In this section the heat kernel is meant for the intrinsic sub-Laplacian, i.e. it is computed with respect to the Popp volume. For the cases treated in this section this volume is proportional to the left Haar measure and is proportional to the Lebesgue measure in the standard system of coordinates we are using.

6.1 Formula for the heat kernel in the 2-step case

In this section we recall the expression of the heat kernel of the intrinsic sub-Laplacian associated with a 2-step nilpotent structure, that has been found in [17]. Then we rewrite it to have a convenient expression on the “vertical subspace”.

Consider on \mathbb{R}^n a 2-step nilpotent structure of rank $k < n$, where X_1, \dots, X_k is an orthonormal frame. Once a smooth complement \mathcal{V} for the distribution is chosen (i.e. $T_q\mathbb{R}^n = \mathcal{D}_q \oplus \mathcal{V}_q$, for all $q \in \mathbb{R}^n$) we can complete an orthonormal frame to a global one $X_1, \dots, X_k, Y_1, \dots, Y_m$, where $m = n - k$ and $\mathcal{V}_q = \text{span}_q\{Y_1, \dots, Y_m\}$. Since the structure is nilpotent, we can assume that the only nontrivial commutation relations are

$$[X_i, X_j] = \sum_{h=1}^m b_{ij}^h Y_h, \quad (22)$$

where B_1, \dots, B_m defined by $B_h = (b_{ij}^h)$ are skew-symmetric matrices (see [14] for the role of these matrices in the exponential map).

Due to the group structure, the intrinsic sub-Laplacian takes the form of sum of squares $\Delta = \sum_{i=1}^k X_i^2$ (see Remark 14). The group structure also implies that the heat kernel is invariant with respect to the group operation hence it is enough to consider the heat kernel $p_t(0, q)$ starting from the identity of the group, which we also denote $p_t(q)$. The heat kernel is written as follows (see again [17, 21])

$$p_t(q) = \frac{2}{(4\pi t)^{Q/2}} \int_{\mathbb{R}^m} V(B(\tau)) \exp\left(-\frac{W(B(\tau))x \cdot x}{4t}\right) \cos\left(\frac{z \cdot \tau}{t}\right) d\tau,$$

where $q = (x, z)$, $x \in \mathbb{R}^k, z \in \mathbb{R}^m$, and $B(\tau) := \sum_{i=1}^m \tau_i B_i$. Moreover $V : \mathbb{R}^{n \times n} \rightarrow \mathbb{C}$ and $W : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n}$ are the matrix functions defined by

$$V(A) = \sqrt{\det\left(\frac{A}{\sin A}\right)}, \quad W(A) = \frac{A}{\tan A}.$$

Here Q is the Hausdorff dimension of the sub-Riemannian structure.

Notice that (23) differs by some constant factors from the formulas contained in [17] since there the heat kernel is the solution of the equation $\partial_t = \frac{1}{2}\Delta$.

Remark 27. Assume that the real skew-symmetric matrix $B(\tau)$ is diagonalizable and denote by $\pm i\lambda_j(\tau)$, for $j = 1, \dots, \ell$, its non zero eigenvalues. Then we have the formula for the expansion on the “vertical subspace” (i.e. where $x = 0$)

$$p_t((0, z)) = \frac{2}{(4\pi t)^{Q/2}} \int_{\mathbb{R}^m} \prod_{j=1}^{\ell} \frac{\lambda_j(\tau)}{\sinh \lambda_j(\tau)} \cos\left(\frac{z \cdot \tau}{t}\right) d\tau. \quad (23)$$

6.2 The Heisenberg group

The Heisenberg group is the simplest example of sub-Riemannian manifold. It is defined by the orthonormal frame $\mathcal{D} = \text{span}\{X_1, X_2\}$ on \mathbb{R}^3 (with coordinates (x, y, z)) defined by

$$X_1 = \partial_x - \frac{y}{2}\partial_z, \quad X_2 = \partial_y + \frac{x}{2}\partial_z.$$

Defining $Z = \partial_z$, we have the commutation relations $[X_1, X_2] = Z$ and $[X_1, Z] = [X_2, Z] = 0$. Denote by $\mathcal{E}_0 : \Lambda_0 \times \mathbb{R}^+ \rightarrow M$ the exponential map starting from the origin, where

$$\Lambda_0 = \{p_0 = (\theta, w) \in T_0^*M \mid \theta \in S^1, w \in \mathbb{R}\}.$$

For every $p_0 = (\theta, w) \in \Lambda_0$ with $|w| \neq 0$, the arclength geodesic $\gamma(t) = \mathcal{E}_0(p_0, t) = (x(t), y(t), z(t))$ associated with the initial covector p_0 is described by the equations

$$\begin{aligned} x(t) &= \frac{1}{w}(\cos(wt + \theta) - \cos \theta), \\ y(t) &= \frac{1}{w}(\sin(wt + \theta) - \sin \theta), \\ z(t) &= \frac{1}{2w^2}(wt - \sin wt). \end{aligned} \tag{24}$$

and is optimal up to its cut time $t_{cut} = 2\pi/w$, with $\gamma(t_{cut}) = (0, 0, \pi/w^2)$. If $w = 0$, the geodesic is a straight line contained in the xy -plane and $t_{cut} = +\infty$.

From these properties it follows that the cut locus starting from the origin coincides with the z -axis, and for every point $\zeta = (0, 0, z)$ in this set we have $d^2(0, \zeta) = 4\pi|z|$.

