COMMUNICATIONS IN
ANALYSIS AND GEOMETRY
Volume 15, Number 4, 845–890, 2007

The small-time asymptotics of the heat kernel at the cut locus

Robert Neel

We study the small-time asymptotics of the gradient and Hessian of the logarithm of the heat kernel at the cut locus, giving, in principle, complete expansions for both quantities. We relate the leading terms of the expansions to the structure of the cut locus, especially to conjugacy, and we provide a probabilistic interpretation in terms of the Brownian bridge. In particular, we show that the cut locus is the set of points where the Hessian blows up faster than $1/t$. We also study the distributional asymptotics and use them to compute the distributional Hessian of the energy function (that is, one-half the distance function squared).

1. Introduction

Let M be a compact, connected, smooth Riemannian manifold of dimension n. For any point $x \in M$, we use $\mathrm{Cut}(x) \subset M$ to denote the cut locus of x. In particular, since M is compact, $\mathrm{Cut}(x)$ will be non-empty for every x. The Riemannian metric induces a distance function $\mathrm{dist}(x, y)$. We will also need to consider the energy function, $E(x, y) = \frac{1}{2}\,\mathrm{dist}(x, y)^2$. Let Δ be the Laplace–Beltrami operator on M, that is, if $x_1, \ldots, x_n$ are normal coordinates centered at a point p, then

$$\Delta = \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2} \text{ at } p.$$

The heat kernel $p_t(x, y)$ is the fundamental solution to the heat equation $\partial_t u(t, x) = \frac{1}{2}\Delta u(t, x)$.

The author gratefully acknowledges support from an NSF Graduate Research Fellowship and a Clay Liftoff Fellowship. This paper is partially based on the author's Ph.D. thesis.

A well-known result of Varadhan states that $t \log p_t(x,y)$ converges to $-E(x,y)$ as $t \searrow 0$ uniformly on all of M. Motivated by this result, we define

$$E_t(x,y) = -t \log p_t(x,y)$$

and observe that $E_t(x,y) \to E(x,y)$ uniformly. Malliavin and Stroock [12] have shown, using path-space methods, that away from the cut locus, spatial derivatives of $E_t(x,y)$ commute with taking the limit as $t \searrow 0$ (for an analytic proof, see [3]). Clearly, the lack of differentiability of $E(x,y)$ at the cut locus means that something else must be occurring there (see Bishop [5] for a brief discussion of the smoothness properties of the distance function at the cut locus). Indeed, in the same paper, Malliavin and Stroock use path-space integration to show that, if the set of minimal geodesics connecting x and y is sufficiently "nice", then $\nabla^2 E_t(x,y)$ is asymptotic to $-1/t$ times the variance of a random variable on path space as $t \searrow 0$. Unfortunately, their analysis is too complicated to obtain more detailed information.

In the present paper, we develop an analogous, but purely finite-dimensional approach which allows a much more detailed analysis of the small time asymptotics of the gradient and Hessian of $E_t(x,y)$. In particular, we will show how complete asymptotic expansions of the gradient and the Hessian of $E_t(x,y)$ can be represented as an integrals over the set of midpoints of minimal geodesics from x to y. These general expansions are given in Theorems 2.3 and 2.4, respectively. The leading term in each of these expansions, given at the beginning of section 3.1, is fairly accessible both to analysis and to a probabilistic interpretation, and we will show how the small-time behavior of $\nabla^2 E_t(x,y)$ reflects the structure of minimal geodesics from x to y. In particular, we will show (see Theorem 3.5 below) that $\nabla^2 E_t(x,y)$, as a quadratic form on $T_y M$, is unbounded as $t \searrow 0$ if and only if $y \in \mathrm{Cut}(x)$. Finally, in Theorems 4.1 and 4.2, we show how the asymptotic expansion of $\nabla^2 E_t(x,y)$ can be used to determine the distributional Hessian of $E(x,y)$.

We note that our methods are an extension of a procedure originally due to Molchanov [15] of representing the heat kernel itself as an integral over the midpoints of minimal geodesics, allowing him to determine the rate of decay of the heat kernel at points $y \in \mathrm{Cut}(x)$ in a variety of cases.

A survey of the results in the present paper can be found in [16].

2. The representation as Laplace integrals

Our main tool will be a pair of formulas which express the gradient and the Hessian of $E_t(x,y)$ as integrals over the set of midpoints of minimal

geodesics from x to y, the asymptotics of which are amenable to study. The present section is devoted to the derivation of these formulas.

2.1. Preliminary results

We will need various facts about the small-time asymptotics of the heat kernel, which we present here.

Let $C_M \subset M \times M$ be the set of pairs of points (x, y) such that $y \in \mathrm{Cut}(x)$. Away from the cut locus, we have the following asymptotic expansion of the heat kernel, due to Minakshisundaram and Pleijel [14] (see [6] for a more modern development).

Theorem 2.1. *Let M be a smooth, complete Riemannian manifold of dimension n. Then there are smooth functions $H_i(x, y)$ defined on $(M \times M)\backslash C_M$ such that the asymptotic expansion*

$$p_t(x, y) \sim \left(\frac{1}{2\pi t}\right)^{n/2} e^{-E(x,y)/t} \sum_{i=0}^{\infty} H_i(x, y) t^i$$

holds uniformly as $t \searrow 0$ on compact subsets of $(M \times M)\backslash C_M$. Further, if $y = \exp_x(Y)$, then $H_0(x, y)$ is given by the reciprocal of the square root of the Jacobian of $\exp_x$ at Y.

For future use, let $k(t, x, y) = (2\pi t)^{n/2} e^{E(x,y)/t} p_t(x, y)$ be defined away from the cut locus such that $k(t, x, y) \sim \sum_{i=0}^{\infty} H_i(x, y) t^i$. Recall the result of Varadhan mentioned in introduction; namely that E_t converges to E uniformly on all of M. We will rewrite this as

$$p_t(x, y) = \exp\left[\frac{-E(x, y) + \delta(t, x, y)}{t}\right], \tag{2.1}$$

where $\delta(t, x, y)$ is some function which goes to 0 uniformly in t on all of M. Incidentally, one cannot hope to replace $\delta(t, x, y)$ with a power series expansion. Even in the simplest case of the heat kernel on $\mathbb{S}^1$, we see that $\delta(t, x, y)$ fails to be $\mathrm{O}(t)$ for any x and y.

Having summarized the small-time asymptotics of the heat kernel itself both away from and on the cut locus, we now turn to the log derivatives. As mentioned, Malliavin and Stroock [12] (for $m = 1, 2$) and Stroock and

Turetsky [18] (for $m > 2$) have proved that

$$\nabla^m E_t(x,y) \to \nabla^m E(x,y) \tag{2.2}$$

uniformly on compact subsets of $M \setminus \mathrm{Cut}(x)$, where ∇^m is the mth covariant derivative and all derivatives are taken in the y variable. Next, we will need better control over the convergence of $\nabla^m E_t(x,y)$ away from the cut locus.

Lemma 2.2. *Let M be a smooth, complete Riemannian manifold of dimension n. Then there are smooth functions $G_i(x,y)$ defined on $(M \times M)\setminus C_M$ such that, for any positive integer m, the asymptotic expansion*

$$t\nabla^m \log p_t(x,y) \sim \sum_{i=0}^{\infty} \nabla^m G_i(x,y) t^i$$

holds uniformly as $t \searrow 0$ on compact subsets of $(M \times M)\setminus C_M$. Further, the G_i are given in terms of the H_i by taking the log derivatives of the Minakshisundaram–Pleijel expansion of Theorem 2.1, *and in particular, $G_0(x,y) = -E(x,y)$.*

This lemma is a direct consequence of the fact that the Minakshi–sundaram–Pleijel expansion can be differentiated. For a probabilistic proof of this fact see [2]; for an analytic proof see [3]. Note that we get a power series expansion only if we take at least one spatial derivative. Again for future use, let $l(t,x,y,A) = t\nabla_A \log p_t(x,y)$ be defined away from the cut locus, so that it has the above expansion.

Finally, we have a result of Stroock and Turetsky [19] which states that

$$|\nabla^m p_t(x,y)| \le D_m \left[\frac{\mathrm{dist}(x,y)}{t} + \frac{1}{\sqrt{t}}\right]^m p_t(x,y), \tag{2.3}$$

where the D_m are some constants depending only on M. The point is that this estimate is valid even when $y \in \mathrm{Cut}(x)$. Note that the results of Stroock and Malliavin mentioned above for $m = 2$, and the results of Stroock and Turetsky for $m > 2$, show that the power of t in Equation (2.3) is sharp.

2.2. The gradient

In this section, we will prove a theorem describing the asymptotic expansion of the gradient of $E_t(x,y)$ which is valid everywhere on M. We begin by introducing some notation. Fix any two distinct points x and y on M. Let

Γ be the set of midpoints of minimal geodesics from x to y (for example, in the case $M = \mathbb{S}^n$ with x and y the north and south poles, Γ is the equator). By compactness, there exists an $\epsilon > 0$ such that the ϵ-neighborhood of Γ, denoted by Γ_ϵ, is strictly positive distance from x, y, and both of their cut loci. Also, we define the *hinged energy function* $h_{x,y}(z) = E(x,z) + E(y,z)$. Note that $h_{x,y}$ obtains its minimum precisely on Γ and that this minimum is equal to $E(x,y)/2$.

Theorem 2.3. *Let M be a smooth, compact, connected Riemannian manifold. Choose any two distinct points x and y. Then there exist positive constants C and λ such that, for any $A \in T_yM$, we have*

$$\nabla_A E_{2t}(x,y) = -\frac{\int_{\Gamma_\epsilon} l(t,z,y,A)\exp\left[-(h_{x,y}(z)/t)\right]k(t,x,z)k(t,y,z)dz}{\int_{\Gamma_\epsilon}\exp\left[-(h_{x,y}(z)/t)\right]k(t,x,z)k(t,y,z)dz} + \hat{e}(t,x,y),$$

where $l(t,x,y,A)$ and $k(t,x,y)$ are as above and $|\hat{e}(t,x,y)| \le C\exp(-\lambda/t)$.

Proof. Choose and fix some $(x,y) \in M \times M$. Let Γ and Γ_ϵ be as above with ϵ small enough such that Γ_ϵ is a strictly positive distance from $\mathrm{Cut}(x)$ and $\mathrm{Cut}(y)$. The Chapman–Kolmogorov equation gives

$$p_{2t}(x,y) = \int_{\Gamma_\epsilon} p_t(x,z)p_t(z,y)dz + \int_{M\setminus\Gamma_\epsilon} p_t(x,z)p_t(z,y)dz.$$

Using Theorem 2.1 and Equation (2.1), we have

$$p_{2t}(x,y) = \int_{\Gamma_\epsilon}\left(\frac{1}{2\pi t}\right)^n \exp\left[-\frac{h_{x,y}(z)}{t}\right]k(t,x,z)k(t,y,z)dz + \int_{M\setminus\Gamma_\epsilon}\exp\left[-\frac{1}{t}\left(h_{x,y}(z)+\delta(t,x,y)\right)\right]dz.$$

Observe that $h_{x,y}(z)$ is continuous and achieves its minimum precisely on Γ. Hence its minimum on $M\setminus\Gamma_\epsilon$ is strictly greater than its minimum on $\Gamma_{\epsilon/2}$. It follows that there exist positive λ and C such that

$$p_{2t}(x,y) = [1+e(t,x,y)]\int_{\Gamma_\epsilon}\left(\frac{1}{2\pi T}\right)^n \times \exp\left[-\frac{h_{x,y}(z)}{t}\right]k(t,x,z)k(t,y,z)dz, \tag{2.4}$$

where $|e(t,x,y)| \leq C\exp(-\lambda/t)$ (see [10, Lemma 5.3.1] of Hsu's book, where he gives essentially this result with a more detailed proof).

Again use the Chapman–Kolmogorov equation to write

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

Taking derivatives (all derivatives are on the y variable) gives

$$\nabla_A p_{2t}(x,y) = \int_M p_t(x,z)\left[\nabla_A p_t(z,y)\right]dz.$$

We divide the manifold into three regions. Let $\epsilon > 0$, Γ and Γ_ϵ be as before. Let C_ϵ be an ϵ-neighborhood around the cut locus of x. We now demand that ϵ also be small enough that these sets are a strictly positive distance from one another. Finally, let $R_\epsilon = M\backslash(C_\epsilon \cup \Gamma_\epsilon)$ be the rest of M. Then

$$\nabla_A p_{2t}(x,y) = \int_{\Gamma_\epsilon \cup R_\epsilon} p_t(x,z)\left[\nabla_A p_t(z,y)\right]dz + \int_{C_\epsilon} p_t(x,z)\left[\nabla_A p_t(z,y)\right]dz.$$

On Γ_ϵ and R_ϵ we are at a strictly positive distance from the cut locus. Also, note that

$$\nabla_A p_t(z,y) = p_t(z,y)\nabla_A \log p_t(z,y).$$

Then we have

$$\begin{aligned}\nabla_A p_{2t}(x,y) = &\int_{C_\epsilon} p_t(x,z)\left[\nabla_A p_t(z,y)\right]dz + \int_{\Gamma_\epsilon \cup R_\epsilon} \frac{1}{t} l(t,z,y,A) \\ &\times \left(\frac{1}{2\pi t}\right)^n \exp\left[-\frac{h_{x,y}(z)}{t}\right] k(t,x,z)k(t,y,z)dz.\end{aligned}$$

So now the problem is to control the last integral. For this we use Equations (2.1) and (2.3). This gives us the bound

$$\begin{aligned}\left|\int_{C_\epsilon} p_t(x,z)\left[\nabla_A p_t(z,y)\right]dz\right| \leq D_1 &\left[\frac{\operatorname{diam}(M)}{t} + \frac{1}{\sqrt{t}}\right] \\ &\times \int_{C_\epsilon} \exp\left[-\frac{1}{t}\left(h_{x,y}(z) + \delta(t,x,y)\right)\right]dz.\end{aligned}$$

To get the log gradient, we need to divide through by $p_{2t}(x,y)$, which we write as an integral using Equation (2.4). We now claim that the terms involving R_ϵ and C_ϵ vanish exponentially fast. Indeed, we have already seen

this during the derivation of Equation (2.4). Thus we have

$$\nabla_A \log p_{2t}(x,y) = (1+e(t,x,y))^{-1} \times \left\{ \frac{1}{t} \frac{\int_{\Gamma_\epsilon} l(t,z,y,A) \exp\left[-(h_{x,y}(z)/t)\right] k(t,x,z)k(t,y,z)dz}{\int_{\Gamma_\epsilon} \exp\left[-(h_{x,y}(z)/t)\right] k(t,x,z)k(t,y,z)dz} + \hat{e}(t,x,y) \right\},$$

where $e(t,x,y)$ and $\hat{e}(t,x,y)$ are bounded in absolute value by $C\exp(-\lambda/T)$, for some positive C and λ (perhaps different from above). Further, we have that $1/(1+e(t,x,y)) = 1 + \mathrm{O}(e(t,x,y))$. Hence, by adjusting $\hat{e}(t,x,y)$, C and λ, we can get rid of $e(T,x,y)$. Recalling the definition of $E_t(x,y)$, we have proved the theorem. □

2.3. The Hessian

Here we develop the analogous formula for the Hessian. By polarization, it is sufficient to consider $\nabla^2_{A,A} E_{2t}(x,y)$.

