Stat260/CS294: Spectral Graph Methods **Lecture 22 - 04/14/2015**

Lecture: Some Statistical Inference Issues (1 of 3)

Lecturer: Michael Mahoney *Scribe: Michael Mahoney*

Warning: these notes are still very rough. They provide more details on what we discussed in class, but there may still be some errors, incomplete/imprecise statements, etc. in them.

22 Overview of some statistical inference issues

So far, most of what we have been doing on spectral methods has focused on various sorts of algorithms—often but not necessarily worst-case algorithms. That is, there has been a bias toward algorithms that are more rather than less well-motivated statistically—but there hasn't been a lot statistical emphasis *per se*. Instead, most of the statistical arguments have been informal and by analogy, e.g., if the data are nice, then one should obtain some sort of smoothness, and Laplacians achieve that in a certain sense; or diffusions on graphs should look like diffusions on low-dimensional spaces or a complete graph; or diffusions are robust analogues of eigenvectors, which we illustrated in several ways; and so on.

Now, we will spend a few classes trying to make this statistical connection a little more precise. As you can imagine, this is a large area, and we will only be able to scratch the surface, but we will try to give an idea of the space, as well as some of the gotchas of naively applying existing statistical or algorithmic methods here—so think of this as pointing to lots of interesting open questions to do statistically-principled large-scale computing, rather than the final word on the topic.

From a statistical perspective, many of the issues that arise are somewhat different than much of what we have been considering.

- Computation is much less important (but perhaps it should be much more so).
- Typically, one has some sort of model (usually explicit, but sometimes implicit, as we saw with the statistical characterization of the implicit regularization of diffusion-based methods), and one wants to compute something that is optimal for that model.
- In this case, one might want to show things like convergence or consistence (basically, that what is being computed on the empirical data converges to the answer that is expected, as the number of data points $n \to \infty$).

For spectral methods, at a high level, there are basically two types of reference states or classes of models that are commonly-used: one is with respect to some sort of very low-dimensional space; and the other is with respect to some sort of random graph model.

- **Low-dimensional spaces.** In this simplest case, this is a line; more generally, this is a low-dimensional linear subspace; and, more generally, this is a low-dimensional manifold.

Informally, one should think of low-dimensional manifolds in this context as basically low-dimensional spaces that are curved a bit; or, relatedly, that the data are low-dimensional, but perhaps not in the original representation. This manifold perspective provides added descriptive flexibility, and it permits one to take advantage of connections between the geometry of continuous spaces and graphs (which in very special cases are a discretization of those continuous places).

- **Random graphs.** In the simples case, this is simply the G_{nm} or G_{np} Erdos-Renyi (ER) random graph. More generally, one is interested in finding clusters, and so one works with the stochastic blockmodel (which can be thought of as a bunch of ER graphs pasted together). Of course, there are many other extensions of basic random graph models, e.g., to include degree variability, latent factors, multiple cluster types, etc.

These two places provide two simple reference states for statistical claims about spectral graph methods; and the types of guarantees one obtains are somewhat different, depending on which of these reference states is assumed. Interestingly, (and, perhaps, not surprisingly) these two places have a direct connection with the two complementary places (line graphs and expanders) that spectral methods implicitly embed the data.

In both of these cases, one looks for theorems of the form: "If the data are drawn from this place and things are extremely nice (e.g., lots of data and not too much noise) then good things happen (e.g., finding the leading vector, recovering hypothesized clusters, etc.) if you run a spectral method. We will cover several examples of this. A real challenge arises when you have realistic noise and sparsity properties in the data, and this is a topic of ongoing research.

As just alluded to, another issue that arises is that one needs to specify not only the hypothesized statistical model (some type of low-dimensional manifold or some type of random graph model here) but also one needs to specify exactly what is the problem one wants to solve. Here are several examples.

- One can ask to recover the objective function value of the objective you write down.
- One can ask to recover the leading nontrivial eigenvector of the data.
- One can ask to converge to the Laplacian of the hypothesized model.
- One can ask to find clusters that are present in the hypothesized model.