Remark 28. The expression of the heat kernel $p_t(q)$ for the Heisenberg group is well known and was first computed by Gaveau [31] and Hulanicki [34]. The integral formula for p_t can be directly recovered from (23) since in this case there is a single skew-symmetric matrix B

$$B = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad \text{eig}(B(\tau)) = \{\pm i\tau\}.$$

Hence it follows

$$p_t(0, q) = \frac{2}{(4\pi t)^2} \int_{-\infty}^{\infty} \frac{\tau}{\sinh \tau} \exp\left(-\frac{x^2 + y^2}{4t} \frac{\tau}{\tanh \tau}\right) \cos\left(\frac{z\tau}{t}\right) d\tau.$$

On the vertical axis the integral can be explicitly computed

$$p_t(0, \zeta) = \frac{2}{(4\pi t)^2} \int_{-\infty}^{\infty} \frac{\tau}{\sinh \tau} \cos\left(\frac{z\tau}{t}\right) d\tau = \frac{1}{8t^2} \frac{1}{1 + \cosh\left(\frac{\pi z}{t}\right)}.$$

Hence, using that $d^2(0, \zeta) = 4\pi z$ we have

$$p_t(0, \zeta) = \frac{1}{t^2} \exp\left(-\frac{\pi z}{t}\right) \psi(t) = \frac{1}{t^2} \exp\left(-\frac{d^2(0, \zeta)}{4t}\right) \psi(t), \tag{25}$$

where $\psi(t)$ is a smooth function of t , nonvanishing at 0. (Here z is fixed.)

In what follows, we recover the expansion (25) computing the expansion of the hinged energy function and applying Corollary 1. For reasons of symmetry it is not restrictive to consider only points $\hat{\zeta} = (0, 0, \hat{z})$ such that $\hat{z} > 0$ (the on-diagonal expansion is a different situation).

The set of minimal geodesics joining 0 to $\hat{\zeta} = (0, 0, \hat{z})$ is parametrized by the covectors $p_0 = (\theta, \hat{w})$ where $\theta \in S^1$, $\hat{z} = \pi/\hat{w}^2$. For each p_0 , the associated geodesic γ_{p_0} satisfies $\gamma_{p_0}(0) = 0$ and $\gamma_{p_0}(2\pi/\hat{w}) = \hat{\zeta}$. Further, we have that the set of midpoints Γ is characterized as follows

$$\Gamma = \mathcal{E}_0 \left(S^1, \hat{w}, \frac{\pi}{\hat{w}} \right) = \left\{ \left(x, y, \frac{\hat{z}}{2} \right) : x^2 + y^2 = 2/\hat{w} \right\}.$$

We introduce cylindrical coordinates (ρ, φ, z) , where $x = \rho \cos \varphi$, $y = \rho \sin \varphi$. We have that (t, θ, w) forms a smooth coordinate system on $N(\Gamma)$ (for $N(\Gamma)$ small), where t represents the distance from the origin. Because of the invariance with respect to rotation around the z axis, to compute the Hessian of the hinged energy function $h_{0, \hat{\zeta}}$ we are left to study the relationship between (ρ, z) and (t, w) near Γ . We have

$$\rho(t, w) = \frac{2}{w} \sin \left(\frac{wt}{2} \right), \quad \text{and} \quad z(t, w) = \frac{1}{2w^2} (wt - \sin wt).$$

Recall that

$$h_{0, \hat{\zeta}}(\rho, z) = \frac{1}{2} \left(d^2(0, (\rho, z)) + d^2((\rho, z), \hat{\zeta}) \right).$$

Using that t represents the distance from the origin and exchanging the role of 0 and $\hat{\zeta}$, one can get with some implicit differentiation for the matrix element of the Hessian of $h_{0, \hat{\zeta}}$

$$\left. \frac{\partial^2}{\partial z^2} h_{0, \hat{\zeta}}(\rho, z) \right|_{\Gamma} = 2\hat{w}^2, \quad \left. \frac{\partial^2}{\partial \rho^2} h_{0, \hat{\zeta}}(\rho, z) \right|_{\Gamma} = \frac{\pi^2}{2}, \quad \left. \frac{\partial^2}{\partial \rho \partial z} h_{0, \hat{\zeta}}(\rho, z) \right|_{\Gamma} = 0.$$

It follows that there exists a smooth change of coordinates $(\rho, z) \mapsto (u, v)$ on a small disk perpendicular to Γ (with respect to the usual \mathbb{R}^3 metric) with the following three properties. First, Γ corresponds to the set where u and v are both zero. Second, $h_{0, \hat{\zeta}}(u, v) = \frac{\pi^2}{\hat{w}^2} + u^2 + v^2$ on $N(\Gamma)$. Third, $du = \frac{\pi}{2} d\rho$ on Γ and $dv = \hat{w} dz$ on Γ . Applying Theorem 25 and keeping track of all the constants one gets

$$p_t(0, \hat{\zeta}) = \frac{1}{t^2} \exp \left(-\frac{\pi^2/\hat{w}^2}{t} \right) \left(\frac{48\pi (c_0(0, \Gamma))^2}{\hat{w}^2} + O(t) \right).$$

where $c_0(0, \Gamma)$ is the constant appearing in the Ben Arous expansion. Taking into account that $\hat{z} = \pi/\hat{w}^2$, the heat kernel decays like a constant times $t^{-2} \exp(-4\pi\hat{z}/4t)$, which agrees with what one obtains from equation (25).

6.3 (3,6) case

The free nilpotent Lie group (3,6) is the sub-Riemannian structure on \mathbb{R}^6 (with coordinates $(x_1, x_2, x_3, z_1, z_2, z_3)$) defined by the distribution $\mathcal{D} = \text{span}\{X_1, X_2, X_3\}$, where the vector fields

$$X_1 = \partial_{x_1} - \frac{1}{2}x_2\partial_{z_3} + \frac{1}{2}x_3\partial_{z_2},$$

$$\begin{aligned} X_2 &= \partial_{x_2} + \frac{1}{2}x_1\partial_{z_3} - \frac{1}{2}x_3\partial_{z_1}, \\ X_3 &= \partial_{x_3} + \frac{1}{2}x_2\partial_{z_1} - \frac{1}{2}x_1\partial_{z_2}, \end{aligned}$$

define an orthonormal frame. If we set $Z_i = \partial_{z_i}$ for $i = 1, 2, 3$ we have $[X_1, X_2] = Z_3$, $[X_2, X_3] = Z_1$, and $[X_3, X_1] = Z_2$.