Theorem 2.4. *Let M be a smooth, compact, connected Riemannian manifold. Choose any two distinct points x and y. Then there exist positive constants C and λ (possibly different from the constants in Theorem* 2.3*) such that, for any $A \in T_yM$, we have*

$$\begin{aligned}\nabla^2_{A,A} E_{2t}(x,y) = -\frac{1}{t} \Bigg\{ & \frac{\int_{\Gamma_\epsilon} l^2(t,z,y,A) \exp\left[-(h_{x,y}(z)/t)\right] k(t,x,z)k(t,y,z)dz}{\int_{\Gamma_\epsilon} \exp\left[-(h_{x,y}(z)/t)\right] k(t,x,z)k(t,y,z)dz} \\ & - \left[\frac{\int_{\Gamma_\epsilon} l(t,z,y,A) \exp[-(h_{x,y}(z)/t)] k(t,x,z)k(t,y,z)dz}{\int_{\Gamma_\epsilon} \exp\left[-(h_{x,y}(z)/t)\right] k(t,x,z)k(t,y,z)dz} \right]^2 \Bigg\} \\ & - \frac{\int_{\Gamma_\epsilon} \nabla_A l(t,z,y,A) \exp\left[-(h_{x,y}(z)/t)\right] k(t,x,z)k(t,y,z)dz}{\int_{\Gamma_\epsilon} \exp\left[-(h_{x,y}(z)/t)\right] k(t,x,z)k(t,y,z)dz} \\ & + \hat{e}(T,x,y),\end{aligned}$$

where $l(t,x,y,A)$ and $k(t,x,y)$ are as above and $|\hat{e}(t,x,y)| \le C\exp(-\lambda/t)$.

Proof. We begin by observing that

$$\nabla^2_{A,A} \log f = \frac{\nabla^2_{A,A} f}{f} - (\nabla_A \log f)^2. \tag{2.5}$$

Since we have computed the log gradient of the heat kernel in our proof of Theorem 2.3, the only thing remaining is for us to compute the Hessian of

the heat kernel. Again we start with the Chapman–Kolmogorov equation and differentiate under the integral sign to get

$$\begin{aligned}
\nabla^2_{A,A} p_{2t}(x,y) &= \int_M p_t(x,z)\nabla^2_{A,A} p_t(z,y)dz \\
&= \int_{\Gamma_\epsilon \cup R_\epsilon} \left[\nabla^2_{A,A}\log p_t(z,y) + (\log p_t(z,y))^2\right] p_t(z,y)p_t(z,x)dz \\
&\quad + \int_{C_\epsilon} p_t(z,x)\nabla^2_{A,A} p_t(z,y)dz \\
&= \int_{\Gamma_\epsilon \cup R_\epsilon} \left[\frac{1}{t}\nabla_A l(t,z,y,A) + \frac{1}{t^2}\left(l(t,z,y,A)\right)^2\right] \\
&\quad \times \left(\frac{1}{2\pi t}\right)^n \exp\left[-\frac{h_{x,y}(z)}{t}\right] k(t,x,z)k(t,y,z)dz \\
&\quad + \int_{C_\epsilon} p_t(z,x)\nabla^2_{A,A} p_t(z,y)dz.
\end{aligned}$$

We wish to divide both sides by $p_{2t}(x,y)$, since that is what appears in the expansion of the log Hessian. We can use Equation (2.3) to control the integral over C_ϵ (in particular, it decays exponentially), and the integral over R_ϵ also decays exponentially, for the same reasons as before. Thus (refer to Theorem 2.3)

$$\begin{aligned}
&\frac{\nabla^2_{A,A} p_{2t}(x,y)}{p_{2t}(x,y)} \\
&= \frac{1}{1+e(t,x,y)}\left\{\frac{1}{t}\frac{\int_{\Gamma_\epsilon}\nabla_A l(t,z,y,A)\exp\left[-(h_{x,y}(z)/t)\right]k(t,x,z)k(t,y,z)dz}{\int_{\Gamma_\epsilon}\exp\left[-(h_{x,y}(z)/t)\right]k(t,x,z)k(t,y,z)dz}\right. \\
&\quad \left. + \frac{1}{t^2}\frac{\int_{\Gamma_\epsilon}\left[l(t,z,y,A)\right]^2\exp\left[-(h_{x,y}(z)/t)\right]k(t,x,z)k(t,y,z)dz}{\int_{\Gamma_\epsilon}\exp\left[-(h_{x,y}(z)/t)\right]k(t,x,z)k(t,y,z)dz} + \hat{e}(t,x,y)\right\}
\end{aligned}$$

where we change C and λ as necessary.

We now plug all of our results from above into Equation (2.5). The theorem then follows. □

These theorems work well when we wish to compute the asymptotics of the log gradient or log Hessian with respect to fixed x and y. However, if we wish to study how the asymptotics change as y moves, say into the cut locus, then they will not be of much use. This is because Γ_ϵ and λ can change discontinuously in y, which we can see just by looking at $\mathbb{S}^2$ and letting y move into $\mathrm{Cut}(x)$. Fortunately, our derivation of the theorem

makes it clear how to solve this problem. Let $\mathcal{O}$ be the union of the sets Γ associated to every $y \in \mathrm{Cut}(x)$. Then $\mathcal{O}$ is still a uniformly positive distance from $\mathrm{Cut}(x)$. Let $\mathcal{O}_\epsilon$ be an ϵ-neighborhood, chosen small enough to still be a uniformly positive distance from $\mathrm{Cut}(x)$. Then if we choose $y \in B_\epsilon(\mathrm{Cut}(x))$, everything we have done above works with Γ and Γ_ϵ replaced by $\mathcal{O}$ and $\mathcal{O}_\epsilon$. This will allow us to reduce all questions of what happens when y is moved to studying how our integral operators change. Said informally, all we have done is take all the parts of R which will become relevant as we move y and make them part of $\mathcal{O}$, our new region of interest. Thus all of the important behavior takes places in $\mathcal{O}$. This modification is worth formalizing.

Corrolary 2.5. *Let M and x be as above. Then for any $y \in B_\epsilon(\mathrm{Cut}(x))$, the expansions in Theorems* 2.3 *and* 2.4 *hold with Γ_ϵ replaced throughout by $\mathcal{O}_\epsilon$.*

3. The leading terms

Theorems 2.3 and 2.4 give, in principle, the complete asymptotic expansions of the gradient and Hessian of $E_t(x, y)$, up to terms which vanish faster than any power of t. However, if we restrict our attention to the first few terms of these expansions, the formulas simplify considerably.

3.1. Formulation in terms of expectation and variance

Considering the leading terms for the gradient and Hessian of $E_t(x, y)$, we have

$$\nabla_A E_t(x,y) = 2\left\{\frac{\int_{\Gamma_\epsilon} \nabla_A E(z,y)\exp[-(2/t)h_{x,y}(z)]\,H_0(x,z)H_0(y,z)dz}{\int_{\Gamma_\epsilon}\exp[-(2/t)h_{x,y}(z)]\,H_0(x,z)H_0(y,z)dz}\right\} + \mathrm{O}(t).$$

and

$$\nabla^2_{A,A} E_t(x,y) = -\frac{4}{t}\left\{\frac{\int_{\Gamma_\epsilon}(\nabla_A E(z,y))^2\exp[-(2/t)h_{x,y}(z)]\,H_0(x,z)H_0(y,z)dz}{\int_{\Gamma_\epsilon}\exp[-(2/t)h_{x,y}(z)]\,H_0(x,z)H_0(y,z)dz} - \left[\frac{\int_{\Gamma_\epsilon}\nabla_A E(z,y)\exp[-(2/t)h_{x,y}(z)]\,H_0(x,z)H_0(y,z)dz}{\int_{\Gamma_\epsilon}\exp[-(2/t)h_{x,y}(z)]\,H_0(x,z)H_0(y,z)dz}\right]^2\right\} + \mathrm{O}(1).$$

Note that we are now looking at $E_t(x,y)$ at time t rather than at time $2t$.

To begin, we can give a probabilistic interpretation of the constant term of the gradient and the $1/t$ term of the Hessian. Let

$$\mu_t(dz) = \frac{1_{\Gamma_\epsilon}(z)}{Z_t} H(x,z)H(y,z)\exp\left(-\frac{2h_{x,y}(z)}{t}\right)dz, \tag{3.1}$$

where

$$Z_t = \int_{\Gamma_\epsilon} H(x,z)H(y,z)\exp\left(-\frac{2h_{x,y}(z)}{t}\right)dz.$$

Then μ_t is a probability measure supported on Γ_ϵ, and the above becomes

$$\nabla_A E_t(x,y) = 2\mathbb{E}^{\mu_t}\left[\nabla_A E(z,y)\right] + \mathrm{O}(t)$$

and

$$\nabla^2_{A,A} E_t(x,y) = -\frac{4}{t}\operatorname{Var}^{\mu_t}\left[\nabla_A E(z,y)\right] + \mathrm{O}(1).$$

Since Γ_ϵ is compact, so is the space of probability measures supported on Γ_ϵ (in the weak topology). In particular, if we take any sequence of times decreasing to 0, then it will have a subsequence t_i such that μ_{t_i} converges to some limit probability measure μ supported on Γ. Let M_0 be the set of all such limit measures. For any $\mu \in M_0$, we will say that t_i is an associated sequence of times if μ_{t_i} converges (weakly) to μ.

We know that

$$\begin{aligned}\nabla_A E(z,y) &= \operatorname{dist}(z,y)\left\langle A, \nabla \operatorname{dist}(z,y)\right\rangle \\ &= \tfrac{1}{2}\operatorname{dist}(x,y)\left\langle A, Y(z)\right\rangle\end{aligned}$$

for $z \in \Gamma$, where $Y(z)$ is the (unique) unit vector at y such that

$$\exp_y\left[-Y(z)\operatorname{dist}(z,y)\right] = z.$$

Further, if we let $\theta_A(z)$ be the angle between A and $Y(z)$, then we can write $\langle A, Y(z)\rangle$ as $|A|\cos\theta_A(z)$. Thus, if we choose any $\mu \in M_0$ and let t_i be an associated sequence of times, we have

$$\lim_{i\to\infty} \nabla_A E_{t_i}(x,y) = |A|\operatorname{dist}(x,y)\mathbb{E}^{\mu}\left[\cos\theta_A(z)\right]$$

and

$$\lim_{i\to\infty} t_i \nabla^2_{A,A} E_{t_i}(x,y) = -|A|^2\operatorname{dist}(x,y)^2\operatorname{Var}^{\mu}\left[\cos\theta_A(z)\right].$$

Some elementary facts about the log gradient and log Hessian follow immediately. If we homothetically scale M by a factor of $a > 0$, then $\lim_{i\to\infty} \nabla_A E_{t_i}(x,y)$ is multiplied by a and $\lim_{i\to\infty} t_i \nabla^2_{A,A} E_{t_i}(x,y)$ is multiplied by a^2. Also, we have the pair of inequalities

$$-|A|\,\mathrm{dist}(x,y) \le \liminf_{t\searrow 0} \nabla_A E_t(x,y) \le \limsup_{t\searrow 0} \nabla_A E_t(x,y) \le |A|\,\mathrm{dist}(x,y)$$

and

$$-|A|^2\,\mathrm{dist}(x,y)^2 \le \liminf_{t\searrow 0} t\nabla^2_{A,A} E_t(x,y) \le \limsup_{t\searrow 0} t\nabla^2_{A,A} E_t(x,y) \le 0.$$

3.2. Relation to path-space integration

There is a one-to-one correspondence between Γ and the set of minimal geodesics from x to y. This suggests that we think of $z \in \Gamma$ as parameterizing these minimal geodesics, of any $\mu \in M_0$ as a measure on the set of minimal geodesics, and of $\cos\theta_A(z)$ as a function on the minimal geodesics (in particular, $\cos\theta_A(z)$ is the cosine of the angle between A and the geodesic corresponding to z). This viewpoint can be fleshed out by considering the Brownian bridge. (Intuitively, the Brownian bridge from x to y at time t is the stochastic process obtained from Brownian motion by conditioning on the particle starting at x and being at y at time t; for a more detailed discussion see, for example, Hsu's book [10].) In particular, fix x and y, and let P_t be the measure on path space corresponding to the Brownian bridge from x to y at time t. It is well known that P_t exists and that its finite marginal distributions are given in terms of the heat kernel. To be precise, let X_τ be the map from path space to M which sends each path to its position at time $\tau \in [0,t]$. Then for any finite sequence of times $0 = \tau_0 < \cdots < \tau_m < \tau_{m+1} = t$, the joint distribution of $X_{\tau_1}, \ldots, X_{\tau_m}$ under P_t is

$$\frac{1}{p_t(x,y)} \prod_{i=0}^{m} p_{\tau_{i+1}-\tau_i}(x_i, x_{i+1})$$

where, of course, $x_0 = x$ and $x_{m+1} = y$. Now consider $X_{t/2}$, and let ν_t be the distribution of $X_{t/2}$ under P_t. Then the density of ν_t with respect to the volume measure on M is

$$\frac{d\nu_t}{d\ \mathrm{vol}}(z) = \frac{p_{t/2}(x,z)p_{t/2}(z,y)}{p_t(x,y)}.$$

In order to study the limiting behavior, we integrate against a smooth test function $\varphi(z)$. Also, we use Equation (2.4) in the denominator to get

$$\mathbb{E}^{\nu_t}[\varphi] = \frac{\int_M \varphi(z) p_{t/2}(x,z) p_{t/2}(z,y) dz}{[1+e(t,x,y)] \int_{\Gamma_\epsilon} (1/\pi t)^n \exp[-(2/t) h_{x,y}(z)]\, k(t/2,x,z) k(t/2,y,z) dz}.$$

The contribution from the integral over $M \backslash \Gamma_\epsilon$ in the numerator vanishes exponentially, and on Γ_ϵ we can use the Pleijel expansion. Proceeding as before, we conclude that

$$\mathbb{E}^{\nu_t}[\varphi] = \frac{\int_{\Gamma_\epsilon} \varphi(z) \exp[-(2/t) h_{x,y}(z)]\, H_0(x,z) H_0(y,z) dz}{\int_{\Gamma_\epsilon} \exp[-(2/t) h_{x,y}(z)]\, H_0(x,z) H_0(y,z) dz} + \mathrm{O}(t).$$

It follows that $\mu_{t_i} \to \mu$ if and only if $\nu_{t_i} \to \mu$. So not only can we think of μ as a measure on the set of minimal geodesics, but we can also say that it is the natural such measure induced by the Brownian bridge.

The preceding allows us to view our method as a finite-dimensional analog of that used by Stroock and Malliavin, in which they work directly with the measure P_t on path space. According to the heuristics of Feynman-type path integrals, Wiener measure should be thought of as the probability measure on path space given by weighting each path $w(\tau)$ by a weight proportional to

$$\exp\left(-\frac{1}{2t}\int_0^1 |w'(\tau)|^2 d\tau\right)$$

(even though this is not possible in a rigorous sense). If we consider paths from x to y at time t, then as $t \searrow 0$ the above indicates that P_t, which is Wiener measure conditioned to require that the path be at y at time t, should be concentrating on paths that minimize energy. These paths are, of course, the minimal geodesics joining x and y. Thus in the limit, we expect the infinite dimensional path-space picture to collapse down to a finite-dimensional picture on minimal geodesics. Understanding this collapse on path space is somewhat difficult, but the present approach avoids this by working directly on the manifold from the beginning.