The first bullet above is most like what we have been discussing so far. In most cases, however, people want to use the solution to that objective for something else, and the other bullets are examples of that. Typically in these cases one is asking for a lot more than the objective function value, e.g., one wants to recover the "certificate" or actual solution vector achieving the optimum, or some function of it like the clusters that are found by sweeping along it, and so one needs stronger assumptions. Importantly, many of the convergence and statistical issues are quite different, depending on the exact problem being considered.

- **Today**, we will assume that the data points are drawn from a low-dimensional manifold and that from the empirical point cloud of data we construct an empirical graph Laplacian; and we will ask how this empirical Laplacian relates to the Laplacian operator on the manifold.

- **Next time**, we will ask whether spectral clustering is consistent in the sense that it converges to something meaningful and $n \to \infty$, and we will provide sufficient conditions for this (and we will see that the seemingly-minor details of the differences between unnormalized spectral clustering and normalized spectral clustering lead to very different statistical results).
- **On the day after that**, we will consider results for spectral clustering in the stochastic blockmodel, for both vanilla situations as well as for situations in which the data are very sparse.

22.1 Introduction to manifold issues

Manifold-based ML is an area that has received a lot of attention recently, but for what we will discuss today one should think back to the discussion we had of Laplacian Eigenmaps. At root, this method defines a set of features that can then be used for various tasks such as data set parametrization, clustering, classification, etc. Often the features are useful, but sometimes they are not; here are several examples of when the features developed by LE and related methods are often less than useful.

- **Global eigenvectors are localized.** In this case, "slowly-varying" functions (by the usual precise definition) are not so slowly-varying (in a sense that most people would find intuitive).
- **Global eigenvectors are not useful.** This may arise if one is interested in a small local part of the graph and if information of interest is not well-correlated with the leading or with any eigenvector.
- **Data are not meaningfully low-dimensional.** Even if one believes that there is some sort of hypothesized curved low-dimensional space, there may not be a small number of eigenvectors that capture most of this information. (This does *not* necessarily mean that the data are "high rank," since it is possible that the spectrum decays, just very slowly.) This is more common for very sparse and noisy data, which are of course very common.

Note that the locally-biased learning methods we described, e.g., the LocalSpectral procedure, the PPR procedure, etc., was motivated by one common situation when the global methods such as LE and related methods had challenges.

While it may be fine to have a "feature generation machine," most people prefer some sort of theoretical justification that says when a method works in some idealized situation. To that end, many of the methods like LE assume that the data are drawn from some sort of low-dimensional manifold. Today, we will talk about one statistical aspect of that having to do with converging to the manifold.

To start, here is a simple version of the "manifold story" for a classification problem. Consider a 2-class classification problem with classes C_1 and C_2, where the data elements are drawn from some space $\mathcal{X}$, whose elements are to be classified. A statistical or probabilistic model typically includes the following two ingredients: a probability density $p(x)$ on $\mathcal{X}$; and class densities $\{p\left(C_i|x \in \mathcal{X}\right)\}$, for $i \in \{1, 2\}$. Importantly, if there are unlabeled data, then the unlabeled data don't tell us much about the conditional class distributions, as we can't identify classes without labels, but the unlabeled data can help us to improve our estimate of the probability distribution $p(x)$. That is, the unlabeled data tell us about $p(x)$, and the labeled data tell us about $\{p\left(C_i|x \in \mathcal{X}\right)\}$.