In this case the matrices $B_k = (b_{ij}^k)$ defined by the identities $[X_i, X_j] = b_{ij}^k Y_k$ are

$$B_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \quad B_3 = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

and for their linear combination $B(\tau) = \sum_{j=1}^3 \tau_j B_j$ we have $\text{eig}(B(\tau)) = \{0, \pm i|\tau|\}$, where we denote by $|\cdot|$ the standard norm on \mathbb{R}^3 .

Using (23) the explicit expression on the ‘‘vertical’’ subspace, i.e. at a point $\zeta = (0, 0, 0, z_1, z_2, z_3)$ is written as follows

$$p_t(\zeta) = \frac{2}{(4\pi t)^{9/2}} \int_{\mathbb{R}^3} \frac{|\tau|}{\sinh|\tau|} \cos\left(\frac{\tau \cdot z}{t}\right) d\tau, \quad (26)$$

To compute the expansion of the heat kernel for $t \rightarrow 0$ we use the fact that (26) is the Fourier transform of the radial function $f(\tau) = |\tau|/\sinh|\tau|$.

Recall that, if $F(x) = f(|x|)$ is a radial function defined on \mathbb{R}^m , its Fourier transform $\widehat{F}(\xi)$ is itself a radial function, i.e. it is defined by $\widehat{F}(\xi) = g(|\xi|)$, where g is the function of one variable that satisfies

$$g(\rho) = \frac{(2\pi)^{m/2}}{\rho^{\frac{m-2}{2}}} \int_0^\infty J_{\frac{m-2}{2}}(\tau\rho) \tau^{m/2} f(\tau) d\tau, \quad \rho = |\xi|,$$

and J denotes the Bessel function. In our case $m = 3$, we have $J_{1/2}(s) = \sqrt{\frac{2}{\pi s}} \sin s$ and

$$g(\rho) = 4\pi \int_0^\infty \frac{\sin \rho\tau}{\rho\tau} f(\tau) \tau^2 d\tau,$$

Then we can rewrite our heat kernel as the 1-dimensional integral

$$p_t(\zeta) = \frac{8\pi}{(4\pi t)^{9/2}} \int_0^\infty \frac{\tau^2 \sin \rho\tau}{\rho \sinh \tau} d\tau, \quad \text{where} \quad \rho = \frac{|z|}{t}.$$

Using that

$$\int_0^\infty \frac{\tau^2 \sin \rho\tau}{\rho \sinh \tau} d\tau = \frac{2\pi^3 \sinh^4\left(\frac{\pi\rho}{2}\right)}{\rho \sinh^3(\pi\rho)}, \quad \rho \in \mathbb{R},$$

we can explicitly write the expression of the heat kernel for ζ such that $|z| = 1$

$$p_t(\zeta) = \frac{\sinh^4\left(\frac{\pi}{2t}\right)}{32\sqrt{\pi}t^{7/2}\sinh^3\left(\frac{\pi}{t}\right)}. \quad (27)$$

From (27) one can immediately show that for such ζ

$$\lim_{t \rightarrow 0} t^{7/2} e^{\frac{\pi}{t}} p_t(\zeta) = C > 0. \quad (28)$$

The following lemma is a direct consequence of Theorem 17:

Lemma 29. *Assume that there exist $\alpha, K > 0, t_0 > 0$ and constants $C_1, C_2 > 0$ such that*

$$\frac{C_1}{t^\alpha} e^{-\frac{K}{4t}} \leq p_t(x, y) \leq \frac{C_2}{t^\alpha} e^{-\frac{K}{4t}}, \quad \forall 0 \leq t \leq t_0. \quad (29)$$

Then $K = d^2(x, y)$.

Proof. Since \log is a monotone function, we can apply $4t \log$ both inequalities in (29) and letting $t \rightarrow 0^+$ (which is allowed since the estimate is uniform for small t) we have $\lim_{t \rightarrow 0^+} 4t \log p_t(x, y) = -K$ and the statement follows from Theorem 17. \square

Proposition 30. *Let $\zeta = (0, 0, 0, z_1, z_2, z_3)$ with $|z| = 1$. Then $d^2(0, \zeta) = 4\pi$ and the following asymptotic expansion holds*

$$p_t(\zeta) = \frac{1}{t^{7/2}} \exp\left(-\frac{d^2(0, \zeta)}{4t}\right) \varphi(t),$$

where $\varphi(t)$ is a smooth function nonvanishing at $t = 0$. Moreover ζ is a conjugate point.

Proof. This follows directly from (28), Lemma 29 and Corollary 2. \square

Remark 31. From this analysis of the heat kernel and the homogeneity of the distance one gets the following information: (i). $d^2(0, \zeta) = 4\pi|z|$ for every $\zeta = (0, 0, 0, z_1, z_2, z_3)$. (ii). The point ζ is reached from the origin by an optimal geodesic that at time $t = \sqrt{4\pi|z|}$ is also conjugate.

These facts were proved in [45] with a detailed analysis of the exponential map. (Notice that by symmetry is not difficult to prove that the point is conjugate to the origin along the geodesic. On the contrary, the difficulty is in proving that the geodesic does not lose optimality before the conjugate locus.) Our method via the analysis of the heat kernel provides a shorter proof.

7 Grushin plane

The Grushin plane is the generalized sub-Riemannian structure on \mathbb{R}^2 for which an orthonormal frame of vector fields is given by

$$X = \partial_x, \quad Y = x\partial_y. \quad (30)$$

Since Y vanishes on the y -axis, this is a rank-varying sub-Riemannian structure and in particular is a 2-dimensional almost-Riemannian structure (see Appendix). One immediately verifies that the Lie bracket generating condition is satisfied since $[X, Y] = \partial_y$.

In this section we compute the expansion of the heat kernel in the Grushin plane at a conjugate point, starting from a Riemannian point.

The interesting feature of this structure is that it provides an example of almost Riemannian geometry in which the geodesic flow is completely integrable by means of trigonometric functions and, at the same time, the conjugate locus has the same structure of the conjugate locus of a generic 2-dimensional Riemannian metric.

