Unspecified Book Proceedings Series

Analysis of the Cut Locus via the Heat Kernel

Robert Neel and Daniel Stroock

Abstract. We study the Hessian of the logarithm of the heat kernel to see what it says about the cut locus of a point. In particular, we show that the cut locus is the set of points at which this Hessian diverges faster than t^{-1} as $t \searrow 0$. In addition, we relate the rate of divergence to the conjugacy and other structural properties.

1. Introduction

Our purpose here is to present some recent research connecting behavior of heat kernel to properties of the cut locus. The unpublished results mentioned below constitute part of the first author's thesis, where they will proved in detail.

To explain what sort of results we have in mind, let M be a compact, connected Riemannian manifold of dimension n, and use $p_t(x,y)$ to denote the heat kernel for the heat equation $\partial_t u = \frac{1}{2}\Delta u$. As a special case of a well-known result due to Varadhan,

$$E_t(x,y) \equiv -t \log p_t(x,y) \to E(x,y) \quad \text{uniformly on } M \times M \text{ as } t \searrow 0,$$

where E is the energy function, given in terms of the Riemannian distance by $E(x,y) = \frac{1}{2}\operatorname{dist}(x,y)^2$. Thus, $(x,y) \rightsquigarrow E_t(x,y)$ can be considered to be a geometrically natural mollification of $(x,y) \rightsquigarrow E(x,y)$. In particular, given a fixed $x \in M$, we can hope to learn something about the cut locus $\operatorname{Cut}(x)$ of x by examining derivatives of $y \rightsquigarrow E_t(x,y)$.

Before going further, we present an example which, although somewhat trivial, may help explain what we have in mind. Namely, take $M = \mathbb{S}^1 \equiv \mathbb{R}/2\pi\mathbb{Z}$. In this case, the heat kernel is a theta function

$$p_t(0,\theta) = \frac{1}{\sqrt{2\pi t}} \sum_{n=-\infty}^{\infty} \exp\left[-\frac{(\theta + 2\pi n)^2}{2t}\right],$$

which is obtained by "wrapping" the heat kernel for $\mathbb{R}$ (i.e., the centered Gaussian kernel with variance t) around $\mathbb{S}^1$. It is clear that, as $t \searrow 0$, for any $m \geq 0$ and

The first author was supported by an NSF Graduate Research Fellowship. The second author thanks the NSF for funds provided in DMS-0244991.

$\theta \in (-\pi, \pi)$:

$$\begin{aligned}\partial_\theta^m E_t &\simeq -\partial_\theta^m t \log\left(e^{-\frac{(\theta+\pi)}{2t}} + e^{-\frac{(\theta-\pi)}{2t}}\right)\\ &= \partial_\theta^m t\left(\frac{\theta^2}{2t} - \log\cosh\frac{\pi\theta}{t}\right),\end{aligned}$$

from which it follows that

$$\lim_{t\searrow 0} \partial_\theta^m E_t(0,\theta) = \partial_\theta^m \frac{(\pi - |\theta|)^2}{2} \quad \text{for } \theta \in (-\pi, \pi).$$

On the other hand,

$$\begin{aligned}\lim_{t\searrow 0} \partial_\theta E_t(0,\theta)|_{\theta=\pi} &= 0,\\ \lim_{t\searrow 0} \partial_\theta^2 E_t(0,\theta)\big|_{\theta=\pi} &= -\infty,\end{aligned}$$

but

$$-\lim_{t\searrow 0} t\partial_\theta^2 E_t(0,\theta)\big|_{\theta=\pi} = \pi^2,$$

and things get worse when $m > 2$.

The lessons to be learned from this example are

- The the behavior as $t \searrow 0$ of derivatives of E_t undergo dramatic change at the cut locus. For $\mathbb{S}^1$, this change is already evident in discontinuity which the first derivative has there, and it becomes even more dramatic in the second derivative, which goes from 1 when $\theta \neq 0$ to $-\infty$ when $\theta = 0$.
- The rate at which the Hessian of $E_t(x, y)$ explodes when $y \in \mathrm{Cut}(x)$ can be as high as t^{-1}.
- The direction in which the Hessian explodes is toward $-\infty$. In the case of $\mathbb{S}^1$, the intuitive explanation for this is easy: there are no strictly convex functions on a compact manifold and the Hessian of $E(0, \cdot)$ is 1 except at π. Hence, for $\mathbb{S}^1$, all the "concavity" of $E(0, \cdot)$ must live at π.

That the behavior of E_t for $\mathbb{S}^1$ is somewhat typical was proved in [**3**]. Namely, the following result was proved there.

THEOREM 1. [1] *Given M and some fixed x as above,*

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

uniformly for y in compact subsets of $M\setminus \mathrm{Cut}(x)$. In addition, when $y \in \mathrm{Cut}(x)$ and there are multiple minimal geodesics from x to y, then, if the initial velocities of these geodesics have a sufficiently nice structure (e.g., if they form a submanifold of the tangent space to x), then there is a (strictly) positive definite, symmetric 2-tensor A to which $t\nabla^2 E_t(x, y)$ converges as $t \searrow 0$.