We will study the limiting measure (or measures) almost entirely in the context of geometric analysis, only occasionally remarking on the probabilistic interpretation. Nonetheless, the probabilistically inclined reader is encouraged to think about the limiting measure μ (when it exists) as giving the probability that a Brownian particle traveling from x to y "instantaneously" does so via a particular minimal geodesic.

3.3. An example

One application of Theorem 2.4 is the explicit computation of the asymptotics of $\nabla^2 E_t(x,y)$ when $y \in \mathrm{Cut}(x)$. (Of course, one expects such an explicit computation to be possible only in special cases.) Here we show how this can be done on $\mathbb{S}^n$. For simplicity, assume that $n \geq 2$. We note that $y \in \mathrm{Cut}(x)$ if and only if x and y are antipodal points. Hence, without loss of generality, we take x and y to be the north and south poles, which we denote N and S. Then Γ is the equatorial sphere $\mathbb{S}^{n-1}(1)$. (By $\mathbb{S}^n(r)$, we mean the standard n-dimensional sphere of radius r.) By symmetry, we see that μ_t converges to the uniform probability measure on the equatorial sphere (with respect to the induced volume measure). Next, let A be any unit vector in $T_y M$ (it doesn't matter which one, again by symmetry). Decomposing the equatorial sphere into level sets of $\theta_A(z)$, we see that the level set for any θ is $\mathbb{S}^{n-2}(\sin\theta)$.

We know that the gradient of $E_t(N,S)$ is zero by symmetry. Thus we proceed to computing the Hessian. Let ω_m denote the volume of the unit sphere of dimension m. We have

$$\mathbb{E}^{\mu}\left[\cos^2\theta_A(z)\right] = \frac{1}{\omega_{n-1}}\int_{\theta=0}^{\pi}\left(\frac{\pi}{2}\cos\theta\right)^2(\sin\theta)^{n-2}\omega_{n-2}d\theta$$

and

$$\mathbb{E}^{\mu}\left[\cos\theta_A(z)\right]^2 = \left(\frac{\omega_{n-2}}{\omega_{n-1}}\right)^2\frac{\pi^2}{4}\left(\int_{\theta=0}^{\pi}\cos\theta(\sin\theta)^{n-2}d\theta\right)^2.$$

This second term vanishes because $\cos\theta$ is anti-symmetric about $\pi/2$ while $\sin\theta$ is symmetric. Using this in our formula for the Hessian gives

$$\begin{aligned}\lim_{t\searrow 0} t\left[\nabla^2_{A,A}E_t(N,S)\right] &= \frac{\omega_{n-2}\pi^2}{\omega_{n-1}}\int_{\theta=0}^{\pi}(\cos\theta)^2(\sin\theta)^{n-2}d\theta\\ &= \frac{\pi^2}{n}.\end{aligned}$$

We conclude that $\nabla^2_{A,A}E_t(N,S) \sim -\frac{\pi^2}{nt}|A|^2$ as $t \searrow 0$ for any $A \in T_S M$ (the above computation assumes that $n \geq 2$, but this formula extends to the case $n = 1$ as can be checked easily by hand).

3.4. The relation to conjugate points

In the case of $\mathbb{S}^n$, we were able to determine the limiting measure μ using only symmetry considerations. In general, this will not be so easy, and the limiting measures (or measures) will depend on the behavior of $h_{x,y}$ near its minima. In particular, the limiting measures will be governed by whether or not these minima are degenerate (in the sense of Morse theory, that is, whether or not the Hessian is positive definite) and, if so, how degenerate they are. However, before discussing the relationship between degeneracy and the asymptotics of μ_t, we wish to relate this degeneracy to the geodesic geometry of the manifold.

We begin by introducing some terminology. Given a smooth, real-valued function f which is defined in a neighborhood of the origin in $\mathbb{R}^n$ and a non-negative integer m, we will say *f is constant to exactly order m at the origin in the direction $\xi \in \mathbb{S}^{n-1}$* if

$$(\partial_t)^i \left[f(t\xi) - f(0)\right]_{t=0}$$

is zero for $0 \leq i \leq m$, but is non-zero for $i = m+1$. In particular, f is constant to exactly order 0 in the direction ξ if its first derivative in the direction ξ is non-zero, f is constant to exactly order 1 if its first derivative is zero but not its second, and so on. We will say that f is constant to finite order in the direction ξ if there exists some non-negative integer m with f constant to exactly order m, and we will say that f is constant to order at least m in the direction ξ if the derivatives above vanish for $0 \leq i \leq m$.

Now, let γ be a minimal geodesic from x to y, and take $(r, \theta_1, \dots, \theta_{n-1})$ to be a polar coordinate system on T_xM such that $\gamma(r) = \exp_x(r, 0, \dots, 0)$ for $r \in [0, \operatorname{dist}(x,y)]$. We then say that γ is *conjugate to exactly order m in the direction $\xi \in \mathbb{S}^{n-2}$* if $\theta \rightsquigarrow \exp_x(\operatorname{dist}(x,y), \theta)$ is constant to exactly order m in the direction ξ (here $\mathbb{S}^{n-2}$ is the unit tangent space in the tangent space to $\mathbb{S}^{n-1}$ at $\theta = 0$ and thus the Jacobi field induced by ξ is perpendicular to γ). In particular, γ is conjugate to exactly order 0 in the direction ξ if it is not conjugate, in the usual sense, in this direction, and γ is conjugate to some positive order if it is conjugate in the usual sense, with the order of conjugacy indicating how many derivatives of the exponential map vanish in that direction. We will use the terms conjugate to finite order and conjugate to order at least m analogously to the above.

The relationship between the degeneracy of the minima of $h_{x,y}$ and the conjugacy of the corresponding geodesics is contained in the following lemma.

Lemma 3.1. *Choose distinct points x and y on M and let $d = \operatorname{dist}(x, y)$. Let $(r, \theta_1, \ldots, \theta_{n-1})$ and γ be as above, and choose some $\xi \in \mathbb{S}^{n-2}$. The coordinates $(r, \theta_1, \ldots, \theta_{n-1})$ on the tangent space induce coordinates in a neighborhood of γ under the exponential map, and thus we can also ask to what order $h_{x,y}$ is constant at $\exp(d/2, 0, \ldots, 0)$ in the direction ξ. Then $h_{x,y}$ is constant to finite order in the direction ξ if and only if γ is conjugate to finite order in the direction ξ. In this case, there exists a non-negative integer m such that $h_{x,y}$ is constant to exactly order $2m + 1$ at $\exp(d/2, 0, \ldots, 0)$ in the direction ξ and γ is conjugate to exactly order $2m$ in the direction ξ.*

Proof. Let $\varphi(t)$ be the angle between the geodesic from $\exp(d/2, t\xi)$ to $\exp(d, t\xi)$ and the geodesic from $\exp(d/2, t\xi)$ to $\exp(d, 0) = y$. Let

$$\rho(t) = \operatorname{dist}(\exp(d, t\xi), y).$$

Then, considering the dependence of geodesics on their starting points in the tangent bundle, we see that $\varphi(t)$ and $\rho(t)$ are comparable for small t. Thus, $\varphi(t)$ is constant to exactly order l at 0 (in the direction ∂_t) if and only $\rho(t)$ is constant to exactly order l at 0 (in the direction ∂_t). From the definition of $\rho(t)$, it is clear that $\rho(t)$ is constant to exactly order l at 0 if and only γ is conjugate to exactly order l in the direction ξ.

The distance from x to $\exp(d/2, t\xi)$ is independent of t, and the vector field given by pushing ξ forward by the exponential map is always perpendicular to the geodesic from x to $\exp(d/2, t\xi)$. Thus, $h_{x,y}(\exp(d/2, t\xi))$ depends only on the distance between $\exp(d/2, t\xi)$ and y. Further, we see that the inner product between the push-forward of ξ and the unit tangent to the geodesic from $\exp(d/2, t\xi)$ to y at $\exp(d/2, t\xi)$ is constant to exactly order l (in the direction ∂_t at $t = 0$) if and only if $\varphi(t)$ is constant to exactly order l. Since this inner product is proportional to the derivative of $h_{x,y}(\exp(d/2, t\xi))$ with respect to t, it follows that $\varphi(t)$ is constant to exactly order l if and only if $h_{x,y}$ is constant to exactly order $l + 1$ in the direction ξ.

Combining these facts, we see that γ is conjugate to exactly order l in the direction ξ if and only if $h_{x,y}$ is constant to exactly order $l + 1$ in the direction ξ. The lemma will be proved once we determine that $h_{x,y}$ can only vanish to odd order. This, however, is just a restatement of the fact that the first non-zero derivative of $h_{x,y}$ in any direction must be even because $h_{x,y}$ has a local minimum at $\exp(d/2, t\xi)$. □

In particular, this lemma implies that if $z \in \Gamma$, then z is a non-degenerate minimum of $h_{x,y}$ if and only if x and y are conjugate to exactly order 0 in all directions, which, as mentioned, is the same as saying that x and y are not

conjugate along γ in the usual sense. On the other hand, if z is a degenerate minimum, then x and y are conjugate (in the usual sense) along the minimal geodesic through z, and furthermore, the index and orders of conjugacy[1] can be determined from the partial derivatives of $h_{x,y}$.

3.5. Laplace asymptotics and the resolution of singularities

Having given a geometric interpretation of the minima of $h_{x,y}$, we now turn to the investigation of the relationship between the minima of $h_{x,y}$ and the limiting measures M_0. To do so, we will need to delve into the theory of Laplace asymptotics. We begin by considering the case when Γ consists of finitely many points, say $z_1, \dots, z_m$. For ease of notation, we write $g(z) = 2h_{x,y}(z) - E(x,y)$; hence g is non-negative and has zeroes precisely at the z_i. By taking ϵ small enough, we can ensure that Γ_ϵ is the union of the disjoint balls $B_\epsilon(z_i)$, where i ranges from 1 to m. In this case, we have that

$$\mathbb{E}^{\mu_t}[\varphi] = \frac{1}{Z(t)} \sum_{i=1}^{m} \int_{B_\epsilon(z_i)} \varphi(z) H_0(x,z) H_0(y,z) e^{-g(z)/t} dz$$

where

$$Z(t) = \sum_{i=1}^{m} \int_{B_\epsilon(z_i)} H_0(x,z) H_0(y,z) e^{-g(z)/t} dz.$$

We are thus lead to study integrals of the form

$$\int_{B_\epsilon(z_i)} \varphi(z) e^{-g(z)/t} dz \tag{3.2}$$

as $t \searrow 0$.

First, suppose that g can be diagonalized at z_i, that is, suppose that we can find coordinates $u_1, \dots, u_n$ around z_i such that

$$g(u_1, \dots, u_n) = \sum_{j=1}^{n} u_j^{2k_j} \tag{3.3}$$

for some positive integers $k_1 \leq \cdots \leq k_n$. Of course, at a non-degenerate minimum, the Morse lemma guarantees the existence of such coordinates

[1]Some authors use the order of a conjugate point to denote the dimension of the null space of the differential of the exponential map. We will, taking our cue from Morse theory, instead call that the index of the conjugate point and reserve the term order to denote the order to which the differential of the exponential map vanishes in some direction.

with $k_j = 1$ for each j, but in general this need not be true. Under the assumption that (3.3) holds, Estrada and Kanwal [7] give the full asymptotic expansion of (3.4); namely,

$$\begin{aligned}(3.4)\qquad &\int_{B_\epsilon(z_i)} \varphi(u_1,\dots,u_n)\exp\left[-\frac{\sum_{j=1}^n u_j^{2k_j}}{t}\right] du_1\cdots du_n \\ &\quad\sim \prod_{j=1}^n \left[\sum_{i_j=0}^\infty \frac{\Gamma\left((2i_j+1)/2k_j\right)}{(2i_j)!k_j} t^{(2i_j+1)/2k_j}\left(\frac{\partial}{\partial u_j}\right)^{2i_j}\right]\varphi(0,\dots,0)\end{aligned}$$

where derivatives on the right-hand side are to be understood as giving partial differential operators which are then applied to φ and evaluated at the origin (that is, at z_i). We will be interested mainly in the first term (that is, the one coming from $i_j = 0$ for all j), in which case Equation (3.4) gives

$$\begin{aligned}(3.5)\qquad &\int_{B_\epsilon(z_i)} \varphi(z)e^{-g(z)/t}dz \\ &\qquad = t^{1/2k_{1,i}+\cdots+1/2k_{n,i}}\left[c_i\ \mathrm{vol}_u(z_i)\varphi(z_i) + \mathrm{O}\left(t^{1/k_{n,i}}\right)\right],\end{aligned}$$

where vol_u is the volume element in the u coordinate chart and c_i is a positive constant which depends only on n and the $k_{j,i}$s.

Equation (3.5) implies that the limit is dominated by those geodesics which are "the most conjugate", in the sense of being conjugate in many directions and/or to high order. More precisely, assume that g can be diagonalized around each of its minima. Then to each z_i we can associate the order of the leading term of the integral over $B_\epsilon(z_i)$, which is $l_i = 1/2k_{1,i} + \cdots + 1/2k_{n,i}$. Then, as $t \searrow 0$, μ_t converges to a limit μ which is supported on those z_i with the smallest leading order (that is, z_i with $l_i = \min\{l_1,\dots,l_m\}$). Further, the mass at these points is given by the coefficient of the leading term of the expansion coming from (3.5), normalized to have total mass one. In terms of the Brownian bridge, this says that the particle prefers to travel along the most conjugate geodesics. More precisely, if we require the particle to travel from x to y "instantaneously", then it travels along the geodesic through z_i with probability $\mu\{z_i\}$. To be a bit more intuitive, it is not be surprising that a particle under Brownian motion prefers conjugate geodesics, since conjugacy should make a geodesic more forgiving of the white noise which the particle experiences as it tries to follow the geodesic.

In order to extend our analysis beyond the diagonalizable case, we will need to introduce resolutions of singularities. Let X be a real-analytic manifold of dimension n (we do not require that X be compact), and let f be a real-analytic function on X which is not identically zero. Then, for our purposes, a resolution of singularities will mean an n-dimensional analytic manifold Y and a proper, surjective analytic map $\xi : Y \to X$ such that ξ possesses the following two properties.

- At each point of the pre-image of the zero level-set of f, there exist local coordinates with respect to which $f \circ \xi$ is a monomial and the Jacobian of ξ is a monomial times a non-zero smooth function.
- The function ξ is a diffeomorphism outside of the zero level-set of f.

A famous theorem of Hironaka states that, for X and f as above, a resolution of singularities always exists.

Remark. The original results of Hironaka are both much more general than what we have stated above and formulated in the language of schemes. A similar definition to the above is given in Arnold, Gusein-Zade, and Varchenko's book [1], which is also concerned with the evaluation of Laplace asymptotics. An "elementary", constructive proof in the case of characteristic zero (which includes the case of real-analytic manifolds) is given by Bierstone and Milman [4]. Finally, a purely analytic statement and proof of a weaker form of the resolution of singularities for real-analytic manifolds (in particular, without the property that φ is a diffeomorphism off of the zero level-set) is given by Sussmann [20].

As a first application of the resolution of singularities, we can prove the following theorem.