The sub-Riemannian Hamiltonian associated with the orthonormal frame (30) (in standard coordinates $\lambda = (p_x, p_y, x, y)$ in $T^*\mathbb{R}^2$) is the smooth function

$$H : T^*\mathbb{R}^2 \rightarrow \mathbb{R}, \quad H(p_x, p_y, x, y) = \frac{1}{2}(p_x^2 + x^2 p_y^2). \quad (31)$$

Since in this case there are no abnormal minimizers (see [7]) the arclength geodesic flow starting from the Riemannian point $q_0 = (-1, -\pi/4)$ is computed as the solution of the Hamiltonian system associated with H , with initial condition $(x_0, y_0) = (-1, -\pi/4)$ and $(p_x(0), p_y(0)) = (\cos \theta, \sin \theta)$, where $\theta \in S^1$. The exponential map $\mathcal{E} : \mathbb{R}_+ \times S^1 \rightarrow \mathbb{R}^2$ starting from q_0 , is computed as follows (we omit the base point q_0 in the notation)

$$\begin{aligned} \mathcal{E}(t, \theta) &= (x(t, \theta), y(t, \theta)), \\ x(t, \theta) &= -\frac{\sin(\theta - t \sin \theta)}{\sin \theta}, \\ y(t, \theta) &= -\frac{\pi}{4} + \frac{1}{4 \sin \theta} \left(2t - 2 \cos \theta + \frac{\sin(2\theta - 2t \sin \theta)}{\sin \theta} \right), \end{aligned} \tag{32}$$

with the understanding $\mathcal{E}(t, 0) = \lim_{\theta \rightarrow 0} \mathcal{E}(t, \theta) = (t - 1, -\pi/4)$.

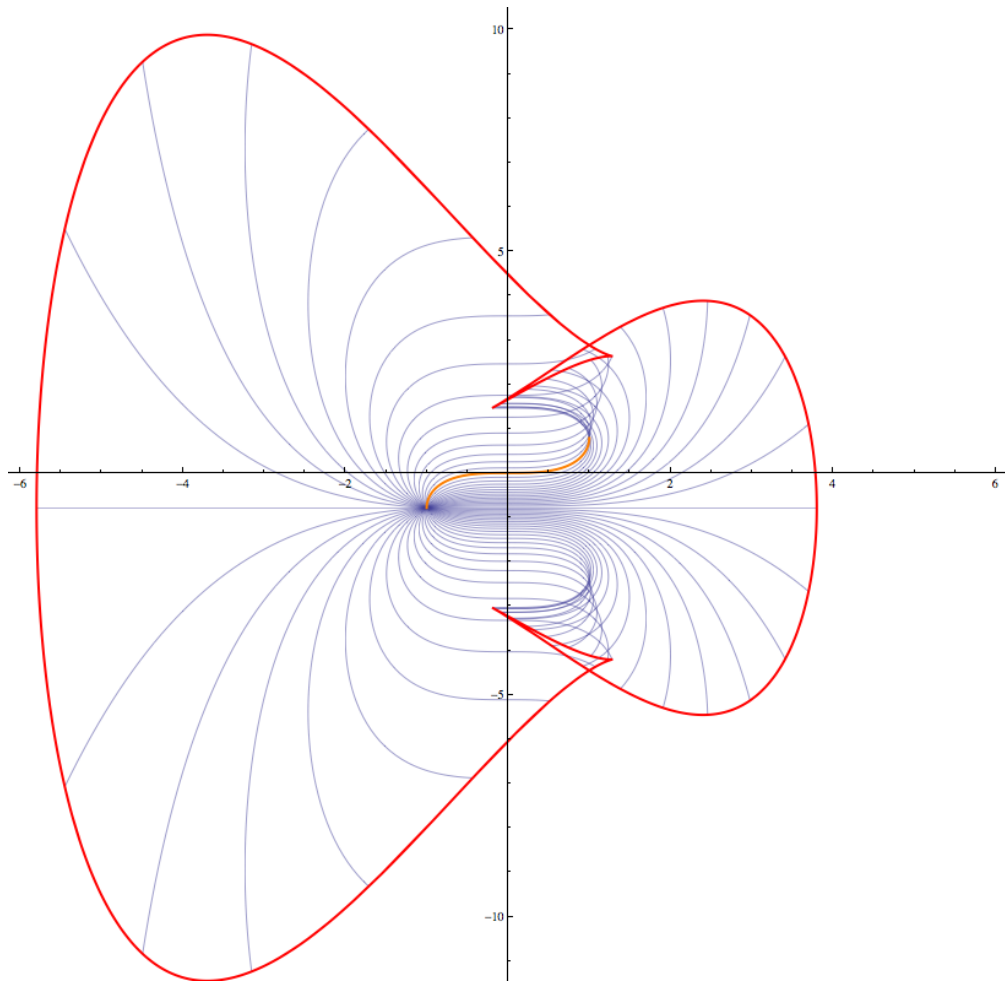


Figure 1: Geodesics starting from q_0

Let us consider the point $q_1 = (1, \pi/4)$, the symmetric of q_0 with respect to the origin. The point q_1 is both a cut and a conjugate point from q_0 . Indeed from the results of [7] immediately

follows that the cut locus from q_0 is the set $\text{Cut}(q_0) = \{(1, \pi/4 + s), s \geq 0\}$. Moreover

$$\mathcal{E}(\pi, \pi/2) = q_1, \quad \left. \frac{d}{d\theta} \right|_{\theta=\pi/2} \mathcal{E}(\pi, \theta) = (0, 0), \quad (33)$$

shows that q_1 is also conjugate to q_0 . Figure 1 shows some geodesics starting from the point q_0 and the endpoints of all geodesics starting from q_0 at time $T = 4.8$.

Remark 32. Notice that the geodesic with initial covector $\theta = \pi/2$ is the only one that reach q_1 optimally in time $T = \pi/2$. The midpoint of the geodesic is the origin $\mathcal{E}(\pi/4, \pi/2) = (0, 0)$. (See also Figure 2.)