Some aspects of Theorem 1 are quite general. For example, if we think about the Hessian $\nabla^2 E(x, \cdot)$ of E as a distribution (in the sense of Laurent Schwartz), then it is relatively easy to show that its singular support is contained in $\mathrm{Cut}(x)$ and that, as a distribution, $\nabla^2 E(x, \cdot)$ can (cf. [**6**]) be estimated from above in terms of the uniform lower bound on the sectional curvature. Alternatively, the key ingredient in the proof of Theorem 1 is the use of pathspace integration to express

[1] An extension of this result to higher derivatives was given in [**7**], where it was shown that, at the cut locus, the mth order of E_t may explode as fast as $t^{-\frac{m}{2}}$.

$\nabla^2 E_t(x,y)$ as the sum of two terms, one of which stays bounded as $t \searrow 0$ and the other of which is $-t^{-1}$ times the variance of a random variable. Because when $y \notin \mathrm{Cut}(x)$ there is only one minimal geodesic and the first variation around this unique minimal geodesic is non-degenerate, an application of Laplace asymptotics to the pathspace integral allows one to show that the distribution of this random variable degenerates fast enough as $t \searrow 0$ to kill off the variance term. On the other hand, when $y \in \mathrm{Cut}(x)$ for the reason that there are multiple minimal geodesics, there can be residual variance, and this is what accounts for the final statement in Theorem 1.[2]

Unfortunately, the method employed in [**3**] is too cumbersome to encourage any attempt to extend it to more delicate situations or obtain more detailed information. For this reason, it seems wise to attempt seek alternative approaches. Perhaps the most geometrically natural alternative would be to see if one can mimic the computation in the preceding example writing M as the quotient of $\mathbb{R}^n$ by some sufficiently nice group of transformations. For example, it is not hard to analyze the flat torus in the same way as we did circle. More generally, one might hope to get something out of writing M as the quotient of its universal cover by the group of deck transformations. In particular, if M has non-positive sectional curvature, then the Cartan–Hadamard Theorem says that its universal cover will have no cut locus, and so everything should come down to an analysis of the way the deck transformations act on the heat kernel on the universal cover. However, except in very special case, like $\mathbb{S}^1$, such an analysis appears to be very difficult. Even worse, when M can have positive curvature, the structure of geodesics on the universal cover may not be essentially simpler than it is on M itself.

2. Another Approach

For the reasons alluded to toward the end of the preceding section, we will now discuss another, and enormously simpler, way to think about the sort of analysis on which Theorem 1 rests. Namely, it has been realized for some time (cf., for example, Pinsky [**4**]) that more precise asymptotics for $p_t(x,y)$ with $y \in \mathrm{Cut}(x)$ can be obtained by the following method, which we sketch here. By the Chapman–Kolmogorov equation, we can write $p_t(x,y)$ as the integral of $p_{\frac{t}{2}}(x,\cdot)p_{\frac{t}{2}}(\cdot,y)$ over M. Loosely speaking, as $t \searrow 0$,

$$p_{\frac{t}{2}}(x,z)p_{\frac{t}{2}}(z,y) \simeq \frac{1}{\sqrt{V(x,\frac{t}{2})V(y,\frac{t}{2})}} \exp\left(-\frac{h_{x,y}(z)}{2t}\right),$$

where $h_{x,y}(z) \equiv E(x,z) + E(z,y)$ is the *hinged energy function* and $V(p,r)$ denotes the volume of the ball of radius r centered at p. In particular, a naive Laplace asymptotic argument indicates the integral should be getting more and more concentrated on the set Γ of z's where $h_{x,y}(z)$ achieves its minimum value. Equivalently, $\Gamma = \{z : h_{x,y}(z) = E(x,y)\}$ and, as such, is the set of mid-points of minimal geodesics running from x to y. Thus, because Γ is a uniformly positive distance from both $\mathrm{Cut}(x)$ and $\mathrm{Cut}(y)$, one can apply the Pleijel expansion to each of the factors $p_{\frac{t}{2}}(x,\cdot)$ and $p_{\frac{t}{2}}(\cdot,y)$. The result is an expression for the asymptotics

[2] In the extension in [**7**], the coefficient of $t^{-\frac{m}{2}}$ can be interpreted as the mth cumulant of the random variable whose variance appears in this discussion.

of $p_t(x,y)$ in terms of a Laplace integral of the asymptotics of the heat kernel near Γ, which is valid whether or not $y \in \mathrm{Cut}(x)$.