If we say that the data come from a low-dimensional manifold $\mathcal{X}$, then a natural geometric object to consider is the Laplace-Beltrami operator on $\mathcal{X}$. In particular, let $\mathcal{M} \subset \mathbb{R}^n$ be an n-dimensional compact manifold isometrically embedded in $\mathbb{R}^k$. (Think of this as an n-dimensional "surface" in $\mathbb{R}^k$.) The Riemannian structure on $\mathcal{M}$ induces a volume form that allows us to integrate functions defined on $\mathcal{M}$. The square-integrable functions form a Hilbert space $\mathcal{L}^2(\mathcal{M})$. Let $C^\infty(\mathcal{M})$ be the space of infinitely-differentiable functions on $\mathcal{M}$. Then, the Laplace-Beltrami operator is a second order differentiable operator $\Delta_{\mathcal{M}} : C^\infty(\mathcal{M}) \to C^\infty(\mathcal{M})$. We will define this in more detail below; for now, just note that if the manifold is $\mathbb{R}^n$, then the Laplace-Beltrami operator is $\Delta = -\frac{\partial^2}{\partial x_1^2}$. There are two important properties of the Laplace-Beltrami operator.

- **It provides a basis for $\mathcal{L}^2(\mathcal{M})$.** In general, Δ is a PSD self-adjoint operator (w.r.t. the $\mathcal{L}^2$ inner product) on twice differentiable functions. In addition, if $\mathcal{M}$ is a *compact* manifold, then Δ has a discrete spectrum, the smallest eigenvalue of Δ equals 0 and the associated eigenfunction is the constant eigenfunction, and the eigenfunctions of Δ provide an orthonormal basis for the Hilbert space $\mathcal{L}^2(\mathcal{M})$. In that case, any function $f \in \mathcal{L}^2(\mathcal{M})$ can be written as $f(x) = \sum_{i=1}^{\infty} a_i e_i(x)$, where e_i are the eigenfunctions of Δ, i.e., where $\Delta e_i = \lambda_i e_i$.

  In this case, then the simplest model for the classification problem is that the class membership is a square-integrable function, call it $m : \mathcal{M} \to \{-1, +1\}$, in which case the classification problem can be interpreted as interpolating a function on the manifold. Then we can choose the coefficients to get an optimal fit, $m(x) = \sum_{i=1}^{n} a_i e_i$, in the same way as we might approximate a signal with a Fourier series. (In fact, if $\mathcal{M}$ is a unit circle, call it S^1, then $\Delta_{S_1} f(\theta) = -\frac{d^2 f(\theta)}{d\theta^2}$, and the eigenfunctions are sinusoids with eigenvalues $\{1^2, 2^2, \ldots\}$. and we get the usual Fourier series.)

- **It provides a smoothness functional.** Recall that a simple measure of the degree of smoothness for a function f on the unit circle S^1 is

$$S(f) = \int_{S^1} |f(\theta)'|^2 d\theta.$$

  In particular, f is smooth iff this is close to zero. If we take this expression and integrate by parts, then we get

$$S(f) = \int_{S^1} f'(\theta) d\theta = \int_{S^1} f \Delta f d\theta = \langle \Delta f, f \rangle_{\mathcal{L}^2(S^1)}.$$

  More generally, if $f : \mathcal{M} \to \mathbb{R}$, then it follows that

$$S(f) = \int_{\mathcal{M}} |\nabla f|^2 d\mu = \int_{\mathcal{M}} f \Delta f d\mu = \langle \Delta f, f \rangle_{\mathcal{L}^2(\mathcal{M})}.$$

  So, in particular, the smoothness of the eigenfunction is controlled by the eigenvalue, i.e.,

$$S(e_i) = \langle \Delta e_i, e_i \rangle_{\mathbb{L}^2(\mathcal{M})} = \lambda_i,$$

  and for arbitrary f that can be expressed as $f = \sum_i \alpha_i e_i$, we have that

$$S(f) = \langle \Delta f, f \rangle = \left\langle \sum_i \alpha_i \Delta e_i, \sum_i \alpha_i e_i \right\rangle = \sum_i \lambda_i \alpha_i^2.$$

(So, in particular, approximating a function f by its first k eigenfunctions is a way to control the smoothness of the eigenfunctions; and the linear subspace where the smoothness functions is finite is a RKHS.)