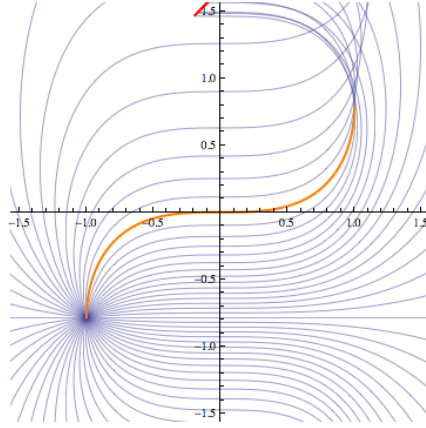


Figure 2: The conjugate geodesic

We are interested in the small time asymptotic expansion of $p_t(q_0, q_1)$, where p_t denotes the heat kernel of the sub-Riemannian heat equation

$$\partial_t \varphi = \Delta \varphi, \quad \Delta = X^2 + Y^2 = \partial_x^2 + x^2 \partial_y^2.$$

Here the sub-Laplacian is not the intrinsic one but is computed with respect to the standard Lebesgue measure of \mathbb{R}^2 . Indeed in this case the intrinsic volume $\mu = \frac{1}{|x|} dx dy$ is diverging along the singular set $\mathcal{Z} = \{x = 0\}$ hence our results does not apply since μ is not smooth. (See [22] for a discussion of the intrinsic heat equation in the Grushin plane.)

An integral representation for the heat kernel for the operator $\partial_x^2 + x^2 \partial_y^2$ can be easily obtained by computing the Fourier transform with respect to the y variable and then using the Mehler kernel for the quantum harmonic oscillator. Its expression, given $q = (x, y)$, $q' = (x', y')$ is

$$p_t(q, q') = \frac{1}{(2\pi t)^{3/2}} \int_{-\infty}^{\infty} \exp \left(\frac{xx'}{t} \frac{\tau}{\sinh(2\tau)} - \frac{(x^2 + x'^2)}{2t} \frac{\tau}{\tanh(2\tau)} \right) \times \\ \times \sqrt{\frac{\tau}{\sinh(2\tau)}} \exp \left(\frac{i(y - y') \tau}{t} \right) d\tau.$$

However from this formula it seems hard to find an asymptotic expansion for t small except on the diagonal at the origin.

Thanks to Corollary 1, to compute the asymptotic expansion of the heat kernel $p_t(q_0, q_1)$ we are reduced to study the expansion of the hinged energy function h_{q_0, q_1} near the origin (we omit the points in the notation in what follows)

$$\begin{aligned} h(x, y) &= \frac{1}{2} (d^2(q_0, (x, y)) + d^2(q_1, (x, y))) \\ &= \frac{1}{2} (d^2(q_0, (x, y)) + d^2(q_0, (-x, -y))), \end{aligned}$$

where the last identity follows from the symmetries of the structure and implies that the expansion of h at the origin contains only even order terms in (x, y) and we are reduced to compute the even terms of the expansion of the function $(x, y) \mapsto d^2(q_0, (x, y))$.

Remark 33. By (33) and the proof of Theorem 22 it follows that the Hessian of the hinged energy function h is degenerate along the direction $(-1, 1)$ since

$$\left. \frac{d}{d\theta} \right|_{\theta=\pi/2} \mathcal{E}(\pi/2, \theta) = (-1, 1).$$

For this reason we consider the new coordinate system (\bar{x}, \bar{y}) around 0 defined by $\bar{x} = \frac{x+y}{2}$, $\bar{y} = \frac{x-y}{2}$. The Hessian of h is diagonal in these coordinates.

Using the fact that the geodesics defined by (32) are parametrized by arclength, we can compute the derivatives of the distance with respect to (\bar{x}, \bar{y}) by computing derivatives of t from (32) with implicit differentiation (as in Section 6.2). After some computations one finds the following expansion for h (we omit the bar in \bar{x}, \bar{y} for the new system of coordinates)

$$\begin{aligned} h(x, y) &= 4x^2 + \frac{\alpha - 32}{24}x^4 - \frac{\alpha}{6}x^3y + \frac{\alpha - 32}{4}x^2y^2 \\ &\quad - \frac{\alpha}{6}xy^3 + \frac{\alpha}{24}y^4 + O(\|(x, y)\|^5), \end{aligned} \tag{34}$$

where $\alpha = \frac{3}{2}\pi^2$.

Concerning our hinged energy function (34) one can also show that the following explicit change of coordinates

$$\varphi(u, v) = \left(u + \frac{32 - \alpha}{192}u^3 + \frac{\alpha}{48}v^3 + \frac{\alpha}{48}u^2v + \frac{16 - \alpha}{32}uv^2, v\right).$$

diagonalizes h up to order 5. Namely

$$h(\varphi(u, v)) = 8u^2 + \frac{\alpha}{24}v^4 + O(\|(u, v)\|^5).$$

A direct application of Corollary 1 (recall also Lemma 21), together with $d(q_0, q_1) = \pi/2$, gives

Theorem 34. *The heat kernel $p_t(q_0, q_1)$ satisfies the following asymptotic expansion*

$$p_t(q_0, q_1) = \frac{1}{t^{5/4}} e^{-\frac{\pi^2}{16t}} (C + O(t)). \tag{35}$$

Remark 35. Notice that the same expansion as in (35) holds for the symmetric point $q_2 = (1, -3\pi/4)$. If $q \notin \{q_1, q_2\}$

$$p_t(q_0, q) \sim \frac{1}{t} e^{-\frac{d^2(q_0, q)}{4t}} (C + O(t)).$$

Remark 36. Corollary 2 can be applied to compute the heat kernel asymptotics starting from the origin. In this case the cut locus is the y axes and these points are not conjugate. On the diagonal, applying the Leandre - Ben Arous result (1) with $Q = 3$ (or using the explicit formula for the heat kernel given above), one gets $p_t((0,0), (0,0)) \sim C/t^{3/2}$ with $C > 0$. Off diagonal, applying Corollary 2, one gets $p_t((0,0), (x,y)) \sim C(x,y)/t$ for some $C(x,y) > 0$.