Using more recent results on the heat kernel, including Theorem 1 and the estimates in [**8**], it is possible to employ this method to study logarithmic derivatives of the heat kernel. In order to give a precise statement of what the method says when applied to the Hessian of $E_t(x,\cdot)$ at $\mathrm{Cut}(x)$, we need to introduce a little notation. We have already introduced Γ, the set of midpoints of minimal geodesics from x to y, and the hinged energy function $h_{x,y}(z) = E(x,z) + E(z,y)$. As we said, Γ is precisely the place where $z \rightsquigarrow h_{x,y}(z)$ achieves its minimum value $E(x,y)$ Now let Γ_ϵ be an ϵ-neighborhood of Γ, where we implicitly think of $\epsilon > 0$ as being strictly smaller than $\frac{1}{2}\,\mathrm{dist}(x,y)$. Further, given $z \in M \setminus \mathrm{Cut}(x)$, there is a unique $Z \in T_xM$ such that $s \in [0,1] \longmapsto \exp_x(sZ)$ is the minimal geodesic from x to z, and we will use $H(x,z)$ to denote the Jacobian of $\exp_x$ at Z. For fixed x, $z \in M \setminus \mathrm{Cut}(x) \longmapsto H(x,z) \in \mathbb{R}$ is a smooth function on $M \setminus \mathrm{Cut}(x)$.

THEOREM 2. *Let M be a smooth, compact, connected Riemannian manifold. Choose any two distinct points x and y on M any $A \in T_yM$. Then there exists a positive constant ϵ such that Γ_ϵ is a strictly positive distance from $\{x,y\} \cup \mathrm{Cut}(x) \cup \mathrm{Cut}(y)$ and*

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

where $\nabla_A E(z,y)$ stands for differentiation in the second variable, evaluated at y.

In many ways, the formula in Theorem 2 is an exact replica of the formula on which Theorem 1 was based. Indeed, here, like there, the coefficient of $-t^{-1}$ is a variance. In addition, as was the case there, all the integrals in this formula lend themselves to analysis via Laplace asymptotics as $t \searrow 0$. The difference is that here Laplace asymptotics is for finite dimensional integrals, whereas there it was for integrals in pathspace. Thus, everything should be simpler here. On the other hand, even though we are now working in finite dimensions, the asymptotics can be far from trivial. Indeed, the set Γ onto which the integral is being forced to collapse can be very complicated and ugly!

3. Preliminary Conclusions

We begin our discussion of Theorem 2 by making it explicit that the coefficient of $-t^{-1}$ is a variance. For this purpose, set

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

Clearly, the coefficient of $-t^{-1}$ is the variance $\mathrm{Var}^{\mu_t}(\nabla_A E(\cdot,y))$ of $\nabla_A E(\cdot,y)$ with respect to μ_t. Moreover, because Γ_ϵ is compact, we know that $\{\mu_t : t > 0\}$ is relatively compact in the weak topology, and it is clear that the set L_0 of limit

points as $t \searrow 0$ consists of probability measures which are supported on Γ. In particular, if $\mu \in L_0$ comes from $t_i \searrow 0$, then we have that

$$\lim_{i\to\infty} t_i \nabla^2_{A,A} E_{t_i}(x,y) = -4 \operatorname{Var}^{\mu} \left[\nabla_A E(\cdot, y)\right].$$

In order to get a more explicit expression for $\nabla_A E(z,y)$, let $z \in \Gamma$ be given and take $Y(z)$ be the (unique) unit vector at y such that $\exp_y \left[\operatorname{dist}(z,y) Y(z)\right] = z$. Then we know that

$$\nabla_A E(z,y) = -\frac{1}{2} \operatorname{dist}(x,y) \left\langle A, Y(z) \right\rangle.$$

Finally, use $\theta_A(z)$ to denote the angle between A and $Y(z)$. Then, for any $A \in T_y M$, we have

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

Remark: We digress here in order to expand on the connections between the approach which we are taking here and the one taken in [**3**]. In [**3**], the integrals were taken with respect to Brownian paths on M which start at x and are conditioned to arrive at y at time 1. Using the intuition which comes from the Feynman picture (cf. [**5**]) of Brownian integrals as being Gaussian integrals in which the weight is given by[3]

$$\exp\left(-\frac{1}{2}\int_0^1 |\dot{w}(t)|^2 \, dt\right),$$

the heuristic expression for the heat kernel is

$$p_t(x,y) = \frac{1}{Z(t)} \int\limits_{w(0)=x \;\&\; w(1)=y} \exp\left(-\frac{1}{2t}\int_0^1 |\dot{w}(t)|^2 \, dt\right) dw,$$

where the "dw" is supposed to indicate that the integral is taken with respect to the (non-existent) Lebesgue measure on pathspace and the constant out in front is a (equally non-existent) normalizing factor. Fanciful as this expression may be, it strongly indicates that, as $t \searrow 0$, the overwhelming contribution to the integral will come from those paths w for whose energy is nearly minimal, and, in the limit, one should expect that the integral will be over minimal geodesics. Of course, this is exactly what (2) says. Namely, because Γ parameterizes the minimal geodesics from x to y, the measure μ can be thought of as a probability measure on the space of these minimal geodesics and the function $\cos\theta_A$ can be thought of a function there.