Theorem 3.2. *Let M be a real-analytic, compact Riemannian manifold. Then for any distinct points x and y of M, the corresponding sequence of measures μ_t (see Equation (3.1)) described above converges to a unique limit μ as $t \searrow 0$.*

Proof. That M is real-analytic implies that the function g defined above is real-analytic in a neighborhood of Γ, and thus for an appropriate choice of ϵ, on Γ_ϵ. Let φ be any smooth function with compact support in Γ_ϵ. It is enough to show that $\mathbb{E}^{\mu_t}[\varphi]$ converges as $t \searrow 0$, since the limit of μ_t is determined by its action on smooth test functions.

Following [1], we define an elementary Laplace integral over a bounded, open neighborhood of $0 \in \mathbb{R}^n$ as an integral of the form

$$\int_U \exp\left(-x_1^{k_1} \cdots x_n^{k_n}/t\right) \left|x_1^{l_1} \cdots x_n^{l_n}\right| \psi(x_1, \ldots, x_n) dx_1 \ldots dx_n$$

where the k_i are non-negative, even integers, at least one of which is non-zero, the m_i are non-negative integers, and ψ is a smooth function of compact support in U (we call ψ the amplitude function). By [1, Theorem 7.4], for any elementary Laplace integral there exists a distribution a, a positive rational number α and a non-negative integer m such that this integral is asymptotic to $a(\psi)t^\alpha |\log t|^m$ as $t \searrow 0$. Further, if ψ is non-negative on U and positive at the origin, then $a(\psi)$ is positive.

We now apply the resolution of singularities to g and Γ_ϵ (where we think of Γ_ϵ as a non-compact real-analytic manifold). Thus we have a real-analytic manifold Y and a map $\xi : Y \to \Gamma_\epsilon$ with the properties listed above. Because Γ (which is the zero level-set of g) has measure zero and ξ is a diffeomorphism elsewhere, it follows that an integral over Γ_ϵ can be written as an integral over Y by pulling back everything on $\Gamma_\epsilon \setminus \Gamma$ (and ignoring the preimage of Γ in Y, which has measure zero under the pull-back of the volume form). In particular, we have that

$$Z(t) = \int_Y (H_0(x, \cdot) \circ \xi)(u)(H_0(y, \cdot) \circ \xi)(u) \exp\left(-g \circ \xi(u)/t\right) \xi^*(d \text{ vol})(u).$$

Next, observe that restricting the integral to the preimage of the closed $\epsilon/2$ neighborhood of Γ only introduces an exponentially small error; and further, the same is true if we integrate over any set intermediate between an $\epsilon/2$ neighborhood and Γ_ϵ. Since this closed neighborhood is compact, so is its preimage under ξ. Hence it can be covered by finitely many coordinate charts (each of which is also contained in the preimage of Γ_ϵ) of the type described above (namely, g is a monomial and the Jacobian of ξ is a monomial times a smooth, non-zero function). Using a partition of unity subordinate to this cover, we can write $Z(t)$ as a sum of finitely many, say l, elementary Laplace integrals (up to exponentially small error). In addition, we see that the amplitude function ψ_i in each of these integrals is non-negative and positive at the origin (in the local coordinates) because H_0, the volume form, and the partition functions have these properties. Thus, we see that the integrals over these charts are asymptotic to $a_i(\psi_i)t^{\alpha_i}|\log t|^{m_i}$, respectively, for $i = 0, \ldots, l$, where the $a_i(\psi_i)$ are all positive. Picking out the dominant such term or terms, we conclude that there is some positive real number

c, some positive rational number α, and some non-negative integer m such that

$$Z(t) \sim ct^{\alpha} |\log t|^m \quad \text{as } t \searrow 0.$$

In order to determine the numerator in $\mathbb{E}^{\mu_t}[\varphi]$, we follow the same procedure. The only difference is that now the amplitude functions in the elementary Laplace integrals also include the pull-back of the test function φ as a factor. In particular, we have the same coordinate charts as before, and now our amplitude functions are given by $\tilde{\psi}_i = \psi_i(\varphi \circ \xi)$. We can no longer guarantee that the $\tilde{\psi}_i$ are non-negative and positive at the origins of our charts, and thus we cannot guarantee that $a_i(\tilde{\psi}_i)$ is non-zero. However, note that we need only concern ourselves with the charts where $\alpha_i = \alpha$ and $m_i = m$. If the sum of the $a_i(\tilde{\varphi}_i)$ corresponding to these charts is non-zero, then call it $\tilde{c}$ and observe that

$$\int_{\Gamma_\epsilon} \varphi(z) H_0(x,z) H_0(y,z) e^{-g(z)/t} dz \sim \tilde{c} t^{\alpha} |\log t|^m \quad \text{as } t \searrow 0.$$

On the other hand, if the sum of these $a_i(\tilde{\varphi}_i)$ is zero, we cannot necessarily determine the leading term of the expansion for this integral; however, we can assert that this integral is $o(t^{\alpha}|\log t|^m)$. In either case, dividing by $Z(t)$ and letting $t \searrow 0$ shows that $\mathbb{E}^{\mu_t}[\varphi]$ converges to some finite limit. As mentioned above, this completes the proof. □

Note that this result says nothing about how to determine the limit μ, or more generally, about how it relates to the geodesic geometry of M. We suspect, although we have not proven, that if M is only assumed to be smooth, the limiting measure need not be unique.

3.6. Newton polyhedra and evaluation of the limit

Our use of the resolution of singularities in the last section was non-constructive. In the present section, we discuss how additional assumptions about g allow one to say much more about the limit measure μ.

We return to the case where M is smooth and g has finitely many minima, which we denote by z_i, although now we drop the assumption that g can be diagonalized near the z_i. In a series of papers (for example, [22]) which culminate in the monograph [1], Arnold and his school have provided a fairly complete analysis of the asymptotic expansion of Equation (3.2) in this case. We summarize the needed results. First, we need to assume that g vanishes to finite order at z_i. This is always true in the real-analytic category, but in the smooth category it need not be the case. Given that g

vanishes to finite order at z_i, we can in fact assume that g is real-analytic in a neighborhood of g by taking an appropriate change of coordinates. In particular, there exist coordinates u_j around z_i such that g is equal to its Taylor expansion in these coordinates.

Before stating the results, we need to introduce some notation. Let $\mathbb{Z}_+^n$ be the set of all n-tuples of non-negative integers. Also, we will use the multi-index notation such that for $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{Z}_+^n$ we interpret u^α as the monomial $u_1^{\alpha_1} \cdots u_n^{\alpha_n}$. Then we have that

$$g = \sum_{\alpha \in \mathbb{Z}_+^n} g_\alpha u^\alpha,$$

where the $g_\alpha \in \mathbb{R}$ are the coefficients in the Taylor expansion. Now let $\mathbb{R}_+^n$ be the set of all n-tuples of non-negative reals, and consider $\mathbb{Z}_+^n$ to be a subset of $\mathbb{R}_+^n$ in the natural way. Then the Newton polytope of g is defined to be the subset of $\mathbb{R}_+^n$ determined by taking the convex hull of $\cup(\alpha + \mathbb{R}_+^n)$ where the union is over all $\alpha \in \mathbb{Z}_+^n$ such that $g_\alpha \neq 0$. Further, the Newton diagram of g, which we will denote by $\Delta(g)$, is defined to be the union of the compact faces of the Newton polytope. (Strictly speaking, both the Newton polytope and Newton diagram of g also depend on our choice of coordinates u_i, but for now we will simply consider the coordinates to be given.)

Let γ be a face of $\Delta(g)$. Then we define g_γ to be the power series consisting of the monomials whose multi-indices lie on γ with the same coefficients as appear in the power series of g. That is,

$$g_\gamma = \sum_{\alpha \in \gamma \cap \mathbb{Z}_+^n} g_\alpha u^\alpha.$$

Note that if γ is compact, then g_γ is a polynomial. Similarly, we define the principal part of g, denoted g_Δ, by

$$g_\Delta = \sum_{\alpha \in \Delta(g) \cap \mathbb{Z}_+^n} g_\alpha u^\alpha.$$

We say that g is non-degenerate if for every compact face γ of $\Delta(g)$, the polynomials

$$\frac{\partial g_\gamma}{\partial x_1}, \ldots, \frac{\partial g_\gamma}{\partial x_n}$$

have no common zeroes in $(\mathbb{R} \setminus 0)^n$. We now take a closer look at $\Delta(g)$. Consider the ray $(r, \ldots, r)$ for $r > 0$. This ray intersects $\Delta(g)$ in exactly

one point, say $(r_0, \dots, r_0)$. We define $p(\Delta(g)) = 1/r_0$ to be the remoteness of the Newton diagram of g. Now let $k(\Delta(g))$ be the number of degrees of freedom of the supporting hyperplane of $\Delta(g)$ at the point $(r_0, \dots, r_0)$; we call $k(\Delta(g))$ the multiplicity of the Newton diagram of g.

We are now in a position to state the main result. Combining [1, Theorems 7.6 and 8.6] and [22, Theorem 2.6], we have the following.

Theorem 3.3. *Let g be an analytic function defined in a closed neighborhood $\overline{B_\epsilon(0)} \subset \mathbb{R}^n$ of the origin with a unique minimum of finite order at the origin. Assume that $g(0) = 0$. Let φ be a smooth function on $\overline{B_\epsilon(0)}$ such that $\varphi(0) \neq 0$. Then there exist an integer m between 0 and $n-1$ inclusive, a positive rational number α and a positive real number c, all depending only on g, such that*

$$\int_{B_\epsilon(0)} \varphi(z) e^{-g(z)/t} dz = c\varphi(0) t^\alpha |\log t|^m + \mathrm{O}(t^\alpha |\log t|^{m-1})$$

as $t \searrow 0$. Consider the Newton diagram of g and its remoteness and multiplicity, as defined above. Then

$$t^\alpha |\log t|^m \geq t^{p(\Delta(g))} |\log t|^{k(\Delta(g))}$$

for small enough t. Finally, if g is non-degenerate, then in fact $\alpha = p(\Delta(g))$ and $m = k(\Delta(g))$.

While we will not prove this theorem, we will comment on its proof and application. As might be expected, the resolution of singularities plays an important role. In fact, the central observation is that in the non-degenerate case one can construct a resolution of singularities from the Newton polytope such that the leading term has the properties given in the theorem. As such, the above represents a special case of the results of the preceding section. A more direct approach is taken in [22], in which the lead term of the integral is computed directly under the assumption of non-degeneracy. The approach taken in [22] also proves the one-sided estimate on the lead term given in the theorem. On the other hand, this approach fails to show that the Laplace integral has an expansion of the desired kind in the degenerate case.

It is clear that Theorem 3.3 provides us with more precise information about Laplace integrals when the function g is non-degenerate. For this reason, it is helpful to know that degeneracy is rare, in some sense. In particular, [1, Lemma 6.1] asserts that, given a Newton diagram, the set of degenerate principal parts is a proper semi-algebraic subset of the space of all

principal parts corresponding to the given diagram, the complement of which is everywhere dense. Approaching the issue of degeneracy from a different direction, we could ask to what extent a change of coordinates might help. Recall that the Newton diagram of g and its associated features, including degeneracy, depend on the chosen coordinates. It is easy to come up with functions which are degenerate in one coordinate system, but which are non-degenerate in another. Unfortunately, trying to eliminate degeneracy by changing coordinates does not work in general; there exist functions which are degenerate in any coordinate system.[2]

We now return to the problem at hand, namely determining the limiting behavior of μ_t. If we assume that g vanishes to finite order at each of the points z_i, then we can apply Theorem 3.3 to conclude that around each z_i the Laplace integral (3.2) is asymptotic to a constant times $t^{\alpha_i}|\log t|^{m_i}$. Since $Z(t)$ is the sum of these terms, we see that μ_t converges to a limiting measure μ which is supported on those z_i with dominant leading term.

So far we have been assuming that Γ consists of a finite number of points. In general, though, Γ can be quite complex. While we are far from a result which covers all possibilities, we can generalize the preceding a bit more. In particular, suppose that Γ consists of a finite collection of isolated, smooth submanifolds (possibly with boundary) $N_1, \ldots, N_m$ of M. Then the integral over Γ_ϵ can be decomposed into integrals over the ϵ-neighborhoods $(N_i)_\epsilon$ of the individual N_i. Further, the integral over each $(N_i)_\epsilon$ can be written as an integral in the normal direction followed by an integral in the tangent direction. Hence, if g vanishes to finite order in the normal directions, we can treat the integral in the normal direction at each $z \in N_i$ by the above methods. In particular, if we assume that for each N_i the asymptotics of the integral in the normal direction has the same leading term $t^{\alpha_i}|\log t|^{m_i}$ at all points, then the integral over $(N_i)_\epsilon$ will have leading term $t^{\alpha_i}|\log t|^{m_i}$, and μ_t will converge to a measure μ which is supported on the union of those N_is with dominant leading term. Further, on each such N_i, μ will be absolutely continuous with respect to the induced volume measure on N_i.

While this analysis is fairly general, it certainly does not cover every case. One can easily come up with cases where, for example, Γ has accumulation

[2]The two-dimensional case is almost an exception. In [21], Varchenko proves that for an analytic function with an isolated minimum, there always exists what he calls *adapted coordinates*, in which the leading power of t is equal to the remoteness of the Newton diagram.

points or g vanishes to infinite order, and where determining the limiting behavior of μ_t would be quite difficult, if not impossible.

In a different vein, we can also use Theorem 3.3 to prove the following lemma, which we will need later.

Lemma 3.4. *Let M be a smooth, compact, Riemannian manifold, and let x and y be points in M such that $y \in \mathrm{Cut}(x)$. Suppose there is a point $w \in \Gamma$ and an associated sequence of times t_i decreasing to zero such that μ_{t_i} converges to the point mass at w as $i \to \infty$. Then $\nabla^2 h_{x,y}(w)$ is degenerate.*

Proof. Let γ be the minimal geodesic from x to y passing through w (so that w is the midpoint of γ). It follows from Lemma 3.1 that $\nabla^2 h_{x,y}(w)$ is degenerate if and only if γ is conjugate. For the remainder of the proof, we assume that $\nabla^2 h_{x,y}(w)$ is non-degenerate. Hence γ is not conjugate, and because $y \in \mathrm{Cut}(x)$, there must be at least one other minimal geodesic from x to y. Let $\hat{\gamma}$ be such a geodesic and let $\hat{w}$ be its midpoint. Then, because $\nabla^2 h_{x,y}(w)$ is non-degenerate, there are some positive constants δ and C such that

$$\begin{aligned}\frac{\mathbb{E}^{\mu_t}\left[B_\delta(\hat{w})\right]}{\mathbb{E}^{\mu_t}\left[B_\delta(w)\right]} &= \frac{\int_{B_\delta(\hat{w})} H_0(x,z)H_0(z,y)e^{-g(z)/t}dz}{\int_{B_\delta(w)} H_0(x,z)H_0(z,y)e^{-g(z)/t}dz} \\ &\geq C \quad \text{for all sufficiently small } t.\end{aligned}$$

This follows from the fact that, by multiplying g on $B_\delta(\hat{w})$ by some positive constant, g on $B_\delta(\hat{w})$ can be made smaller than g on $B_\delta(w)$ (where this comparison can be made by introducing normal coordinates on each of these balls), and, by further reducing this constant, any problem arising from the H_0 or the volume form can be overcome. Since δ can be chosen such that $B_\delta(\hat{w})$ and $B_\delta(w)$ are disjoint, this inequality shows that no limiting measure can be supported only at w. We have shown that if $\nabla^2 h_{x,y}(w)$ is non-degenerate, then no limit measure can be the point mass at w. This proves the lemma. □

3.7. A characterization of the cut locus

The results of the previous section show that μ can be a point mass, even though there are multiple minimal geodesics from x to y. For example, consider the case when there are two minimal geodesics from x to y, one of which is conjugate and one of which is not. Then the limiting measure will be a point mass at the middle of the conjugate geodesic, and the $1/t$

term will vanish. (In terms of the Brownian bridge, if we require the particle to travel from x to y "instantaneously", then with probability 1 it travels along the conjugate geodesic.) Thus, it is clear that the $1/t$ term in the asymptotics of $\nabla^2_{A,A} E_t(x,y)$ is not sufficient to determine when $y \in \mathrm{Cut}(x)$. On the other hand, we have the following result characterizing $\mathrm{Cut}(x)$ in terms of the asymptotics of $\nabla^2 E_t(x,y)$.