This has strong connections with a range of RKHS problems. (Recall that a RKHS is a Hilbert space of functions where the evaluation functionals, the functionals that evaluate functions at a point, are bounded linear functionals.) Since the Laplace-Beltrami operator on $\mathcal{M}$ can be used to provide a basis for $\mathcal{L}^2(\mathcal{M})$, we can take various classes of functions that are defined on the manifold and solve problems of the form

$$\min_{f \in H} \sum_i (y_i - f(x_i))^2 + \lambda G(f), \tag{1}$$

where $H : \mathcal{M} \to \mathbb{R}$. In general, the first term is the empirical risk, and the second term is a stabilizer or regularization term. As an example, one could choose $G(f) = \int_{\mathcal{M}} \langle \nabla f, \nabla f \rangle = \sum_i \alpha_i^2 \lambda_i$ (since $f = \sum_i \alpha_i e_i(x)$), and $H = \{f = \sum_i \alpha_i e_i | G(f) < \infty\}$, in which case one gets an optimization problem that is quadratic in the α variables.

As an aside that is relevant to what we discussed last week with the heat kernel, let's go through the construction of a RKHS that is invariantly defined on the manifold $\mathcal{M}$. To do so, let's fix an infinite sequence of non-negative numbers $\{\mu_i | i \in \mathbb{Z}^+\}$ s.t. $\sum_i \mu_i < \infty$ (as we will consider in the examples below). Then, define the following linear space of continuous functions

$$H = \left\{ f = \sum_i \alpha_i f_i | \sum_i \frac{\alpha_i^2}{\mu_i} < \infty \right\}.$$

Then, we can define the inner product as: for $f = \sum_i \alpha_i f_i$ and $g = \sum_i \beta_i g_i$, we have $\langle f, g \rangle = \sum_i \frac{\alpha_i \beta_i}{\mu_i}$. Then, H is a RKHS with the following kernel: $K(p, q) = \sum_i \mu_i e_i(p) e_i(q)$. Then, given this, we can solve regularized optimization problems of the form given in Eqn. (**??**) above. In addition, we can get other choices of kernels by using different choices of μ vectors. For example, if we let $\mu_i = e^{-t\lambda_i}$, where λ_i are the eigenvalues of Δ, then we get the heat kernel corresponding to heat diffusion on the manifold; if we let $\mu_i = 0$, for all $i > i^*$, then we are solving an optimization problem in a finite dimensional space; and so on.

All of this discussion has been for data drawn from an hypothesized manifold $\mathcal{M}$. Since we are interested in a smoothness measure for functions for a graph, then if we think of the graph as a model for the manifold, then we want the value of a function not to change too much between points. In that case, we get

$$S_G(f) = \sum_{i\ j} W_{ij} (f_i - f_j),$$

and it can be shown that

$$S_G(f) = fLf^T = \langle f, Lf \rangle_G = \sum_{i=1}^{n} \lambda_i \langle f, e_i \rangle_G.$$

Of course, this is the discrete object with which we have been working all along. Viewed from the manifold perspective, this corresponds to the discrete analogue of the integration by parts we performed above. In addition, we can use all of this to consider questions having to do with "regularization on manifolds and graphs," as we have allude to in the past.

To make this connection somewhat more precise, recall that for a RKHS, there exists a kernel $K : X \times X \to \mathbb{R}$ such that $f(x) = \langle f(\cdot), K(x, \cdot) \rangle_H$. For us today, the domain X could be a manifold $\mathcal{M}$ (in which case we are interested in kernels $K : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$), or it could be points from $\mathbb{R}^n$ (say, on the nodes of the graph that was constructed from the empirical original data by a nearest neighbor rule, in which case we are interested in kernels $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$). We haven't said anything precise yet about how these two relate, so now let's turn to that and ask about connections between kernels constructed from these two different places, as $n \to \infty$.

22.2 Convergence of Laplacians, setup and background

Now let's look at questions of convergence.