Some simple facts about the log Hessian follow immediately from equation (2). In the first place, if we homothetically scale M by a factor of $a > 0$,

$$\lim_{i\to\infty} t_i \nabla^2_{A,A} E_{t_i}(x,y)$$

is multiplied by a^2. Secondly, we have the inequality

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

Before looking more closely at $h_{x,y}$ and its accompanying Laplace asymptotics, we take a moment to compute a specific example. Earlier, we observed that our explicit computation for $\mathbb{S}^1$ would not extend to higher dimensional spheres. Using

[3]The use of w to denote a generic path is in honor of Wiener.

Theorem 2, however, we can easily compute the leading term of the log Hessian of the heat kernel on the spheres. Choose any point $x \in \mathbb{S}^n$ (here n can be any integer greater than or equal to 2). Then $\mathrm{Cut}(x)$ consists of a single point, namely, the antipodal point to x. Thus, we may as well let x and y be the north and south pole, respectively, which we will denote N and S. In this case, Γ is the equatorial sphere $\mathbb{S}^{n-1}$. Further, by symmetry we see that μ_t converges to uniform probability measure on the equatorial sphere (with respect to the induced volume measure). Next, let A be any unit vector in $T_y\mathbb{S}^n$ (it doesn't matter which one, again by symmetry). The equatorial sphere decomposes nicely into level sets of $\theta_A(z)$. In particular, the level set for a given θ is $\mathbb{S}^{n-2}(\sin\theta)$.[4]

Given the preceding, we can compute the relevant integrals. If we let ω_m denote the volume of the unit sphere of dimension m, we then have that

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

and

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

The expectation-squared term vanishes because $\cos\theta$ is anti-symmetric about $\pi/2$ while $\sin\theta$ is symmetric, and thus the relevant integral vanishes. Plugging this in gives

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

where this last quotient of integrals can be evaluated using integration by parts. Thus, for any $n \geq 1$, we have now shown that $\nabla^2_{A,A}E_t(N,S) \sim -\frac{\pi^2}{nt}|A|^2$ as $t \searrow 0$ for any $A \in T_SM$.

4. Degenerate Minima and Conjugacy

In general it will not be so easy to determine the limiting measure, or measures, $\mu \in L_0$. Even when the set Γ is rather simple (e.g., a finite set of points), one needs information about the nature of the minima which $h_{x,y}$ has on Γ in order to understand L_0. That is, are some or all of the minima degenerate and, if they are, how degenerate are they?

Before getting into a discussion of how degeneracy manifests itself in the asymptotics of $\nabla^2 E_t$, we take a moment to give, in terms of more familiar geometric quantities, an interpretation of what it means for $h_{x,y}$ to have a degenerate or non-degenerate minimum at a point $z \in \Gamma$. Namely, we want to show that the degeneracy of $h_{x,y}$ at $z \in \Gamma$ gives precise information about the conjugacy of the minimal geodesic from x to y which runs through z. That something of this sort ought to be true is clear. To wit, the most extreme degeneracy of $h_{x,y}$ occurs when z is one of a whole submanifold $M' \subseteq \Gamma$ having dimension $n' \geq 1$, as will be the

[4]By $\mathbb{S}^n(a)$ we mean the standard n-dimensional sphere scaled by a factor of a, that is, $\mathbb{S}^n(a)$ is the set of points a distance a from the origin in $\mathbb{R}^{n+1}$.

case when $M = \mathbb{S}^n$ for some $n \geq 2$. Because, in this case, the exponential map will be constant as one moves away from z in any direction $A \in T_z M'$, the geodesic through z will certainly be conjugate. A less extreme case occurs when z is an isolated point of Γ (equivalently, an isolated minimum of $h_{x,y}$). If we think about how $h_{x,y}$ behaves as one moves away from z in some direction, then high order vanishing of $h_{x,y}$ in that direction should indicate the presence of nearby points which are "almost" the midpoints of minimal geodesics from x to y. In other words, we should expect that in this case the minimal geodesic through z is conjugate, although now the conjugacy will usually be a consequence of finite order degeneracy of the exponential map.

To make the preceding precise, given a smooth, real-valued function f which is defined in a neighborhood of the origin in $\mathbb{R}^N$, we will say *f is constant to exactly order m at the origin in the direction $\xi \in \mathbb{S}^{N-1}$ if $(\partial_t)^i f(t\xi)\big|_{t=0}$ is zero for $1 \leq i < m$ but is non-zero for $i = m$*. Now, let γ be a minimal geodesic connecting a point x and y in M, and take $(r, \theta_1, \ldots, \theta_{n-1})$ to be the polar coordinate system on $T_x M$ such that $\gamma(r) = \exp_x(r, 0, \ldots, 0)$ for $r \in [0, \mathrm{dist}(x, y)]$. We then say that γ is *conjugate to exactly order m in the direction $\xi \in \mathbb{S}^{n-2}$ if $\theta \rightsquigarrow \exp_x(r, \theta)$ is constant to exactly order m in the direction ξ*. Notice that this terminology has the annoying feature that geodesics which are conjugate of order 1 are *not* conjugate in the usual sense!