Theorem 3.5. *Let M be a compact, smooth Riemannian manifold, and let x and y be any two distinct points of M. We have that $y \notin \mathrm{Cut}(x)$ if and only if*

$$\lim_{t \searrow 0} \nabla^2 E_t(x,y) = \nabla^2 E(x,y)$$

and $y \in \mathrm{Cut}(x)$ if and only if

$$\limsup_{t \searrow 0} \|\nabla^2 E_t(x,y)\| = \infty,$$

where $\|\nabla^2 E_t(x,y)\|$ is the operator norm, that is, the max of $|\nabla^2_{A,A} E_t(x,y)|$ over all $A \in T_y M$ with unit length. Further, if M is real-analytic, we have the stronger result that $y \in \mathrm{Cut}(x)$ if and only if

$$\lim_{t \searrow 0} \|\nabla^2 E_t(x,y)\| = \infty.$$

Proof. The case where $y \notin \mathrm{Cut}(x)$ is just a restatement of the result of Stroock and Malliavin mentioned in Equation (2.2). So we consider the case when $y \in \mathrm{Cut}(x)$. First suppose that the set of accumulation points of μ_t contains at least one measure μ which is not a point mass. Let t_i be an associated sequence of times such that $\mu_{t_i} \to \mu$. Because μ is not a point mass and the support of μ is disjoint from $\mathrm{Cut}(y)$, there is some unit vector $A \in T_y M$ such that $\nabla_A E(z,y)$ is not almost surely μ-constant. Hence, for this A, we will have a non-zero variance and so $\nabla^2_{A,A} \log p_{t_i}(x,y)$ will blow up like $1/t_i^2$.

Next, suppose that all of the accumulation points of μ_t are point masses. This will be the situation for the remainder of the proof. Take any w such that the point mass at w is an accumulation point of μ_t, and let t_i be an associated sequence of times (that is, the t_i are a sequence of times decreasing monotonely to zero such that μ_{t_i} converges to a point mass at w). Take a small ball $B_\delta(w)$ around w, and let $x_1, \ldots, x_n$ be any smooth coordinates around w defined on this ball such that $\partial_{x_1}, \ldots, \partial_{x_n}$ form an orthonormal

basis for T_wM. Then we wish to consider the function

$$V(t_i) = \max_{j\in\{1,\dots,n\}} \mathrm{Var}^{\mu_{t_i}}(x_j);$$

here we consider the x_j to be equal to zero outside of $B_\delta(w)$.

Define $\rho_{w,\delta}$ to be equal to $2[h_{x,y} - h_{x,y}(w)]$ on $B_\delta(w)$ and zero elsewhere, and let $\pi_{w,\delta}(s)$ be the measure (with respect to Riemannian volume) of the subset of $B_\delta(w)$ where $\rho_{w,\delta}$ is less than s. Note that $H_0(x,\cdot)H_0(\cdot,y)$ is bounded above and below by positive constants on Γ_ϵ. In particular, this means that, for the purposes of integration, we can absorb this factor into the volume form. Doing so greatly simplifies the formulas, and thus we will assume this is the case for the remainder of the proof. As a result, for the remainder of the proof we let $Z(t)$ be the integral of $\exp[-2(h_{x,y} - h_{x,y}(w))/t]$ over Γ_ϵ. (Note that $Z(t)$ is defined in terms of the integral over all of Γ_ϵ, instead of restricting to $B_\delta(w)$.) Finally, we set

$$f_{t_i}(x_1,\dots,x_n) = \exp[-\rho_{w,\delta}(x_1,\dots,x_n)/t_i]/Z(t_i)$$

on $B_\delta(w)$ and equal to zero elsewhere, so that f_{t_i} is the density of μ_{t_i} with respect to Riemannian volume on $B_\delta(w)$. (Technically, f_{t_i} also depends on w and δ, but to include this in the notation would be unmanageable.)

We wish to estimate $V(t_i)$ in terms of the measure of the set where f_{t_i} is at least half of its maximum value, $f_{t_i}(0,\dots,0) = 1/Z(t_i)$. Call this set $S_{1/2}(t_i)$. First, we observe that the measure of $S_{1/2}(t_i)$ is equal to $\pi_{w,\delta}(t_i \log 2)$. This follows from noting that

$$\begin{aligned}
&\left\{(x_1,\dots,x_n) : f_t(x_1,\dots,x_n) \geq \frac{1}{2} f_t(0,\dots,0)\right\} \\
&\quad = \left\{(x_1,\dots,x_n) : \frac{\exp[\rho_{w,\delta}(x_1,\dots,x_n)]}{Z(t)} \geq \frac{1}{2Z(t)}\right\} \\
&\quad = \{(x_1,\dots,x_n) : \rho_{w,\delta}(x_1,\dots,x_n) \leq t\log 2\}.
\end{aligned}$$

Next, we have that

$$V(t_i) \geq \max_{j\in\{1,\dots,n\}} \left[\min_{\alpha\in\mathbb{R}} \int_{S_{1/2}(t_i)} (x_j - \alpha)^2 \frac{f_{t_i}(0)}{2}\, \mathrm{vol}_M(x_1,\dots,x_n)dx_1\cdots dx_n\right].$$

Because the x_j are smooth and orthonormal at w, we can choose δ small enough so that the ratio of the Euclidean volume form determined by the x_j to the Riemannian volume form (which, we recall, includes the factor

$H_0(x,\cdot)H_0(\cdot,y))$ is bounded from below by $1/\sqrt{D}$ and from above by $\sqrt{D}$ on $B_\delta(w)$, for some positive constant D. Assuming that this is the case, we have that

$$V(t_i) \geq \max_{j\in\{1,\dots,n\}} \frac{1}{D} f_{t_i}(0) \left[\min_{\alpha\in\mathbb{R}} \int_{S_{1/2}(t_i)} (x_j - \alpha)^2 dx_1 \cdots dx_n \right]$$

and the measure of $S_{1/2}(t_i)$ with respect to the Euclidean volume induced by the x_j (which we are now integrating against) is at least $\pi_\delta(t_i \log 2)/2$.

Now suppose we take the Steiner symmetrization (see [17] for the definition and basic properties of Steiner symmetrization) of $S_{1/2}(t_i)$ with respect to x_1. Then it is clear that the above integral with respect to x_1 is now minimized by taking $\alpha = 0$ and that it can only have decreased after the symmetrization, while the integral with respect to any other coordinate remains unchanged. Hence, if we symmetrize with respect to all of the coordinates and call the resulting set $\mathrm{Sym}(S_{1/2}(t_i))$, we see that

$$V(t_i) \geq \max_{j\in\{1,\dots,n\}} \frac{1}{D} f_{t_i}(0) \int_{\mathrm{Sym}(S_{1/2}(t_i))} x_j^2 dx_1 \cdots dx_n.$$

Summing over j and writing $r = x_1^2 + \cdots + x_n^2$, we have that

$$V(t_i) \geq \frac{1}{Dn} f_{t_i}(0) \int_{\mathrm{Sym}(S_{1/2}(t_i))} r^2 dx_1 \cdots dx_n.$$

We recall that Steiner symmetrization preserves the Lebesgue measure of sets, and thus the above integral is being taken over a set of measure at least $\pi_{w,\delta}(t_i \log 2)/2$. On the other hand, it is clear that for a given measure for $\mathrm{Sym}(S_{1/2}(t_i))$, this integral is minimized when $\mathrm{Sym}(S_{1/2}(t_i))$ is the ball of the appropriate measure centered at the origin. Computing the integral of r^2 over such a ball, we conclude that

$$V(t_i) \geq \frac{\mathrm{Area}(\mathbb{S}^{n-1})}{Dn^2(n+2)(2\omega_n)^{(n+2)/n}} \left(\pi_{w,\delta}(t_i \log 2)\right)^{(n+2)/n} f_{t_i}(0). \tag{3.6}$$

Note that Equation (3.6) is valid (possibly with different constants) for any w with associated times t_i (meaning that μ_{t_i} converges to a point mass at w) and smooth coordinates x_j which are orthonormal at w, once we choose small enough δ and small enough t_i.

Our next task is to compare the asymptotics of $\pi_{w,\delta}(t_i \log 2)$ and $f_{t_i}(0) = 1/Z(t_i)$ as $t_i \searrow 0$. Let $\pi(s)$ be the analog of $\pi_{w,\delta}(s)$ over all of Γ_ϵ. Then for

any large enough $\alpha > 0$, we have (up to exponentially small error, which we ignore)

$$Z(t) = \pi(\alpha)e^{-\alpha/t} + \frac{1}{t}\int_0^\alpha \pi(s)e^{-s/t}ds.$$

Choose any $\beta > 0$. Then for small enough t,

$$\begin{aligned}\frac{1}{t}\int_0^\alpha \pi(s)e^{-s/t}ds &= \frac{1}{t}\int_0^{-\beta t\log t} \pi(s)e^{-s/t}ds + \frac{1}{t}\int_{-\beta t\log t}^\alpha \pi(s)e^{-s/t}ds \\ &\leq \pi(-\beta t\log t)\left(1-t^\beta\right) + \pi(\alpha)\left(t^\beta - e^{-\alpha/t}\right).\end{aligned}$$

It follows that, for small enough t, we have

$$Z(t) \leq \frac{1}{2}\pi(-\beta t\log t) + \pi(\alpha)t^\beta. \tag{3.7}$$

Next, we claim that $\pi(s)$ is bounded from above and below by positive powers of s (times a constant) for small enough s. To see that, first note that ρ is bounded from above by some non-negative function with a single zero and a non-degenerate Hessian at that zero, because the Hessian of ρ over all of Γ_ϵ is controlled from above by compactness. The volume of sublevel sets of this comparison function is easily seen to be asymptotic to a positive power of s, and thus $\pi(s)$ is bounded from below by this power of s. Next, we observe that the Hessian of ρ is always non-degenerate in the radial direction (in polar coordinates around x) and that the second derivative in this direction is bounded from below by a positive constant, again by compactness. Hence ρ is bounded from below by some quadratic function of the radius alone. This comparison function also has the volume of its sublevel sets asymptotic to a positive power of s, and this provides an upper bound for $\pi(s)$.

Now let a be the infimum of all positive reals b such that, for some constant c and small enough t, $ct^b \leq \pi(t)$. Choose any small $\overline{\epsilon} > 0$. Then there must be some sequence of times $\tau_i \searrow 0$ such that

$$\frac{\pi(\tau_i)}{\tau_i^{a-\overline{\epsilon}}} \searrow 0.$$

Let the β from Equation (3.7) be given by $\beta = a - \overline{\epsilon}$, and let $\overline{\tau}_i$ be a sequence such that $\tau_i = -\beta\overline{\tau}_i\log\overline{\tau}_i$ for all i. Then, applying Equation (3.7), we

have that

$$\begin{aligned} Z(\overline{\tau}_i) &\leq \pi(\tau_i) + \pi(\alpha)\overline{\tau}_i^{\beta} \\ &\leq \tau_i^{a-\overline{\epsilon}} + \pi(\alpha)\overline{\tau}_i^{a-\overline{\epsilon}}, \end{aligned}$$

which implies that, for some constant C, we have

$$\frac{1}{Z(\overline{\tau}_i)} \geq \frac{1}{C\overline{\tau}_i^{a-\overline{\epsilon}}|\log \overline{\tau}_i|^{a-\overline{\epsilon}}}. \tag{3.8}$$

By compactness, we can assume, after passing to a subsequence, that $\mu_{\overline{\tau}_i}$ converges to some limit measure, which must be a point mass. Thus, we can take $t_i = \overline{\tau}_i$.

In the remainder of the proof, we will assume that C is some appropriate constant, the exact value of which may vary from appearance to appearance. Using that $t_i = \overline{\tau}_i$ and applying Equations (3.6) and (3.7), we have that

$$V(t_i) \geq C\frac{\pi_{w,\delta}(t_i \log 2)^{(n+2)/n}}{t_i^{a-\overline{\epsilon}}|\log t_i|^{a-\overline{\epsilon}}}.$$

By the same argument as before, we know that $\pi(t\log 2)$ is the measure of the subset of Γ_ϵ where the density of μ_t is at least half of its maximum value. Then, because μ_{t_i} converges to a point mass at w, we must have that $2\pi_{w,\delta}(t_i \log 2) \geq \pi(t_i \log 2)$ for small enough t_i. Thus $\pi_{w,\delta}(t_i \log 2) \geq Ct_i^{a+\overline{\epsilon}}$ (for some potentially different constant C), and we conclude that

$$V(t_i) \geq C\frac{t_i^{2\overline{\epsilon}}}{|\log t_i|^{a-\overline{\epsilon}}}\pi_{w,\delta}(t_i \log 2)^{2/n}.$$

Finally, it remains to estimate the asymptotics of $\pi_{w,\delta}(t)$. Because we are assuming that μ_{t_i} converges to a point mass at w, Lemma 3.4 implies that the Hessian of ρ must be degenerate at w. Since the Hessian of ρ is degenerate at w, its Newton diagram must be dominated by the diagram of $x_1^4 + \sum_{i=2}^n x_i^2$, possibly after relabeling the coordinates. Hence our discussion of Laplace asymptotics (in particular, Theorem 3.3) implies that $\pi_{w,\delta}(t) \geq Ct^{(n/2)-D_n}$, where D_n is some positive constant depending only on the dimension n of M. If we choose $\overline{\epsilon} \leq D_n/2n$, then we see that

$$V(t_i) \geq C\frac{t_i^{1-D_n/n}}{|\log t_i|^{a-\overline{\epsilon}}}.$$

This proves that there is some point w and an associated sequence of times t_i such that, for any smooth coordinates x_j orthonormal at w, at least one of the x_j has variance that goes to zero slower than t_i.

Because $y \notin \mathrm{Cut}(w)$, we can choose vectors $A_1, \dots, A_n$ in T_yM such that the functions $\nabla_{A_j} E(z,y) - \nabla_{A_j} E(w,y)$ are smooth coordinates around w which are orthonormal at w. Thus the above argument applies, and we conclude that for some A_j, the variance of $\nabla_{A_j} E(z,y) - \nabla_{A_j} E(w,y)$ goes to zero slower than t_i. Normalize this A_j to have length one, and call the result A. Then $\nabla_A E(z,y)$ differs from $\nabla_{A_j} E(z,y) - \nabla_{A_j} E(w,y)$ only by an affine transformation, and thus its variance differs only by multiplication by some (positive) constant. We conclude that $\nabla_A E(z,y)$ has variance that goes to zero slower than t_i, and this completes the proof of the theorem for smooth manifolds.