The expansion of the heat kernel for the Grushin plane is summarized in the following table:

	diagonal (Leandre) (Ben Arous)	off diagonal off cut locus (Ben Arous)	off diagonal cut (non-conjugate) (Corollary 1)	off diagonal cut conjugate (Corollary 2)
$p_t(q, q')$ q Riemannian point	$\sim \frac{C}{t}$	$\sim \frac{C}{t} e^{-d^2(q,q')/(4t)}$	$\sim \frac{C}{t} e^{-d^2(q,q')/(4t)}$	$\sim \frac{C}{t^{5/4}} e^{-d^2(q,q')/(4t)}$
$p_t(q, q')$ q degenerate point	$\sim \frac{C}{t^{3/2}}$	$\sim \frac{C}{t} e^{-d^2(q,q')/(4t)}$	$\sim \frac{C}{t} e^{-d^2(q,q')/(4t)}$	—

A Extension to rank-varying sub-Riemannian structures

In this section we give a more general definition of sub-Riemannian manifold (that we call rank-varying sub-Riemannian manifold). This definition includes also as a particular case Riemannian manifolds. For a more complete presentation one can see [3]. All the results of the paper hold for this more general structure.

Let M be an n -dimensional smooth manifold. Given a vector bundle \mathbf{U} over M , the $\mathcal{C}^\infty(M)$ -module of smooth sections of \mathbf{U} is denoted by $\Gamma(\mathbf{U})$. For the particular case $\mathbf{U} = TM$, the set of smooth vector fields on M is denoted by $\text{Vec}(M)$.

Definition 37. An (n, k) -rank-varying distribution on an n -dimensional manifold M is a pair (\mathbf{U}, f) where \mathbf{U} is a vector bundle of rank k over M and $f : \mathbf{U} \rightarrow TM$ is a morphism of vector bundles, i.e. (i) the diagram

$$\begin{array}{ccc} \mathbf{U} & \xrightarrow{f} & TM \\ & \searrow \pi_{\mathbf{U}} & \downarrow \pi \\ & & M \end{array}$$

commutes, where $\pi : TM \rightarrow M$ and $\pi_{\mathbf{U}} : \mathbf{U} \rightarrow M$ denote the canonical projections and (ii) f is linear on fibers. Moreover, we require the map $\sigma \mapsto f \circ \sigma$ from $\Gamma(\mathbf{U})$ to $\text{Vec}(M)$ to be injective.

Let (\mathbf{U}, f) be an (n, k) -rank-varying distribution, $\Delta = \{f \circ \sigma \mid \sigma \in \Gamma(\mathbf{U})\}$ be its associated submodule and denote by Δ_q the linear subspace $\{V(q) \mid V \in \Delta\} = f(\mathbf{U}_q) \subseteq T_q M$. Let $\text{Lie}(\Delta)$ be the smallest Lie subalgebra of $\text{Vec}(M)$ containing Δ and, for every $q \in M$, let $\text{Lie}_q(\Delta)$ be the linear subspace of $T_q M$ whose elements are the evaluation at q of elements belonging to $\text{Lie}(\Delta)$. We say that (\mathbf{U}, f) satisfies the *Hörmander condition* if $\text{Lie}_q(\Delta) = T_q M$ for every $q \in M$.

Definition 38. An (n, k) -rank-varying sub-Riemannian structure is a triple $(\mathbf{U}, f, \langle \cdot, \cdot \rangle)$ where (\mathbf{U}, f) is a Lie bracket generating (n, k) -rank-varying distribution on a manifold M and $\langle \cdot, \cdot \rangle_q$ is a scalar product on \mathbf{U}_q smoothly depending on q .

Several classical structures can be seen as particular cases of rank-varying sub-Riemannian structures, e.g., Riemannian structures (when $\mathbf{U} = TM$ and $f = id$) and constant-rank sub-Riemannian structures (as defined in Section 2). An (n, n) -rank-varying sub-Riemannian structure is called an *n-dimensional almost-Riemannian structure*. An example of 2-almost Riemannian structure is provided by the Grushin plane, see [7, ?].

If $\sigma_1, \dots, \sigma_k$ is an orthonormal frame for $\langle \cdot, \cdot \rangle$ on an open subset Ω of M , an *orthonormal frame* in Ω for the rank-varying sub-Riemannian structure is given by X_1, \dots, X_k , where $X_i := f \circ \sigma_i$. Orthonormal frames are systems of local generators of Δ . For every $q \in M$ and every $v \in \Delta_q$ define

$$\mathbf{G}_q(v) = \inf\{\langle u, u \rangle_q \mid u \in \mathbf{U}_q, f(u) = v\}.$$

Notice that if X_1, \dots, X_k is an orthonormal frame for the rank-varying sub-Riemannian structure in Ω , then it may happen that there exist a $q \in \Omega$ such that $\dim \text{span}\{X_1(q), \dots, X_k(q)\} < k$ and that $\mathbf{G}_q(X_i(q)) < 1$ for some i .

A Lipschitz continuous curve $\gamma : [0, T] \rightarrow M$ is said to be *horizontal* (or *admissible*) if there exists a measurable essentially bounded function

$$[0, T] \ni t \mapsto u(t) \in \mathbf{U}_{\gamma(t)},$$

called *control function*, such that $\dot{\gamma}(t) = f(u(t))$ for almost every $t \in [0, T]$. Given an admissible curve $\gamma : [0, T] \rightarrow M$, the *length* of γ is

$$\ell(\gamma) = \int_0^T \sqrt{\mathbf{G}_{\gamma(t)}(\dot{\gamma}(t))} dt.$$

The *Carnot–Caratheodory distance* is defined as

$$d(q_0, q_1) = \inf\{\ell(\gamma) \mid \gamma(0) = q_0, \gamma(T) = q_1, \gamma \text{ admissible}\}.$$

As in the classical sub-Riemannian case, the hypothesis of connectedness of M and the Hörmander condition guarantees the finiteness and the continuity of $d(\cdot, \cdot)$ with respect to the topology of M .