We can now make a precise statement about the relationship between the degeneracy of $h_{x,y}$ and conjugacy of geodesics.

LEMMA 3. *Choose distinct points x and y on M. Let $(r, \theta_1, \ldots, \theta_{n-1})$ and γ be as above. Then $h_{x,y}$ vanishes to exactly order $2m$ at $(\mathrm{dist}(x, y)/2, 0, \ldots, 0)$ in the direction ξ if and only if γ is conjugate to exactly order $2m - 1$ in that direction.*

Thus if $z \in \Gamma$, then z is a non-degenerate minimum of $h_{x,y}$ (i.e., $h_{x,y}$ vanishes to exact order 2 in all directions) if and only if x and y are not conjugate along the minimal geodesic γ passing through z. On the other hand, if z is a degenerate minimum, then x and y are conjugate, and furthermore, the index and orders of conjugacy can be determined from information about which partial derivatives of $h_{x,y}$ are zero.

5. More Refined Laplace Asymptotics when Γ is Discrete

Having related the degeneracy properties $h_{x,y}$ to geodesic geometry, we now return to the problem of understanding the set L_0 of limits, as $t \searrow 0$, of (cf. equation (1)) $\{\mu_t : t > 0\}$, and we begin by considering the case when Γ consists of finitely many points, say $z_1, \ldots, z_m$. Obviously, by taking ϵ small, we can write the integrals with respect to the μ_t's as a sum of integrals over neighborhoods of the individual z_i's. Thus, we can study the asymptotics around each z_i separately.

In order to understand what is happening to μ_t near z_i as $t \searrow 0$, we must understand the structure of the Laplace asymptotics of integrals of the form

$$e^{-2h_{x,y}(z_i)/t} \int_{B_\epsilon(z_i)} \varphi(z) e^{-g(z)/t}\, dz \tag{3}$$

as $t \searrow 0$, where $g(z) \equiv 2h_{x,y}(z) - 2h_{x,y}(z_i)$ and φ is a smooth function. By assumption, g is non-negative and has a unique zero at z_i. Laplace determined the first term of the asymptotic expansion of this integral in the case when the region of integration is one-dimensional and where g has a non-degenerate minimum (that

is, $g''(z_i) > 0$). In order to see what happens in n-dimensions, we first suppose that g can be diagonalized, in the sense that we can find coordinates $(u_1, \dots, u_n)$ around z_i so that

$$g(u_1, \dots, u_n) = \sum_{j=1}^{n} u_k^{2k_j} \tag{4}$$

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 with $k_j = 1$ for each j. However, as we will discuss further below, in general diagonalizability represents a serious problem. Be that as it may, when g can be diagonalized at z_i, results of Estrada and Kanwal [**2**] allow us to give a complete expansion of (3). For the present, we will content ourselves with the first term. Namely,

$$\int_{B_\epsilon(z_i)} \varphi(z) e^{-g(z)/t}\, dz = t^{1/2k_1 + \cdots + 1/2k_n} \left[c \operatorname{vol}_u(z_i) \varphi(z_i) + O\left(t^{1/k_n}\right) \right] \tag{5}$$

where vol_u is the volume element in the u coordinate chart and c is a constant which depends only on n and the k_j's.

From (5), we see that a geodesic which is conjugate in many directions and/or to high order contributes more to the integral over Γ_ϵ than a "less conjugate" geodesic. In particular, suppose that g can be diagonalized around each of its minima and that z_i has associated to it the order of its leading term, $l_i = 1/2k_{1,i} + \cdots + 1/2k_{n,i}$. Then we see that, as $t \searrow 0$, μ_t converges to a limit μ which is supported on those z_i with $l_i = \min\{l_1, \dots, l_m\}$ and furthermore, the density at these points is given by the coefficient of the leading term of the expansion coming from (5), normalized to have total mass one. Therefore, not only does the limit exist, but we know we know what it is.

In more general situations, g may not be diagonalizable around some of the z_i. Nonetheless, in a series of papers (for example [**9**]) which culminate in the monograph [**1**], Arnold and his school have provided a fairly complete analysis of the asymptotic expansion of equation (3) . We give a very brief summary of the most relevant results. First, we need[5] to assume that g vanishes to finite order at z_1. This presents no problem if we work in the analytic category, but in the smooth category it need not be the case. Given that g vanishes to finite order at z_i, there exists a resolution of singularities (in the sense of Hironaka) from which one can determine the behavior of the leading term of the expansion around z_i. In particular, there exists a positive rational number α_i, an integer $p_i \in \{0, \dots, n-1\}$, and a positive real number c_i such that

$$\int_{B_\epsilon(z_i)} \varphi(z) e^{-g(z)/t}\, dz = c_i \varphi(z_i) t^{\alpha_i} |\log t|^{p_i} + O\left(t^{\alpha_i} |\log t|^{p_i - 1}\right).$$

Further, for generic g, α_i and p_i can be determined simply by looking at the Newton polytope, which is a finite combinatorial object depending only on finitely many terms in the Taylor expansion of g, associated to g. Thus, at least in principle, if we assume that g vanishes to finite order at each of the z_i, we can associate to each z_i its leading term, and μ_t will converge to a limit μ supported on those z_i with dominant leading term.