In the case when M is real-analytic, the proof follows the same lines, but is somewhat simpler. First of all, we know that the limit measure μ is unique. If it is not a point mass the result follows just as before, only now we do not have to pass to a subsequence of time and thus our result holds for the limit, not just the limit supremum.

Now suppose that μ_t converges to point mass at some point w. Then we have, using Equation (3.6) and noting that we do not need to pass to a subsequence,

$$V(t) \geq \frac{\mathrm{Area}(\mathbb{S}^{n-1})}{4n^2(n+2)(2\omega_n)^{(n+2)/n}} \left(\pi_{w,\delta}(t \log 2)\right)^{(n+2)/n} f_t(0).$$

Again, the point is to compare the asymptotics of $\pi_\delta(t \log 2)$ and $f_t(0) = 1/Z(t)$. We observe that $1/Z$ is essentially determined by the Laplace transform of π_δ (see [22] for a discussion of this fact). So we are comparing a function with its Laplace transform. In the proof of Theorem 3.2, we showed that $Z(t) \sim ct^\alpha |\log t|^m$ for some $c > 0$, some non-negative rational number α, and some non-negative integer m. Direct computation shows that any function of this form is asymptotically equivalent (up to multiplication by some positive real number a) to the reciprocal of its Laplace transform. (That this is true for real-analytic functions but not necessarily for smooth functions is one reason for the increased difficulty in that case.) Thus we have that

$$V(t) \geq \frac{a\,\mathrm{Area}(\mathbb{S}^{n-1})}{4n^2(n+2)(2\omega_n)^{(n+2)/n}} \left(\pi_{w,\delta}(t \log 2)\right)^{2/n}.$$

Again, we know that $\pi_{w,\delta}(t) \geq Ct^{(n/2)-D_n}$. From here, the theorem follows just as above. □

3.8. Lower order leading terms

Theorem 3.5 shows that the Hessian of $E_t(x,y)$ always blows up on the cut locus. However, we have already seen that, if the limiting measure is a point mass, there will be no $1/t$ term in the expansion. In other words, the variance will go to zero, but it will to do more slowly than t. In this case, it is more difficult to determine the leading term of the expansion of $\nabla^2 E_t(x,y)$, since it involves further terms in the Laplace asymptotics. Because of this, we can only discuss the simplest case.

We consider the case where Γ consists of a single point, z_1, such that g can be diagonalized in a neighborhood of z_1. Recall this means that there exist coordinates u_j around z_1 such that

$$g(u_1,\dots,u_n) = \sum_{j=1}^{n} u_j^{2k_j}$$

for some positive integers $k_1 \leq \cdots \leq k_n$. Since we assume $y \in \mathrm{Cut}(x)$, this geodesic must be conjugate and thus $k_n \geq 2$. Further, since the Hessian of $h_{x,y}$ is non-degenerate in the radial direction, $k_1 = 1$. The advantage to this situation is that we have a full asymptotic expansion from Equation (3.4), rather than just the leading term (at least that is the advantage of assuming diagonalizability; assuming that Γ contains just a single point makes the computation tractable). Let l be the smallest index such that $k_l = k_n$. (Note that we allow $l = n$, but we must have $l \geq 2$.) Thus $u_l,\dots,u_n$ correspond to the directions of maximal conjugacy. Keeping the first two terms of the expansion in Equation (3.4), we see that

$$\begin{aligned}
&\int_{B_\epsilon(z_1)} \varphi(u_1,\dots,u_n) \exp\left[-\frac{\sum_{j=1}^n u_j^{2k_j}}{t}\right] du_1\cdots du_n \\
&\quad\sim \frac{\Gamma(1/2k_1)\cdots\Gamma(1/2k_n)}{k_1\cdots k_n} t^{1/2k_1+\cdots+1/2k_n}\varphi(0) \\
&\quad+ \sum_{j=l}^{n} \frac{\Gamma(1/2k_1)\cdots\Gamma(3/2k_j)\cdots\Gamma(1/2k_n)}{k_1\cdots(2k_j)\cdots k_n} t^{1/2k_1+\cdots+(3/2k_j)+\cdots+1/2k_n} \frac{\partial^2}{\partial u_j^2}\varphi(0).
\end{aligned}$$

We will be interested in the cases

$$\varphi = (\nabla_A E(z,y))^* H_0(x,z) H_0(z,y)\ \mathrm{vol}_u(z),$$

where $*$ denotes 1 or 2 (corresponding to the expectation of the square and the square of the expectation which appear in the variance of $\nabla_A E(z,y)$), and $\varphi = H_0(x,z)H_0(z,y)\ \mathrm{vol}_u(z)$ (which gives the expansion of $Z(t)$). Dividing the relevant expansions and keeping the first two terms shows that (here we use that $k_j = k_n$ for all $j = l, \dots, n$)

$$\mathbb{E}^{\mu_t}\left[(\nabla_A E(z,y))^*\right] \sim (\nabla_A E(0,y))^* + \frac{\Gamma(3/2k_n)}{2\Gamma(1/2k_n)} t^{1/k_n} \sum_{j=l}^{n} \left[\frac{\partial^2}{\partial u_j^2} (\nabla_A E(0,y))^* \right.$$
$$\left. + 2\frac{\frac{\partial}{\partial u_j}(\nabla_A E(0,y))^* \frac{\partial}{\partial u_j}\left(H_0(x,0)H_0(0,y)\ \mathrm{vol}_u(0)\right)}{H_0(x,0)H_0(0,y)\ \mathrm{vol}_u(0)} \right].$$

If we now compute the variance, most of this cancels, and we see that

$$\mathrm{Var}^{\mu_t}\left(\nabla_A E(z,y)\right) \sim \frac{\Gamma(3/2k_n)}{\Gamma(1/2k_n)} t^{1/k_n} \sum_{j=l}^{n} \left(\frac{\partial}{\partial u_j} \nabla_A E(0,y) \right)^2.$$

Finally, we compute that

$$\nabla^2_{A,A} E_t(x,y) \sim -\frac{4}{t^{1-1/k_n}} \frac{\Gamma(3/2k_n)}{\Gamma(1/2k_n)} \sum_{j=l}^{n} \left(\frac{\partial}{\partial u_j} \nabla_A E(0,y) \right)^2.$$

There are several things to observe regarding this formula. First of all, it shows that it is easy to produce situations in which $\nabla^2_{A,A} E_t(x,y)$ blows up at a rate intermediate between $1/t$ and 1, and that every rational of the form $-(m-1)/m$ for a positive integer m can be achieved as the order of the leading term. This gives concrete intuition to the results of Theorem 3.5. Second, in the case of a single minimal geodesic, we see that knowing the order of the leading term tells us the maximum order of degeneracy of the Hessian of $h_{x,y}$ and thus also "how conjugate" this minimal geodesic is, at least in the "most conjugate" directions. Finally, it is relatively easy to see that the coefficient of this leading term also has geometric significance. In particular, as a function of $A \in T_y M$, it is a symmetric, non-positive definite quadratic form such that the dimension of its kernel is equal to $n - l$. To see this, let $\gamma(z)$ be the minimal geodesic from $z \in B_\epsilon(z_1)$ to y and let $v_j \in T_y M$ be the derivative of the unit tangent to $\gamma(z_1)$ at y with

respect to $\partial/\partial u_j$. Then $\partial/\partial u_j \nabla_A \operatorname{dist}(z_1, y)$ is non-zero if and only if $\langle v_j, A\rangle$ is non-zero. Because $h_{x,y}$ is non-degenerate in the radial direction, we can assume that the $\partial/\partial u_j$ (for $j = l, \dots, n$) are perpendicular to $\gamma(z_1)$, and thus we see that

$$\frac{\partial}{\partial u_j}\nabla_A E(0, y) = \operatorname{dist}(z_1, y)\frac{\partial}{\partial u_j}\nabla_A \operatorname{dist}(z_1, y) \quad \text{for } j = l, \dots, n.$$

Observing that, for $j = l, \dots, n$, the v_j are linearly independent, the desired result follows. Thus, the leading term not only gives us the maximum order of degeneracy of the Hessian of $h_{x,y}$, but also tells us the dimension of the subspace on which this maximum degeneracy is achieved.

We find it interesting that, at least in this case, a term in the expansion further down than $1/t$ has such a nice geometric interpretation. Unfortunately, these terms are hard to compute in general (the present case of a single minimal geodesic is the easiest case, yet even here completely working out the above calculations is rather laborious), and we do not know anything more about them.

4. Mollification of energy

In the last section, we studied the asymptotics relative to fixed points x and y. Now, we turn our attention to considering how Corollary 2.5 can be used to study the distributional Hessian of $E(x, y)$, where we think of x as a fixed base point and thus of $E(x, y)$ as a function of y. In particular, Varadhan's result (see Equation (2.1)) implies that $E_t(x, y)$ is a smooth mollifier of $E(x, y)$ as $t \searrow 0$. Hence, computing the distributional limit of $\nabla^2_{A,A} E_t(x, y)$ as $t \searrow 0$ gives $\nabla^2_{A,A} E(x, y)$ as a distribution.

We note that the approach below is not the only way to study the distributional Hessian of $E(x, y)$. For example, many of these results are consequences of the fact that the cut locus is rectifiable with respect to $(n-1)$-dimensional Hausdorff measure, as proven by Mennucci [13] using viscosity solutions of Hamilton–Jacobi equations and geometric measure theory.

4.1. The results

Away from $\operatorname{Cut}(x)$, the distribution $\nabla^2_{A,A} E(x, y)$ is just a smooth function, and Equation (2.2) tells us that $\nabla^2_{A,A} E_t(x, y)$ converges to this limit uniformly on compact subsets of $M \setminus \operatorname{Cut}(x)$. This means that the singular part of $\nabla^2_{A,A} E(x, y)$, which we denote $\operatorname{sing}(\nabla^2_{A,A} E(x, y))$, is supported on $\operatorname{Cut}(x)$. Considering Theorem 2.4, we see that any contribution to the singular part

must come from the variance term. In particular, for any smooth function φ we have

$$\left\langle \varphi, \operatorname{sing}(\nabla^2_{A,A} E(x,\cdot)) \right\rangle = -\lim_{\epsilon \searrow 0} \lim_{t \searrow 0} \int_{B_\epsilon(\operatorname{Cut}(x))} \varphi(y) \tfrac{4}{t} \operatorname{Var}^{\mu_{t,y}}(\nabla_A E(\cdot, y)) dy,$$

where $\mu_{t,y}$ is the measure μ_t from above corresponding to the point y, the variance of $\nabla_A E(\cdot, y)$ is taken with respect to the first variable, and we use the notation $\langle \varphi, D \rangle$ to denote the action of the distribution D on the smooth function φ.

Let $(r, \theta) \in \mathbb{R}_+ \times \mathbb{S}^{n-1}$ be (normal) polar coordinates around x, and let $d(\theta)$ be the distance to the cut locus along the geodesic corresponding to θ. Let $U_x = \{(r,\theta) : \theta \in \mathbb{S}^{n-1}, r \in (d(\theta) - \epsilon, d(\theta))\}$. Then the exponential map gives a diffeomorphism from U_x to $M \setminus \operatorname{Cut}(x)$, and $\partial U_x = (d(\theta), \theta)$ is the tangential cut locus, that is, the preimage of $\operatorname{Cut}(x)$ under the exponential map (or more accurately, the connected component of the preimage closest to the origin). This gives a natural identification of $\mathbb{S}^{n-1}$ with the set of minimal geodesics from x to $\operatorname{Cut}(x)$ and with ∂U_x. We will frequently assume this identification, for example, when stating that some $\theta \in \mathbb{S}^{n-1}$ corresponds to a conjugate geodesic. Because $\operatorname{Cut}(x)$ has measure zero, we can write the above integral in polar coordinates on U_x as

(4.1)

$$\begin{aligned} &\left\langle \varphi, \operatorname{sing}(\nabla^2_{A,A} E(x,y)) \right\rangle \\ &\quad = -\lim_{\epsilon \searrow 0} \lim_{t \searrow 0} \int_{\mathbb{S}^{n-1}} \left[\int_{d(\theta)-\epsilon}^{d(\theta)} \varphi(r,\theta) \tfrac{4}{t} \operatorname{Var}^{\mu_{t,(r,\theta)}}(\nabla_A E(z,(r,\theta))) \operatorname{vol}(r,\theta) dr \right] d\theta. \end{aligned}$$

We are now in a position to state the following theorem.

Theorem 4.1. *Let M be a smooth, compact Riemannian manifold and let x be any point in M. Let A be any smooth vector field on M. Choose (normal) polar coordinates on T_xM and define U_x as above. Then the right-hand side of Equation (4.1) defines a negative measure on ∂U_x, which is absolutely continuous with respect to the measure $d\theta$ on ∂U_x obtained by identifying it with $\mathbb{S}^{n-1}$ via polar coordinates. Denote the corresponding Radon–Nikodym derivative by $\rho(\theta)$; then $\rho(\theta)$ is bounded. Thought of as a distribution on M, $\nabla^2_{A,A} E(x,y)$ has as its singular part a negative measure $\nu_{x,A}$ supported on $\operatorname{Cut}(x)$, and further, $\nu_{x,A}$ is given by the pushforward of $\rho(\theta)d\theta$ under the exponential map.*

While Theorem 4.1 shows that the singular part of $\nabla^2_{A,A}E(x,y))$ has a relatively nice structure, it says little about the relationship between ρ and the geodesic geometry of M. In order to describe the relationship, we will need a bit more notation. Let $C \subset \mathbb{S}^{n-1}$ be the set of all θ which correspond to conjugate geodesics. Next, say that the geodesics corresponding to θ and $\tilde{\theta}$ are *associated* if they lead to the same point in $\mathrm{Cut}(x)$ (that is, if $d(\theta) = d(\tilde{\theta})$ and $(d(\theta), \theta)$ and $(d(\tilde{\theta}), \tilde{\theta})$ are mapped to the same point under $\exp_x$). Let $P \subset \mathbb{S}^{n-1}$ be the set of $\theta \in \mathbb{S}^{n-1} \setminus C$ to which there is associated precisely one other $\tilde{\theta} \in \mathbb{S}^{n-1}$ and such that $\tilde{\theta} \notin C$. Finally, let $R = \mathbb{S}^{n-1} \setminus (C \cup P)$ (so R consists of non-conjugate θ which are associated to more than one other geodesic or which are associated to a conjugate geodesic). The three sets C, P and R are disjoint and partition $\mathbb{S}^{n-1}$.