For rank-varying sub-Riemannian structures the definitions of minimizers, geodesics, normal and abnormal extremals and the formulation of the Pontryagin Maximum Principle are the same as in the constant rank case. Also the definition of cut and conjugate loci are the same. Thanks to the injectivity assumption, the definition of the horizontal gradient is still $\nabla\varphi = \sum_{i=1}^k X_i(\varphi)X_i$. The definition of the Popp’s volume is instead more delicate, since the volume diverges while approaching a point in which there is a drop of rank of the distribution. However, for a smooth volume μ the sub-Laplacian still has the form $\Delta = \sum_{i=1}^k X_i^2 + (\text{div}X_i)X_i$, and all the results of the paper hold in this case.

Acknowledgements. The authors would like to thank Fabrice Baudoin and Andrei Agrachev for helpful discussions. The authors also thank IHP for its hospitality during the finishing of this paper.

References

- [1] A. AGRACHEV, *Compactness for sub-Riemannian length-minimizers and subanalyticity*, Rend. Sem. Mat. Univ. Politec. Torino, 56 (1998), pp. 1–12 (2001). Control theory and its applications (Grado, 1998).

- [2] A. AGRACHEV AND D. BARILARI, *Sub-Riemannian structures on 3D Lie groups*, J. Dyn. and Contr. Syst., 18 (2012), pp. 21–44.
- [3] A. AGRACHEV, D. BARILARI, AND U. BOSCAIN, *Introduction to Riemannian and sub-Riemannian geometry (Lecture Notes)*, http://people.sissa.it/agrachev/agrachev_files/notes.html, (2012).
- [4] ———, *On the Hausdorff volume in sub-Riemannian geometry*, Calc. Var. and PDE's, 43 (2012), pp. 355–388.
- [5] A. AGRACHEV, B. BONNARD, M. CHYBA, AND I. KUPKA, *Sub-Riemannian sphere in Martinet flat case*, ESAIM Control Optim. Calc. Var., 2 (1997), pp. 377–448 (electronic).
- [6] A. AGRACHEV, U. BOSCAIN, J.-P. GAUTHIER, AND F. ROSSI, *The intrinsic hypoelliptic Laplacian and its heat kernel on unimodular Lie groups*, J. Funct. Anal., 256 (2009), pp. 2621–2655.
- [7] A. AGRACHEV, U. BOSCAIN, AND M. SIGALOTTI, *A Gauss-Bonnet-like formula on two-dimensional almost-Riemannian manifolds*, Discrete Contin. Dyn. Syst., 20 (2008), pp. 801–822.
- [8] A. AGRACHEV AND J.-P. GAUTHIER, *On the subanalyticity of Carnot-Caratheodory distances*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 18 (2001), pp. 359–382.
- [9] A. A. AGRACHEV, *Exponential mappings for contact sub-Riemannian structures*, J. Dynam. Control Systems, 2 (1996), pp. 321–358.
- [10] A. A. AGRACHEV, *Any sub-Riemannian metric has points of smoothness*, Dokl. Akad. Nauk, 424 (2009), pp. 295–298.
- [11] A. A. AGRACHEV AND Y. L. SACHKOV, *Control theory from the geometric viewpoint*, vol. 87 of Encyclopaedia of Mathematical Sciences, Springer-Verlag, Berlin, 2004. Control Theory and Optimization, II.
- [12] V. I. ARNOL'D, S. M. GUSEĬN ZADE, AND A. N. VARCHENKO, *Singularities of differentiable maps. Vol. II*, vol. 83 of Monographs in Mathematics, Birkhäuser Boston Inc., Boston, MA, 1988.
- [13] D. BARILARI, *Trace heat kernel asymptotics in 3d contact sub-Riemannian geometry*, To appear on Journal of Mathematical Sciences, (2011).
- [14] D. BARILARI, U. BOSCAIN, AND J.-P. GAUTHIER, *On 2-step, corank 2 sub-Riemannian metrics*, SIAM Journal of Control and Optimization, 50 (2012), pp. 559–582.
- [15] F. BAUDOIN, *An introduction to the geometry of stochastic flows*, Imperial College Press, London, 2004.
- [16] F. BAUDOIN AND N. GAROFALO, *Curvature-dimension inequalities and Ricci lower bounds for sub-Riemannian manifolds with transverse symmetries*, arXiv:1101.3590v1.
- [17] R. BEALS, B. GAVEAU, AND P. GREINER, *The Green function of model step two hypoelliptic operators and the analysis of certain tangential Cauchy Riemann complexes*, Adv. Math., 121 (1996), pp. 288–345.
- [18] G. BEN AROUS, *Développement asymptotique du noyau de la chaleur hypoelliptique hors du cut-locus*, Ann. Sci. École Norm. Sup. (4), 21 (1988), pp. 307–331.
- [19] G. BEN AROUS AND R. LÉANDRE, *Décroissance exponentielle du noyau de la chaleur sur la diagonale. II*, Probab. Theory Related Fields, 90 (1991), pp. 377–402.
- [20] J.-M. BISMUT, *Large deviations and the Malliavin calculus*, vol. 45 of Progress in Mathematics, Birkhäuser Boston Inc., Boston, MA, 1984.
- [21] A. BONFIGLIOLI, E. LANCONELLI, AND F. UGUZZONI, *Stratified Lie groups and potential theory for their sub-Laplacians*, Springer Monographs in Mathematics, Springer, Berlin, 2007.
- [22] U. BOSCAIN AND C. LAURENT, *The Laplace-Beltrami operator in almost-Riemannian geometry*, arXiv:1105.4687v1 [math.SP], Preprint, (2011).