[5]In fact, we know of no general results for the infinite order of vanishing case.

6. Laplace Asymptotics when Γ is not Discrete

So far we've restricted our attention to cases in which Γ is composed of a finite number of points. Obviously, this will not always be true. In general, Γ can be quite complex, and we are very far from a result which covers all possibilities. Nonetheless, we will now discuss one fairly broad situation. Namely, suppose that Γ can be decomposed as 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 accordingly into integrals over the ϵ-neighborhood $(N_i)_\epsilon$ of the individual N_i. Further, by Fubini's Theorem, the integral over $(N_i)_\epsilon$ can be written as an iterated integral in which the horizontal and normal directions to N_i are segregated. But this means that, for each $z \in N_i$, we can, when it applies, use the analysis just discussed. In particular, if we assume that, for each N_i, the asymptotics of the integral in the normal direction has the same order l_i at all points, then, just as before, μ_t will converge to a measure μ which is supported on the union of those N_i's for which l_i is minimal. In fact, on each such N_i, μ will be absolutely continuous with respect to the induced Riemann measure on N_i.

7. A Cautionary Example and a Positive Result

In the preceding two sections, we dealt with relatively nice situations in which Γ was given by isolated smooth submanifolds. Here we construct an example which shows that this need not be the case.

Our example involves $\mathbb{S}^2$ with a metric somewhat deformed from the standard one. To be precise, start with the standard $\mathbb{S}^2$, the one which is embedded in $\mathbb{R}^3$ as the set of points of distance one from the origin, and let x be the north pole and y the south pole. Next, parameterize the equator by $\theta \in [-\pi, \pi)$. With the aid of C^∞ bump functions, we can increase the radius in a neighborhood of some sections of the equator so that, after this deformation, $\Gamma = \{0\} \cup \{2^{-m}\pi : m \geq 1\}$. Further we can achieve this in such a way that $h_{x,y}$ has a non-degenerate minimum at $2^{-m}\pi$ for each $m \geq 1$. On the other hand, it is clear that $h_{x,y}$ will have to vanish to infinite order at 0.

Obviously, we are well outside the situations considered here-to-fore, and the methods developed above do not apply. Nonetheless, we can argue as follows. Pick any point $\theta = \pi/2^{-m}$ for some $m \geq 1$, and choose an open set $U \cap \Gamma = \{\theta\}$. When we apply Laplace asymptotics to the integral over U, we see that

$$\int_U e^{-\frac{h_{x,y}(z)}{t}}\, dz \sim t^{-1} e^{-t^{-1}d} c_1 \operatorname{vol}_u(\theta),$$

where d is half the distance from x to y and c_1 is the constant which appeared earlier. After applying this line of reasoning to each $2^{-m}\pi$, we conclude that

$$\limsup_{t \searrow 0} \mu_t(U) \leq \frac{\det\big(\nabla^2 \mathrm{h_{x,y}}(2^{-\mathrm{m}}\pi)\big)}{\sum_{\ell=1}^{\infty} \det\big(\nabla^2 \mathrm{h_{x,y}}(2^{-\ell}\pi)\big)} = 0,$$

since the denominator is infinite. But this means that, as $t \searrow 0$, μ_t converges to the unit point mass at the point on the equator corresponding to $\theta = 0$.

Here, even though Γ had an accumulation point at which $h_{x,y}$ vanished to infinite order, we were still able to determine the limiting behavior of μ_t. One could easily imagine extending the above construction to produce more pathological

examples where determining the limiting behavior of μ_t would be quite difficult, if it could be done at all.

In addition, this example demonstrates that multiplicity of minimal geodesics is no guarantee that μ is not a point mass. In terms of the asymptotics of $\nabla^2 E_t(x,y)$, L_0 contains some μ other than a point mass precisely when

$$\limsup_{t\searrow 0} t\nabla^2_{A,A} E_t(x,y) < 0 \quad \text{for some } A \in T_yM.$$

Thus, it is clear that the $-t^{-1}$ term in the asymptotics of $\nabla^2_{A,A}E_t(x,y)$ is not sufficient to determine when $y \in \mathrm{Cut}(x)$. Nonetheless we have the following positive result about the relationship between $\mathrm{Cut}(x)$ and the asymptotics of $\nabla^2 E_t(x,y)$.

THEOREM 4. *With the same notation as before, 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

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

where $\|\cdot\|_{\mathrm{op}}$ is the operator norm.