If $H : \mathcal{M} \to R$ is a RKHS invariantly defined on $\mathcal{M}$, then the key goal is to minimize regularized risk functionals of the form

$$E_\lambda = \min_{f \in H} \mathbb{E}\left[(y - f(x))^2\right] + \lambda \|f\|_H^2.$$

In principle, we can do this—if we had an infinite amount of data available *and* the true manifold is known. Instead, we minimize the empirical risk which is of the form

$$\hat{E}_{\lambda,n} = \min_{f \in H} \frac{1}{n} \sum (y_i - f(x_i))^2 + \lambda \|f\|_H.$$

The big question is: how far is $\hat{E}_{\lambda,n}$ from E_λ.

The point here is the following: assuming the manifold is known or can be estimated from the data, then making this connection is a relatively-straightforward application of Hoeffding bounds and regularization/stability ideas. But:

- In theory, establishing convergence to the hypothesized manifold is challenging. We will get to this below.
- In practice, testing the hypothesis that the data are drawn from a manifold in some meaningful sense of the word is harder still. (For some reason, this question is not asked in this area. It's worth thinking about what would be test statistics to validate or invalidate the manifold hypothesis, e.g., is that the best conductance clusters are not well balanced sufficient to invalidate it?)

So, the goal here is to describe conditions under which the point cloud in $\mathcal{X}$ of the sample points converges to the Laplace-Beltrami operator on the underlying hypothesized manifold $\mathcal{M}$. From this perspective, the primary data are points in $\mathcal{X}$, that is assumed to be drawn from an underlying manifold, with uniform or nonuniform density, and we want to make the claim that the Adjacency Matrix or Laplacian Matrix of the empirical data converges to that of the manifold. (That is, the data are not a graph, as will arise with the discussion of the stochastic block model.) In particular, the graph is and empirical object, and if we view spectral graph algorithms as applying to that empirical object then they are stochastically justified when they can relate to the underlying processes generating the data.

What we will describe today is the following.

- For data drawn from a uniform distribution on a manifold $\mathcal{M}$, the graph Laplacian converges to the Laplace-Beltrami operator, as $n \to \infty$ and the kernel bandwidth is chosen appropriately (where the convergence is uniform over points on the manifold and for a class of functions).
- The same argument applies for arbitrary probability distributions, except that one converges to a weighted Laplacian; and in this case the weights can be removed to obtain convergence to the normalized Laplacian. (Reweighting can be done in other ways to converge to other quantities of interest, but we won't discuss that in detail.)

Consider a compact smooth manifold $\mathcal{M}$ isometrically embedded in $\mathbb{R}^n$. The embedding induces a measure corresponding to volume form μ on the manifold (e.g., the volume form for a closed curve, i.e., an embedding of the circle, measures the usual curve length in $\mathbb{R}^n$). The Laplace-Beltrami operator $\Delta_{\mathcal{M}}$ is the key geometric object associated to a Riemannian manifold. Given $\rho \in \mathcal{M}$, the tangent space $T_\rho\mathcal{M}$ can be identified with the affine space to tangent vectors to $\mathcal{M}$ at ρ. (This vector space has a natural inner product induced by embedding $\mathcal{M} \subset \mathbb{R}^n$.) So, given a differentiable function $f : \mathcal{M} \to \mathcal{R}$, let $\nabla_{\mathcal{M}} f$ be the gradient vector on $\mathcal{M}$ (where $\nabla_{\mathcal{M}} f(p)$ points in the direction of fastest ascent of f at ρ. Here is the definition.

Definition 1. *The* Laplace-Beltrami operator $\Delta_{\mathcal{M}}$ *is the divergence of the gradient, i.e.,*

$$\Delta_{\mathcal{M}} f = -div(\nabla_{\mathcal{M}} f)\,.$$

Alternatively, $\Delta_{\mathcal{M}}$ *can be defined as the unique operator s.t., for all two differentiable functions* f *and* h*,*

$$\int_{\mathcal{M}} h(x)\Delta_{\mathcal{M}} f(x) d\mu(x) = \int_{\mathcal{M}} \langle \nabla_{\mathcal{M}} h(x), \nabla_{\mathcal{M}} f(x)\rangle \, d\mu,$$