- [23] U. BOSCAIN AND S. POLIDORO, *Gaussian estimates for hypoelliptic operators via optimal control*, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl., 18 (2007), pp. 333–342.
- [24] U. BOSCAIN AND F. ROSSI, *Invariant Carnot-Carathéodory metrics on S^3 , $SO(3)$, $SL(2)$, and lens spaces*, SIAM J. Control Optim., 47 (2008), pp. 1851–1878.
- [25] R. W. BROCKETT AND A. MANSOURI, *Short-time asymptotics of heat kernels for a class of hypoelliptic operators*, Amer. J. Math., 131 (2009), pp. 1795–1814.
- [26] D. BURAGO, Y. BURAGO, AND S. IVANOV, *A course in metric geometry*, vol. 33 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, 2001.
- [27] Y. CHITOUR, F. JEAN, AND E. TRÉLAT, *Genericity results for singular curves*, J. Differential Geom., 73 (2006), pp. 45–73.
- [28] E.-H. C. EL-ALAOU, J.-P. GAUTHIER, AND I. KUPKA, *Small sub-Riemannian balls on \mathbf{R}^3* , J. Dynam. Control Systems, 2 (1996), pp. 359–421.
- [29] R. ESTRADA AND R. P. KANWAL, *A distributional approach to asymptotics*, Birkhäuser Advanced Texts: Basler Lehrbücher. [Birkhäuser Advanced Texts: Basel Textbooks], Birkhäuser Boston Inc., Boston, MA, second ed., 2002. Theory and applications.
- [30] G. B. FOLLAND AND E. M. STEIN, *Estimates for the $\bar{\partial}_b$ complex and analysis on the Heisenberg group*, Comm. Pure Appl. Math., 27 (1974), pp. 429–522.
- [31] B. GAVEAU, *Principe de moindre action, propagation de la chaleur et estimées sous elliptiques sur certains groupes nilpotents*, Acta Math., 139 (1977), pp. 95–153.
- [32] D. GROMOLL AND W. MEYER, *On differentiable functions with isolated critical points*, Topology, 8 (1969), pp. 361–369.
- [33] E. P. HSU, *Stochastic analysis on manifolds*, vol. 38 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, 2002.
- [34] A. HULANICKI, *The distribution of energy in the Brownian motion in the Gaussian field and analytic-hypoellipticity of certain subelliptic operators on the Heisenberg group*, Studia Math., 56 (1976), pp. 165–173.
- [35] D. JERISON AND A. SÁNCHEZ-CALLE, *Subelliptic, second order differential operators*, in Complex analysis, III (College Park, Md., 1985–86), vol. 1277 of Lecture Notes in Math., Springer, Berlin, 1987, pp. 46–77.
- [36] D. S. JERISON AND A. SÁNCHEZ-CALLE, *Estimates for the heat kernel for a sum of squares of vector fields*, Indiana Univ. Math. J., 35 (1986), pp. 835–854.
- [37] S. KUSUOKA AND D. STROOCK, *Applications of the Malliavin calculus. II*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 32 (1985), pp. 1–76.
- [38] R. LÉANDRE, *Majoration en temps petit de la densité d’une diffusion dégénérée*, Probab. Theory Related Fields, 74 (1987), pp. 289–294.
- [39] ———, *Minoration en temps petit de la densité d’une diffusion dégénérée*, J. Funct. Anal., 74 (1987), pp. 399–414.
- [40] ———, *Développement asymptotique de la densité d’une diffusion dégénérée*, Forum Math., 4 (1992), pp. 45–75.
- [41] J. MITCHELL, *On Carnot-Carathéodory metrics*, J. Differential Geom., 21 (1985), pp. 35–45.
- [42] I. MOISEEV AND Y. L. SACHKOV, *Maxwell strata in sub-Riemannian problem on the group of motions of a plane*, ESAIM Control Optim. Calc. Var., 16 (2010), pp. 380–399.

- [43] S. A. MOLČANOV, *Diffusion processes, and Riemannian geometry*, Uspehi Mat. Nauk, 30 (1975), pp. 3–59.
- [44] R. MONTGOMERY, *A tour of subriemannian geometries, their geodesics and applications*, vol. 91 of Mathematical Surveys and Monographs, American Mathematical Society, Providence, RI, 2002.
- [45] O. MYASNICHENKO, *Nilpotent (3, 6) sub-Riemannian problem*, J. Dynam. Control Systems, 8 (2002), pp. 573–597.
- [46] R. NEEL, *The small-time asymptotics of the heat kernel at the cut locus*, Comm. Anal. Geom., 15 (2007), pp. 845–890.
- [47] R. NEEL AND D. STROOCK, *Analysis of the cut locus via the heat kernel*, in Surveys in differential geometry. Vol. IX, Surv. Differ. Geom., IX, Int. Press, Somerville, MA, 2004, pp. 337–349.
- [48] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISHCHENKO, *The mathematical theory of optimal processes*, Translated from the Russian by K. N. Trirogoff; edited by L. W. Neustadt, Interscience Publishers John Wiley & Sons, Inc. New York-London, 1962.
- [49] L. RIFFORD AND E. TRÉLAT, *Morse-Sard type results in sub-Riemannian geometry*, Math. Ann., 332 (2005), pp. 145–159.
- [50] L. P. ROTHSCHILD AND E. M. STEIN, *Hypoelliptic differential operators and nilpotent groups*, Acta Math., 137 (1976), pp. 247–320.
- [51] Y. L. SACHKOV, *Symmetries of flat rank two distributions and sub-Riemannian structures*, Trans. Amer. Math. Soc., 356 (2004), pp. 457–494 (electronic).
- [52] A. SÁNCHEZ-CALLE, *Fundamental solutions and geometry of the sum of squares of vector fields*, Invent. Math., 78 (1984), pp. 143–160.
- [53] R. S. STRICHARTZ, *Sub-Riemannian geometry*, J. Differential Geom., 24 (1986), pp. 221–263.
- [54] T. J. S. TAYLOR, *Off diagonal asymptotics of hypoelliptic diffusion equations and singular Riemannian geometry*, Pacific J. Math., 136 (1989), pp. 379–399.
- [55] S. R. S. VARADHAN, *On the behavior of the fundamental solution of the heat equation with variable coefficients*, Comm. Pure Appl. Math., 20 (1967), pp. 431–455.