Theorem 4.2. *Let the hypotheses be as in Theorem 4.1. If $\theta \in C$, then $\rho(\theta) = 0$. Also, R has measure zero as a subset of $\mathbb{S}^{n-1}$ with respect to $d\theta$, and ρ is continuous except possibly at points of R. Finally, there is an explicit expression for ρ on P (see Equation (4.2) below).*

In order to give the explicit expression for ρ on P, we will need still more notation. Let θ be in P, let $\tilde{\theta}$ be the (one) associated geodesic and let y be their common endpoint. Also, let z be the midpoint of the geodesic corresponding to θ, and let B be the (non-degenerate) Hessian of $h_{x,y}$ at z. Let $\tilde{z}$ and $\tilde{B}$ be the corresponding objects associated to $\tilde{\theta}$. Next, let $A_y \in T_yM$ be the value of the vector field A at y. Then let ψ be the angle between the geodesic given by θ and A_y, $\tilde{\psi}$ be the angle between the geodesic corresponding to $\tilde{\theta}$ and A_y, and φ the angle between the geodesics θ and $\tilde{\theta}$. Then for any $\theta \in P$,

$$\rho(\theta) = -\operatorname{dist}(x,y)|A_y|^2 \left(\cos\psi - \cos\tilde{\psi}\right)^2 \operatorname{vol}(d(\theta),\theta) \times \left[(1-\cos\varphi)\left(1 + \frac{H_0(x,z)H_0(y,z)\sqrt{\det\tilde{B}}}{H_0(x,\tilde{z})H_0(y,\tilde{z})\sqrt{\det B}}\right)\right]^{-1}. \tag{4.2}$$

Note that the volume element, all of the functions H_0 appearing above, and both B and $\tilde{B}$ can be computed from the Jacobi fields along the geodesics given by θ and $\tilde{\theta}$.

Theorem 4.2 tells us that the only contributions to the singular part come from points in P, which on M are places where locally the cut locus looks like a smooth hypersurface and the singular part of $\nabla^2 E(x,y)$ is just given by the jump discontinuity of $\nabla E(x,y)$ across this hypersurface. While the cut locus

itself can be quite complicated (for example, it may not be triangulable, as shown by Gluck and Singer [8]), the singular part of $\nabla^2 E(x,y)$ is supported only at those points with the nicest local structure.

We should point out, however, that even though there may not be any singular part of $\nabla^2 E(x,y)$ in a neighborhood of a conjugate point, the Hessian will not be smooth at a conjugate point. To be precise, suppose that $\gamma : [0, \operatorname{dist}(x,y)] \mapsto M$ is a minimal geodesic (with unit speed parametrization) from x to y. Then if we consider $\nabla^2 E(x, \gamma(s))$ as $s \nearrow \operatorname{dist}(x,y)$, we have that the Hessian, as an operator on vectors fields A, will blow up if and only if γ is a conjugate geodesic. Further, whether or not this blow up occurs for a given A depends on how A relates to the directions in which γ is conjugate (we give the precise formula in the next section).

This leads us to the following picture. The distributional Hessian of $E(x,y)$ is composed of two pieces, an L^1 function, which is just the regular Hessian on $M \setminus \operatorname{Cut}(x)$, and a singular part, which is the measure supported on $\operatorname{Cut}(x)$ as described above. At a point y which is the image on M of a point in P, the L^1 part stays bounded as we approach y along either minimal geodesic, but the singular measure is supported at y, as discussed above. On the other hand, if y is a conjugate point, then the singular measure may not be supported at y, but the L^1 part will blow up as we approach y along any conjugate geodesic. Because the set R has measure zero in the decomposition of the tangential cut locus, these two cases completely describe the non-smooth behavior of the distributional Hessian of $E(x,y)$.

4.2. The proofs

The proof of Theorems 4.1 and 4.2 will require a little preparation. The first thing we need to do is to justify exchanging the integration with the limits in Equation (4.1). Equation (4.1) immediately implies that the singular part is non-positive, and thus is a non-positive measure (as opposed to a higher order distribution). Because the singular part is a non-positive measure, we can estimate the Hessian in terms of the Laplacian. In particular, choose any closed, connected set $\Omega \subset \mathbb{S}^{n-1}$ the boundary of which has finite $(n-2)$-dimensional Hausdorff measure, and let $\Omega(\epsilon)$ be the set

$$\{(r,\theta) : \theta \in \Omega \text{ and } r \in [d(\theta) - \epsilon, d(\theta)]\}.$$

Then we have

$$\left|\int_{\Omega(\epsilon)} \varphi(r,\theta)\nabla^2_{A,A}E_t(x,y)\ \mathrm{vol}(r,\theta)drd\theta\right| \leq C_1|\varphi|_\infty|A|^2 \times\left\{\left|\int_{\Omega(\epsilon)}\Delta E_t(x,y)\ \mathrm{vol}(r,\theta)drd\theta\right| + \int_{\Omega(\epsilon)}\mathrm{vol}(r,\theta)drd\theta\right\}$$

for small enough t and ϵ and some constant C_1. Here $|f|_\infty$ denotes the L^∞ norm of f on the set $\Omega(\epsilon)$. This last integral can be estimated as

$$\int_{\Omega(\epsilon)}\mathrm{vol}(r,\theta)drd\theta \leq \epsilon|\Omega|_{\mathbb{S}^{n-1}}|\ \mathrm{vol}(r,\theta)|_\infty.$$

Let $n(r,\theta)$ be the outward pointing unit normal to $\Omega(\epsilon)$. We use integration by parts to write

$$\int_{\Omega(\epsilon)}\Delta E_t(x,y)\ \mathrm{vol}(r,\theta)drd\theta = -\int_{\partial\Omega(\epsilon)}\langle\nabla E_t(x,y), n(r,\theta)\rangle\, d\mathcal{H}^{n-1}_M,$$

where $\mathcal{H}^{n-1}_M$ denotes $(n-1)$-dimensional Hausdorff measure relative to the Riemannian volume measure. The right-hand side is a smooth function of t, so the only question is what happens to it as $t \searrow 0$. On the "walls" of $\partial\Omega(\epsilon)$, that is, the set

$$W(\Omega(\epsilon)) = \{(r,\theta) : \theta \in \partial\Omega \text{ and } r \in (d(\theta)-\epsilon, d(\theta))\},$$

the results of Stroock and Malliavin show that $\nabla E_t(x,y)$ converges (pointwise) to $\nabla E(x,y) = r\partial_r$. Further, the Gauss lemma shows that, on this set, $n(r,\theta)$ is perpendicular to ∂_r (both in the Riemannian metric and the Euclidean metric). Thus, the integral over the "walls" goes to zero with t, which we will write

$$\int_{W(\Omega(\epsilon))}\langle\nabla E_t(x,y), n(r,\theta)\rangle\, d\mathcal{H}^{n-1}_M = \mathrm{o}(1).$$

Next, we consider the integral over the "top" and "bottom" of $\partial\Omega(\epsilon)$, that is, the sets

$$T(\Omega(\epsilon)) = \{(r,\theta) : \theta \in \partial\Omega \text{ and } r = d(\theta)\}$$

and

$$B(\Omega(\epsilon)) = \{(r,\theta) : \theta \in \partial\Omega \text{ and } r = d(\theta)-\epsilon\}.$$

The results of Stroock and Turetsky show that the norm of $\nabla E_t(x,y)$ is bounded for all small t, and thus $\langle \nabla E_t(x,y), n(r,\theta)\rangle$ is bounded on both $T(\Omega(\epsilon))$ and $B(\Omega(\epsilon))$. Thus it remains only to control the $\mathcal{H}_M^{n-1}$ measure of $T(\Omega(\epsilon))$ and $B(\Omega(\epsilon))$.

First note that the volume density $\mathrm{vol}(r,\theta)$ is bounded, and thus we can estimate the $\mathcal{H}_M^{n-1}$ measure of a set from above by $|\mathrm{vol}(r,\theta)|_\infty$ times the $\mathcal{H}^{n-1}$ measure of the set, where we use $\mathcal{H}^{n-1}$ to indicate the $(n-1)$-dimensional Hausdorff measure relative to the Euclidean volume on the tangent space. The key to estimating this measure is the result of Itoh and Tanaka [11] that the distance to the cut locus (that is, the function $d(\theta)$) is Lipschitz. It follows that

$$\mathcal{H}^{n-1}\left(T(\Omega(\epsilon)) \cup B(\Omega(\epsilon))\right) \leq 2\left(1+\mathrm{Lip}(d)\right)|\Omega|_{\mathbb{S}^{n-1}}|d|_\infty.$$

Combining the above results, we conclude that

$$\begin{aligned}(4.3)\qquad &\left|\int_\Omega \left[\int_{d(\theta)-\epsilon}^{d(\theta)} \varphi(r,\theta)\nabla^2_{A,A}E_t(x,y)\ \mathrm{vol}(r,\theta)dr\right]d\theta\right| \\ &\quad \leq C_2|\varphi|_\infty|A|^2|\ \mathrm{vol}(r,\theta)|_\infty|\Omega|_{\mathbb{S}^{n-1}}\left[\left(1+\mathrm{Lip}(d)\right)|d|_\infty + \mathrm{o}(1)+\epsilon\right]\end{aligned}$$

for small enough t, ϵ and some constant C_2.

This justifies exchanging the limit as $t \searrow 0$ with the integral over θ in Equation (4.1). In addition, because the variance and volume density are always non-negative and φ is smooth, we can also exchange the limit as $\epsilon \searrow 0$ with the integral over θ. This shows that the measure given by the right-hand side of Equation (4.1) is absolutely continuous with respect to the measure $d\theta$ and that its Radon-Nikodym derivative ρ is given by

$$(4.4)\qquad \rho(\theta) = -\lim_{\epsilon\searrow 0}\lim_{t\searrow 0}\int_{d(\theta)-\epsilon}^{d(\theta)} \frac{4}{t}\,\mathrm{Var}^{\mu_{t,(r,\theta)}}(\nabla_A E(z,(r,\theta)))\ \mathrm{vol}(r,\theta)dr.$$

We now turn our attention to the sets C, R and P. On the set C, we want to show that $\rho = 0$. To do so, we will not work directly with (4.1), but rather with the definition of the distributional Hessian. In particular, choose any $\theta \in C$, and let $\Omega_\delta \subset \mathbb{S}^{n-1}$ be a disk of radius δ around θ. Then starting from (4.3) and letting t and ϵ both go to zero, we see that

$$-\int_{\Omega_\delta}\rho(\psi)d\psi \leq C_2|A|^2|\ \mathrm{vol}(d(\psi),\psi)|_\infty|\Omega_\delta|_{\mathbb{S}^{n-1}}\left(1+\mathrm{Lip}(d)\right)|d|_\infty.$$

Everything on the right-hand side is bounded from above, so if we divide both sides by $|\Omega_\delta|_{\mathbb{S}^{n-1}}$ and let δ go to zero, then Lebesgue's differentiation

theorem tells us that $-\rho(\theta)$ is almost everywhere on C less than or equal to a constant times the limit as $\delta \searrow 0$ of the L^∞-norm of $\mathrm{vol}(d(\psi), \psi)$ over Ω_δ. Because $\mathrm{vol}(d(\psi), \psi)$ is continuous and equals zero at θ (because the corresponding geodesic is conjugate), this shows that ρ is equal to zero almost everywhere on C. Because ρ is a density, its value only matters up to almost everywhere equivalence, and thus we can take ρ to be zero on all of C.

Next, we prove a lemma which shows that when computing ρ, we can ignore the set R.

Lemma 4.3. *For any compact manifold M and a basepoint $x \in M$, the corresponding set R, as defined above, has measure zero as a subset of $\mathbb{S}^{n-1}$ with its standard volume measure.*

Proof. As usual, we identify $\mathbb{S}^{n-1}$ with the set of minimal geodesics from x to $\mathrm{Cut}(x)$. Let S_1 be set of non-conjugate θ which are associated to more than one other geodesics, all of which are non-conjugate. Let S_2 be the set of non-conjugate θ which are associated to a conjugate geodesic (and possibly to other geodesics as well). Then $R = S_1 \cup S_2$.

We first consider S_1. In particular, choose any $\theta \in S_1$ and let $\theta_1, \ldots, \theta_k$ for some $k \geq 2$ be the associated geodesics (there are necessarily only finitely many because otherwise they would have an accumulation point, forcing at least one to be conjugate). Let y be their common endpoint. In terms of the exponential map, the fact that θ is not conjugate means that there is a neighborhood U of $(d(\theta), \theta)$ in T_xM which is diffeomorphic to a neighborhood V of y under the exponential map. The same is true for each of the θ_i and we use U_i for the corresponding subsets of T_xM. By choosing these neighborhoods small enough, we can assume that the preimage of V under the exponential map is precisely equal to $U \cup U_1 \cup \cdots \cup U_k$. Now we can define a smooth function f on V to be the length of the corresponding element of the tangent space in U, and we can similarly define the f_i. We will call such functions local distance functions. It follows that, for any point in V, the Riemannian distance is given by $\min\{f, f_1, \ldots, f_k\}$ and further, that the number of minimal geodesics to that point is given by the number of these local distance functions which achieve this minimum. At y, we know that all of these local distance functions achieve their common minimum, that they all have gradient of length one, and that these gradient vectors are all distinct. Given this, it is straight-forward to see that the set of points in V where at least three of the these local distance functions

achieve the common minimum is given by a finite union of smooth submanifolds of dimension no more than $n-2$. Since V and U are diffeomorphic under the exponential map, it follows that the set of points in U which correspond to points with more than two minimal geodesics has $(n-1)$-dimensional Hausdorff measure equal to zero. Projecting this set onto the θ coordinate gives the intersection of S_1 with a neighborhood of θ, and we conclude that this set has measure zero. This shows that every $\theta \in S_1$ has a neighborhood in which S_1 has measure zero, and it follows that all of S_1 has measure zero.