where the inner product is on the tangent space and μ *is the uniform measure.*

In $\mathbb{R}^n$, we have $\Delta f = -\sum_i \frac{\partial^2 f}{\partial x_i^2}$. More generally, on a k-dimensional manifold $\mathcal{M}$, in a local coordinate system $(x_1, \ldots, x_n)$, with a metric tensor g_{ij}, if g^{ij} are the components of the inverse of the metric tensor, then the Laplace-Beltrami operator applied to a function f is

$$\Delta_{\mathcal{M}} f = \frac{1}{\sqrt{\det(g)}} \sum_j \frac{\partial}{\partial x^j} \left(\sqrt{\det(g)} \sum_i g^{ij} \frac{\partial f}{\partial x_i} \right)$$

(If the manifold has nonuniform measure ν, given by $d\nu(x) = P(x)d\mu(x)$, for some function $P(x)$ and with $d\mu$ being the canonical measure corresponding to the volume form, then we have the more general notion of a *weighted manifold Laplacian*: $\Delta_{\mathcal{M},\mu} = \Delta_P f = \frac{1}{P(x)} \mathrm{div}\,(P(x)\nabla_{\mathcal{M}} f)$.)

The question is how to reconstruct $\Delta_{\mathcal{M}}$, given a finite sample of data points from the manifold? Here are the basic objects (in addition to $\Delta_{\mathcal{M}}$) that are used to answer this question.

- **Empirical Graph Laplacian.** Given a sample of n points $x_i, \ldots, x_n$ from $\mathcal{M}$, we can construct a weighted graph with weights $W_{ij} = e^{-\|x_i - x_j\|^2/4t}$, and then

$$\left(L_n^t\right)_{ij} = \begin{cases} -W_{ij} & \text{if } i \neq j \\ \sum_k W_{ik} & \text{if } i = j \end{cases}.$$

Call L_n^t the graph Laplacian matrix. We can think of L_n^t as an operation of functions on the n empirical data points:

$$L_n^t f(x_i) = f(x_i) \sum_j e^{-\|x_i - x_j\|^2/(4t)} - \sum_j f(x_j) e^{-\|x_i - x_j\|^2/(4t)},$$

but this operator operates only on the empirical data, i.e., it says nothing about other points from $\mathcal{M}$ or the ambient space in which $\mathcal{M}$ is embedded.

- **Point Cloud Laplace operator.** This formulation extends the previous results to any function on the ambient space. Denote this by $\underline{\mathrm{L}}_n^t$ to get

$$\underline{\mathrm{L}}_n^t f(x) = f(x) \frac{1}{n} \sum_j e^{-\|x - x_j\|^2/(4t)} - \frac{1}{n} \sum_j f(x_j) e^{-\|x - x_j\|^2/(4t)}$$

(So, in particular, when evaluated on the empirical data points, we have that $\underline{\mathrm{L}}_n^t f(x_i) = \frac{1}{n} L_n^T f(x_i)$.) Call $\underline{\mathrm{L}}_n^t$ the Laplacian associated to the point cloud $x_1, \ldots, x_n$.

- **Functional approximation to the Laplace-Beltrami operator.** Given a measure ν on $\mathcal{M}$, we can construct an operator

$$\underline{\mathrm{L}}^t f(x) = f(x) \int_{\mathcal{M}} e^{-\|x - y\|^2/(4t)} d\nu(y) - \int_{\mathcal{M}} f(y) e^{-\|x - y\|^2/(4t)} d\nu(y).$$

Observe that $\underline{\mathrm{L}}_n^t$ is just a special form of $\underline{\mathrm{L}}^t$, corresponding to the Dirac measure supported on $x_1, \ldots, x_n$.

22.3 Convergence of Laplacians, main result and discussion

The main result they describe is to establish a connection between the graph Laplacian associated to a point cloud (which is an extension of the graph Laplacian from the empirical data points to the ambient space) and the Laplace-Beltrami operator on the underlying manifold $\mathcal{M}$. Here is the main results.