In general, the qualitative result of Theorem 4 is the most we can say about the leading order of $\nabla^2 E_t$ for $y \in \mathrm{Cut}(x)$. However, in the special case where Γ contains only one point, say z_1, around which g is diagonalizable (in particular, we assume equation (4) holds), we can give more detail. Using the further terms in the expansion (5), one can show that, in this case, $\nabla^2_{A,A}E_t(x,y) \sim -Q(A)t^{1/k_n - 1}$ where $Q(A)$ is a symmetric, non-negative definite quadratic form on T_yM. Further, let $Q^\perp$ be the restriction of Q to the $n-1$ dimensional subspace perpendicular to γ, where γ is the unique minimal geodesic between x and y, and let d be the dimension of the kernel of $Q^\perp$. Then the number of i for which $k_i = k_n$ is given by $n-1-d$. In other words, the leading order tells us the highest order of conjugacy of γ, and knowing the leading coefficient as a function of the vector A tells us the number of (independent) directions in which this maximum order of conjugacy is achieved. First, this case shows that every rational of the form $-(m-1)/m$ for a positive integer m can be achieved as the leading order of the expansion of $\nabla^2 E_t$. Second, it indicates that coefficients in the expansion other than that of t^{-1} may have geometric significance. Unfortunately, these coefficients are hard to compute in general, and at the moment we know nothing more about them.

8. Mollification of the Energy Function

So far, we've been concerned with understanding the asymptotics of the log Hessian of the heat kernel for fixed points x and y. However, Theorem 2 can also be used to investigate the distributional Hessian of $E(x,y)$, where we think of x as a fixed base point and y as varying over M. A result of Stroock [**6**] implies that, in the compact case with which we are concerned, $\nabla^2 E(x,y)$, thought of as a distribution, is bounded above by a non-negative constant. It follows that the singular part can be at worst a negative measure, which for fixed base point x and smooth vector field A we denote by $\nu_{x,A}$. It is this measure which we will now investigate. From Varadhan's result, we know that $t\log p_t(x,y)$ gives a smooth mollifier of $-E(x,y)$ as $t \searrow 0$. Thus, computing the (distributional) limit of $-t\nabla^2_{A,A}\log p_t(x,y)$ as $t \searrow 0$

gives the distribution $\nabla^2_{A,A}E(x,y)$. With this in mind, we turn our attention to studying this limit.

We know that in a neighborhood of a point y not in $\mathrm{Cut}(x)$, the distribution $\nabla^2_{A,A}E(x,y)$ is just a smooth function, and Theorem 1 tells us that our mollifier converges to this limit pointwise. In particular, the singular support of $\nabla^2_{A,A}E(x,y)$ is contained in $\mathrm{Cut}(x)$. Looking at Theorem 2, we see that any terms not coming from the variance are $O(t)$ and thus don't contribute to the singular part $\nu_{x,A}$. More concretely, let φ be a smooth function with support in an ϵ-neighborhood of $\mathrm{Cut}(x)$. Then we have

$$\langle \varphi, \nu_{x,A}\rangle = -\lim_{\epsilon \searrow 0}\lim_{t\searrow 0}\int_{B_\epsilon(\mathrm{Cut}(x))}\varphi(y)\frac{4}{t}\operatorname{Var}^{\mu_{t,y}}(\nabla_A E(\cdot,y))\,dy$$

where $\mu_{t,y}$ is the measure μ_t defined by equation (1) corresponding to the point y, with a slight modification. Namely, we need to enlarge the set Γ corresponding to a given y to include not only the midpoints of minimal geodesics to y, but also the midpoints of minimal geodesics from x to any point in a small neighborhood of y (this is so that the error term in Theorem 2 is bounded uniformly as y approaches $\mathrm{Cut}(x)$). We know the limit on the right exists precisely because, by the general theory of distributions, it must be equal to the quantity on the left.

Next, we will lift all of our considerations to the tangent space to M at x. First we recall that $\mathrm{Cut}(x)$ has measure zero, and thus we can ignore it in the preceding integral. Next, observe that $M \setminus \mathrm{Cut}(x)$ is contained in a single coordinate patch. Namely, let (r,θ) 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 $M \setminus \mathrm{Cut}(x)$. We have that

$$\begin{aligned}&\langle \varphi, \nu_{x,A}\rangle = \\ &-\lim_{\epsilon\searrow 0}\lim_{t\searrow 0}\int_{\mathbb{S}^{n-1}}\left[\int_{d(\theta)-\epsilon}^{d(\theta)}\varphi(r,\theta)\frac{4}{t}\operatorname{Var}^{\mu_{t,(r,\theta)}}(\nabla_A E(\cdot,(r,\theta)))\operatorname{vol}(r,\theta)\,dr\right]d\theta.\end{aligned} \tag{6}$$

Note that $(d(\theta),\theta) = \partial U$ parameterizes the preimage of $\mathrm{Cut}(x)$ in the tangent space to x (which is commonly called the tangential cut locus). We can thus identify $\mathbb{S}^{n-1}$ with ∂U using polar coordinates, and then identify $\mathbb{S}^{n-1}$ with the set of minimal geodesics from x to points in $\mathrm{Cut}(x)$ (and also, for that matter, with the union of $\Gamma^{x,y}$ over all $y \in \mathrm{Cut}(x)$). It is this identification of $\mathbb{S}^{n-1}$ with the minimal geodesics from x to $\mathrm{Cut}(x)$ which we will have in mind when, for example, stating that some $\theta \in \mathbb{S}^{n-1}$ corresponds to a conjugate geodesic.