We now consider S_2. This will require some facts about the set of points in M which are conjugate to x along a minimal geodesic; call this set the conjugate-cut locus. Observe that the set of points in M corresponding to S_2 is contained in the conjugate-cut locus. Let C_1 be the set of all vectors in T_xM such that the kernel of $d\exp_x$ has dimension 1. In [23], Warner showed that C_1 is a smooth $(n-1)$-dimensional submanifold of T_xM. Let T be the set of points in C_1 such that the kernel of $d\exp_x$ is contained in the tangent space to C_1 at that point. Let H be the set of all vectors in T_xM such that the kernel of $d\exp_x$ has dimension at least 2. In the proof of [24, Lemma 1.1], Warner showed that the image under the exponential map of $T \cup H$ has $(n-1)$-dimensional Hausdorff measure equal to zero. In the proof of [9, Proposition 3.2], Hebda proved that a point of $C_1 \setminus T$ cannot correspond to a minimal geodesic. This means that the conjugate-cut locus is precisely the image of $T \cup H$ under $\exp_x$ and thus has $(n-1)$-dimensional Hausdorff measure equal to zero. Now choose any $\theta \in S_2$ and let y be the corresponding point in M. As before, we can choose neighborhoods U and V of $(d(\theta), \theta)$ and y such that the exponential map gives a diffeomorphism between them. In particular, this means that there is a neighborhood $W \subset \mathbb{S}^{n-1}$ of θ such that no two points in W correspond to the same point in M. Suppose $S_2 \cap W$ has positive measure. Then the corresponding set of points in U would have positive $(n-1)$-dimensional Hausdorff measure. But these points are contained in the conjugate-cut locus and so this contradicts the above. Hence we conclude that $S_2 \cap W$ has measure zero. Since this holds for any $\theta \in S_2$, it follows that all of S_2 has measure zero. Given the decomposition $R = S_1 \cup S_2$, this completes the proof of the lemma. □

In order to complete the proofs of Theorems 4.1 and 4.2, it suffices to compute ρ from Equation (4.4) on P. This is fairly straight-forward, if somewhat tedious. In this case, $\mathcal{O}_\epsilon$ consists just of balls around z and $\tilde{z}$. We begin by estimating $\mathbb{E}^{\mu_{t,\delta}}[f]$ for any smooth test function f and small

t and δ. First of all, let y_δ be the image under the exponential map of $(d(\theta)-\delta,\theta)$. Then let z_δ be the midpoint of the geodesic from x to y_δ in the direction θ, and let $\tilde{z}_\delta$ be the midpoint of the (non-minimizing, if $\delta>0$) geodesic from x to y_δ in the direction close to $\tilde{\theta}$. Because the exponential map is a diffeomorphism near $(d(\theta),\tilde{\theta})$, it follows that $\tilde{z}_\delta$ is well defined for small enough δ and depends smoothly on δ. We know that h_{x,y_δ} has non-degenerate Hessian at both z_δ and $\tilde{z}_\delta$, and we denote these Hessians by B_δ and $\tilde{B}_\delta$. An easy computation shows that the volume form associated to coordinates which diagonalize $2h_{x,y_\delta}$ around z_δ is $1/\sqrt{\det B_\delta}$, and similarly for $\tilde{z}_\delta$; note that B_δ is the Hessian of h_{x,y_δ}, without the factor of 2 (this is essentially a consequence of the $1/2$ which appears in the second order Taylor expansion). Then Equation (3.4) gives

$$\begin{aligned}
&\int_{\mathcal{O}_\epsilon} f(u)H_0(x,u)H(u,y_\delta)\exp\left[-\frac{2h_{x,y_\delta}(u)}{t}\right]du\\
&\quad=\Gamma\left(\frac{1}{2}\right)^n t^{n/2}\exp\left[-\frac{2h_{x,y_\delta}(z_\delta)}{t}\right]\left[f(z_\delta)\frac{H_0(x,z_\delta)H_0(z_\delta,y_\delta)}{\sqrt{\det B_\delta}}+\mathrm{O}_\delta(t)\right]\\
&\qquad+\Gamma\left(\frac{1}{2}\right)^n t^{n/2}\exp\left[-\frac{2h_{x,y_\delta}(\tilde{z}_\delta)}{t}\right]\left[f(\tilde{z}_\delta)\frac{H_0(x,\tilde{z}_\delta)H_0(\tilde{z}_\delta,y_\delta)}{\sqrt{\det \tilde{B}_\delta}}+\mathrm{O}_\delta(t)\right].
\end{aligned}$$

Here we use the notation $\mathrm{O}_\delta(t)$ to indicate that the error term as $t\searrow 0$ depends on δ, but does so uniformly for all sufficiently small δ; in particular, the integral of $\mathrm{O}_\delta(t)$ with respect to δ is $\mathrm{O}(t)$.

Let

$$F=\frac{H(x,\tilde{z})H(y,\tilde{z})\sqrt{\det B}}{H(x,z)H(y,z)\sqrt{\det \tilde{B}}}.$$

Also note that $f(z_\delta)=f(z)+\mathrm{O}(\delta)$, $f(\tilde{z}_\delta)=f(\tilde{z})+\mathrm{O}(\delta)$, and so on for H_0, h, and B. In addition, we will need to know how $2h_{x,y_\delta}(z_\delta)$ compares with $2h_{x,y_\delta}(\tilde{z}_\delta)$. Elementary trigonometry shows that this difference can be written as

$$2h_{x,y_\delta}(z_\delta)-2h_{x,y_\delta}(\tilde{z}_\delta)=\mathrm{dist}(x,y)\delta(1-\cos\varphi)+\mathrm{O}(\delta^2).$$

Then since $Z(t)$ is obtained simply by taking $f(u) \equiv 1$ in the above, we have that

$$\mathbb{E}^{\mu_{t,\delta}}[f] = \mathrm{O}_\delta(t) + \frac{f(z)+\mathrm{O}(\delta)+(f(\tilde{z})F+\mathrm{O}(\delta))\exp\left[-(1/t)(\delta\,\mathrm{dist}(x,y)(1-\cos\varphi)+\mathrm{O}(\delta^2))\right]}{1+(F+\mathrm{O}(\delta))\exp[-(1/t)(\delta\,\mathrm{dist}(x,y)(1-\cos\varphi)+\mathrm{O}(\delta^2))]}.$$

Using this, we can compute (after doing some algebra) that

$$\mathrm{Var}^{\mu_{t,\delta}}[\nabla_A E(\cdot,y)] = \left(\nabla_A E(z,y) - \nabla_A E(\tilde{z},y)\right)^2 (F+\mathrm{O}(\delta)) \times \frac{\exp\left[-(1/t)(\delta\,\mathrm{dist}(x,y)(1-\cos\varphi)+\mathrm{O}(\delta^2))\right]}{\{1+(F+\mathrm{O}(\delta))\exp[-(1/t)(\delta\,\mathrm{dist}(x,y)(1-\cos\varphi)+\mathrm{O}(\delta^2))]\}^2} + \mathrm{O}_\delta(t).$$

Writing $\mathrm{vol}(d(\theta)-\delta,\theta)$ as $\mathrm{vol}(d(\theta),\theta)+\mathrm{O}(\delta)$ and making the change of variables $\alpha = -\delta\,\mathrm{dist}(x,y)(1-\cos\varphi)/t$, Equation (4.4) gives

$$\rho(\theta) = \frac{4\left(\nabla_A E(z,y) - \nabla_A E(\tilde{z},y)\right)^2}{\mathrm{dist}(x,y)(1-\cos\varphi)}\,\mathrm{vol}(d(\theta),\theta) \times \lim_{\epsilon\searrow 0}\lim_{t\searrow 0}\int_{I(\epsilon,t)}\left[\frac{(F+\mathrm{O}(t\alpha))\,e^{\alpha+\mathrm{O}((t\alpha)^2)}}{\left\{1+(F+\mathrm{O}(t\alpha))\,e^{\alpha+\mathrm{O}((t\alpha)^2)}\right\}^2} + \mathrm{O}_\delta(t)\right][1+\mathrm{O}(t\alpha)]\,d\alpha$$

where $I(\epsilon,t)$ is the interval $[-\epsilon\,\mathrm{dist}(x,y)(1-\cos\varphi)/t, 0]$.

Taking both limits causes the region of integration to become $(-\infty,0]$. Further, it causes all of the $\mathrm{O}(\cdot)$ terms to vanish. To see this first note that $\mathrm{O}_\delta(t)$ vanishes uniformly with t. As for the $\mathrm{O}(t\alpha)$ and $\mathrm{O}((t\alpha)^2)$ terms, $\mathrm{O}(t\alpha)$ is bounded on $I(\epsilon,t)$ for all ϵ and t and goes to zero uniformly on any compact subinterval. Using this, one can show that they do not contribute in the limit (to see this in detail, one can make the further change of variables $\beta = Fe^\alpha$ and compute the integral). Thus, taking the limits, the above integral becomes

$$\int_{-\infty}^{0}\frac{Fe^\alpha}{\{1+Fe^\alpha\}^2}d\alpha = \frac{1}{1+F^{-1}}.$$

Using this in the above expression for ρ, along with the fact that

$$\left(\nabla_A E(z,y) - \nabla_A E(\tilde{z},y)\right)^2 = \tfrac{1}{4}\,\mathrm{dist}(x,y)^2\,|A_y|^2\left(\cos\psi - \cos\tilde{\psi}\right)^2,$$

gives Equation (4.2). This completes the proofs of Theorems 4.1 and 4.2.

Finally, we justify our earlier comments about the L^1 part of the distributional Hessian along conjugate geodesics. We begin by applying Theorem 2.4

at a point $y \notin \mathrm{Cut}(x)$. Because y is not in the cut locus, there is a single, non-conjugate minimal geodesic between x and y. Because this geodesic is not conjugate, $\nabla^2 2h_{x,y}(z)$ is non-degenerate, and we choose coordinates $u_1, \ldots, u_n$ around z such that $2h_{x,y} = u_1^2 + \cdots + u_n^2$. We already know that, since we are not on the cut locus, the leading term in the Hessian will be the constant term, and in order to compute this term we will need the first two terms in the expansion of Theorem 2.4. We begin by computing, using the first two terms of the expansion in Equation (3.4), that

(4.5)
$$\begin{aligned}
&\int_{B_\epsilon(z)} f(u) \exp\left[-\frac{2h_{x,y}(u)}{t}\right] k(t/2,x,u)k(t/2,y,u)du \\
&\quad = \exp\left[-\frac{2h_{x,y}(z)}{t}\right] t^{n/2} \left\{\Gamma(1/2)^n f(z)k(t/2,x,z)k(t/2,y,z)\ \mathrm{vol}_u(z)\right. \\
&\qquad \left. + t\frac{\Gamma(1/2)^{n-1}\Gamma(3/2)}{2}\Delta^u\left(f(z)k(t/2,x,z)k(t/2,y,z)\ \mathrm{vol}_u(z)\right) + \mathrm{O}(t^2)\right\}.
\end{aligned}$$

Here the symbol Δ^u means the operator $\sum_{i=1}^n \partial^2/\partial u_i^2$. Recall that

$$k(t,x,y) = H_0(x,y) + tH_1(x,y) + \mathrm{O}(t^2)$$

and

$$l(t,x,y,A) = -\nabla_A E(x,y) + t\nabla_A G_1(x,y) + \mathrm{O}(t^2).$$

Given Equation (4.5) and the expansions of l and k in terms of the H_i and the G_i, expanding the right-hand side of the equality in Theorem 2.4 becomes simply a lengthy exercise in manipulating Taylor series. We will not reproduce the computation here and will instead merely state the result. The coefficient of the $1/t$ term is zero, as we know it must be, and taking the limit as $t \searrow 0$ gives

$$\nabla^2_{A,A} E(x,y) = 2\left[\nabla^2_{A,A} E(z,y) - \sum_{i=1}^n \left(\frac{\partial}{\partial u_i}\nabla_A E(z,y)\right)^2\right].$$

If we consider what happens as y approaches $\mathrm{Cut}(x)$ along a geodesic γ, we see that $\nabla^2 E(x,y)$ will blow up, as an operator, if and only if γ is conjugate. This is because both $\nabla^2 E(z,y)$ and $\nabla E(z,y)$ remain bounded, and thus the only way for a blow up to occur is if at least one of the ∂_{u_i} has its (Riemannian) length blowing up. This occurs precisely if the corresponding eigenvalue of $\nabla^2 h_{x,y}(z)$ is going to zero, and thus precisely if γ is conjugate in the

direction corresponding to ∂_{u_i}. It is also easy to see that the blow up must be in the negative direction, and that the relationship between a given vector A and the ∂_{u_i} determines whether or not the Hessian blows up for a given A.

Acknowledgments

I would like to thank my advisor, Dan Stroock, for his invaluable suggestions throughout the course of this work. I also thank David Jerison and Joe Harris for helpful discussions and the anonymous referee for his or her comments.

References

[1] V.I. Arnol′d, S.M. Guseĭn-Zade and A.N. Varchenko, *Singularities of differentiable maps*, Vol. II, Monographs in Mathematics, **83**, Birkhäuser Boston Inc., Boston, MA, 1988. *Monodromy and asymptotics of integrals*, Translated from the Russian by Hugh Porteous, Translation revised by the authors and James Montaldi.

[2] G. Ben Arous, *Développement asymptotique du noyau de la chaleur hypoelliptique hors du cut-locus*, Ann. Sci. École Norm. Sup. (4) **21**(3)(1988), 307–331.

[3] Nicole Berline, Ezra Getzler and Michèle Vergne, *Heat kernels and Dirac operators*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], **298**, Springer-Verlag, Berlin, 1992.

[4] Edward Bierstone and Pierre D. Milman, *Canonical desingularization in characteristic zero by blowing up the maximum strata of a local invariant*, Invent. Math. **128**(2) (1997), 207–302.

[5] Richard L. Bishop, *Decomposition of cut loci*, Proc. Amer. Math. Soc. **65** (1) (1977), 133–136.

[6] Isaac Chavel, *Eigenvalues in Riemannian geometry*, Pure and Applied Mathematics, **115**, Academic Press Inc., Orlando, FL, 1984. Including a chapter by Burton Randol, With an appendix by Jozef Dodziuk.

[7] Ricardo Estrada and Ram P. Kanwal, *A distributional approach to asymptotics*: Theory and applications 2nd edn, Birkhäuser Advanced Texts: Basler Lehrbücher [Birkhäuser Advanced Texts: Basel Textbooks], Birkhäuser Boston Inc., Boston, MA, 2002.

[8] Herman Gluck and David Singer, *Deformations of geodesic fields*, Bull. Amer. Math. Soc. **82** (4) (1976), 571–574.

[9] James J. Hebda, *The local homology of cut loci in Riemannian manifolds*, Tôhoku Math. J. (2) **35**(1) (1983), 45–52.

[10] Elton P. Hsu, *Stochastic analysis on manifolds*, Graduate Studies in Mathematics, **38**, American Mathematical Society, Providence, RI, 2002.

[11] Jin-ichi Itoh and Minoru Tanaka, *The Lipschitz continuity of the distance function to the cut locus*, Trans. Amer. Math. Soc. **353**(1) (2001), 21–40.

[12] Paul Malliavin and Daniel W. Stroock, *Short time behavior of the heat kernel and its logarithmic derivatives*, J. Differential Geom. **44**(3) (1996), 550–570.

[13] Andrea C.G. Mennucci, *Regularity and variationality of solutions to Hamilton–Jacobi equations. I. Regularity*, ESAIM Control Optim. Calc. Var. **10**(3) (2004), 426–451 (electronic).

[14] S.Minakshisundaram and Å.Pleijel, *Some properties of the eigenfunctions of the Laplace-operator on Riemannian manifolds*, Canad. J. Math. **1** (1949), 242–256.

[15] S.A. Molčanov, *Diffusion processes, and Riemannian geometry*, Uspehi Mat. Nauk **30**(1) (1975), 3–59.

[16] Robert Neel and Daniel Stroock, *Analysis of the cut locus via the heat kernel*, Surveys in differential geometry, **IX**, Surveys in Differential Geometry, Int. Press, Somerville, MA, 2004, 337–349.

[17] Daniel W. Stroock, *A concise introduction to the theory of integration*, 3rd ed., Birkhäuser Boston Inc., Boston, MA, 1999.

[18] Daniel W. Stroock and James Turetsky, *Short time behavior of logarithmic derivatives of the heat kernel*, Asian J. Math. **1**(1) (1997), 17–33.

[19] Daniel W. Stroock and James Turetsky, *Upper bounds on derivatives of the logarithm of the heat kernel*, Comm. Anal. Geom. **6**(4) (1998), 669–685.

[20] H.J. Sussmann, *Real analytic desingularization and subanalytic sets: an elementary approach*, Trans. Amer. Math. Soc. **317**(2) (1990), 417–461.

[21] A.N. Varčenko, *Newton polyhedra and estimates of oscillatory integrals*, Funktsional. Anal. i Priložhen. **10**(3) (1976), 13–38.

[22] B.A. Vasil′ev, *The asymptotic behavior of exponential integrals, the Newton diagram and the classification of minima*, Funktsional. Anal. i Priložhen. **11**(3) (1977), 1–11, 96.

[23] Frank W. Warner, *The conjugate locus of a Riemannian manifold*, Amer. J. Math. **87** (1965), 575–604.

[24] Frank W. Warner, *Conjugate loci of constant order*, Ann. of Math. (2) **86** (1967), 192–212.

Department of Mathematics
Columbia University
New York, NY
USA
E-mail address: neel@math.columbia.edu

Received June 21, 2006