If we look at the integral in equation (6), we see that the central issue is the fact that, while $\operatorname{Var}^{\mu_{t,(r,\theta)}}(\nabla_A E(\cdot,(r,\theta)))$ converges pointwise at any given r in the region of integration, this convergence is not uniform as $r \nearrow d(\theta)$. Thus one must be able to estimate the variance as a function of t and r in such a way that one can first integrate with respect to r and only then let $t \searrow 0$. Here we will say only that this can be done, the central observation being that the the contribution to the variance from nearby geodesics is determined by the Jacobi fields along the geodesic corresponding to θ, as is the volume form. Carrying these ideas through allows one to prove the a pair of theorems about the singular part of $\nabla^2_{A,A}E(x,y)$.

First, one can show that, in a sense, lifting our concerns to the tangent space is the right thing to do. A priori, the right-hand side of equation (6) only makes

sense for functions φ on the tangent space which are lifts of smooth functions on M. However, one can show that the relevant limits exist for almost every θ. In fact, we have the following theorem.

THEOREM 5. *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 6 defines a negative measure on ∂U, which is absolutely continuous with respect to the measure $d\theta$ on ∂U 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 $\mathrm{Cut}(x)$, and further, $\nu_{x,A}$ is given by the pushforward of $\rho(\theta)\,d\theta$ under the exponential map.*

On M, there need not be any natural reference measure with which to compare $\nu_{x,A}$. Theorem 5 tells us that, on the tangent plane, there is such a natural reference measure, namely the measure induced by identifying the set of directions around x with $\mathbb{S}^{n-1}$.

This leads us to wonder what we can say about $\rho(\theta)$. We have the following result, which will require us to introduce a little 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 to 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 6. *Let the hypotheses be as in Theorem 5. 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.*

In addition, we can give an explicit expression for ρ on P, although this requires introducing more notation. Let θ be in P, and let $\tilde{\theta}$ be the (one) associated geodesic. Let y be their common endpoint. Also, let z be the midpoint of the geodesic corresponding to θ. We know that $h_{x,y}$ has non-degenerate Hessian at z; let B denote this Hessian. 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 $\tilde{\theta}$. Finally, recall that $H(x,z)$ is the Jacobian of $\exp_x$ at the vector Z corresponding to z. Then

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

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

There are a few things for us to observe in regard to Theorem 6. First of all, conjugate geodesics do not contribute to the singular part of the distribution. From the point of view of characterizing the cut locus, this means that simply looking at the singular part of $\nabla^2 E(x,y)$ is insufficient, in contrast to the pointwise limits of Theorem 4. On the other hand, in terms of understanding $\nabla^2 E(x,y)$, this says that its singular part is not too bad, in some sense. While the cut locus itself can be quite complicated (for example, it may not be triangulable), 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. We should point out, however, that even though there may not by any singular part of $\nabla^2 E(x,y)$ in a neighborhood of a conjugate point, in general one expects $\nabla^2 E(x,y)$ to be unbounded near a conjugate point. Finally, note that, for a given x, $\nabla^2 E(x,y)$ has no singular part if and only if $\mathrm{Cut}(x)$ and the first conjugate locus of x coincide.

References

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

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

[3] Paul Malliavin and Daniel W. Stroock, *Short time behavior of the heat kernel and its logarithmic derivatives*, J. Differential Geom. **44** (1996), no. 3, 550–570.

[4] Mark A. Pinsky, *Stochastic Riemannian geometry*, Probabilistic analysis and related topics, Vol. 1, Academic Press, New York, 1978, pp. 199–236.

[5] Daniel W. Stroock, *Gaussian measures in traditional and not so traditional settings*, Bull. Amer. Math. Soc. (N.S.) **33** (1996), no. 2, 135–155.

[6] ———, *Non-divergence form operators and variations on Yau's explosion criterion*, J. Fourier Anal. Appl. **4** (1998), no. 4-5, 565–574.

[7] Daniel W. Stroock and James Turetsky, *Short time behavior of logarithmic derivatives of the heat kernel*, Asian J. Math. **1** (1997), no. 1, 17–33.

[8] ———, *Upper bounds on derivatives of the logarithm of the heat kernel*, Comm. Anal. Geom. **6** (1998), no. 4, 669–685.

[9] B. A. Vasil′ev, *Asymptotic exponential integrals, Newton's diagram, and the classification of minimal points*, Functional Anal. Appl. **11** (1977), no. 3, 163–172.

Harvard University M.I.T.

E-mail address: neel@fas.harvard.edu & dws@math.mit.edu