Theorem 1. *Let $x_1, \ldots, x_n$ be data points sampled from a uniform distribution on the manifold $\mathcal{M} \subset \mathbb{R}^n$. Choose $t_n = n^{-1/(k+2+\alpha)}$, for $\alpha > 0$, and let $f \in C^\infty(\mathcal{M})$. Then*

$$\lim_{n \to \infty} \frac{1}{t_n (4\pi t_n)^{k/2}} \underline{L}_n^{t_n} f(x) = \frac{1}{Vol(\mathcal{M})} \Delta_{\mathcal{M}} f(x),$$

where the limit is taken in probability and $Vol(\mathcal{M})$ is the volume of the manifold with respect to the canonical measure.

We are not going to go through the proof in detail, but we will outline some key ideas used in the proof. Before doing that, here are some things to note.

- This theorem assert pointwise convergence of $\underline{\mathrm{L}}_n^t f(p)$ to $\Delta_{\mathcal{M}} f(p)$, for a fixed function f and a fixed point p.
- Uniformity over all $p \in \mathcal{M}$ follows almost immediately from the compactness of $\mathcal{M}$.

- Uniform convergence over a class of function, e.g., functions $C^k(\mathcal{M})$ with bounded k^{th} derivative, follows with more effort.
- One can consider a more general probability distribution P on $\mathcal{M}$ according to which data points are sampled—we will get back to an example of this below.

For the proof, the easier part is to show that $\underline{\mathrm{L}}_n^t \to \underline{\mathrm{L}}^t$, as $n \to \infty$, if points are samples uniformly: this uses some basic concentration results. The harder part is to connect $\underline{\mathrm{L}}^t$ and $\Delta_{\mathcal{M}}$: what must be shown is that when $t \to 0$, then L^t appropriately scaled converges to $\Delta_{\mathcal{M}}$.

Here are the basic proof ideas, which exploit heavily connections with the heat equation on $\mathcal{M}$.

For simplicity, consider first $\mathbb{R}^n$, where we have the following theorem.

Theorem 2 (Solution to heat equation on $\mathbb{R}^k$)**.** *Let $f(x)$ be a sufficiently differentiable bounded function. Then*

$$H^t f = (4\pi t)^{-k/2} \int_{\mathbb{R}^k} e^{-\frac{\|x-y\|^2}{4t}} f(y) dy,$$

and

$$f(x) = \lim_{t \to 0} H^t f(x) = (4\pi t)^{-k/2} \int_{\mathbb{R}^k} e^{-\frac{\|x-y\|^2}{4t}} f(y) dy,$$

and the function $u(x,t) = H^t f$ satisfies the heat equation

$$\frac{\partial}{\partial t} u(x,t) + \Delta u(x,t) = 0$$

with initial condition $u(x,0) = f(x)$.

This result for the heat equation is the key result for approximating the Laplace operator.

$$\begin{aligned} \Delta f(x) &= -\frac{\partial}{\partial t} u(x,t)|_{t=0} \\ &= -\frac{\partial}{\partial t} H^t f(x)|_{t=0} \\ &= \lim_{t \to 0} \frac{1}{t} \left(f(x) - H^t f(x) \right). \end{aligned}$$

By this last result, we have a scheme for approximating the Laplace operator. To do so, recall that the heat kernel is the Gaussian that integrates to 1, and so

$$\Delta f(x) = \lim_{t \to 0} -\frac{1}{t} \left((4\pi t)^{-k/2} \int_{\mathbb{R}^k} e^{\frac{-\|x-y\|^2}{4t}} f(y) dy - f(x) (4\pi t)^{-k/2} \int_{\mathbb{R}^k} e^{\frac{-\|x-y\|^2}{4t}} dy. \right)$$

It can be shown that this can be approximated by the point cloud $x_1, \ldots, x_n$ by computing the empirical version as

$$\begin{aligned} \hat{\Delta} f(x) &= \frac{1}{t} \frac{(4\pi t)^{-k/2}}{n} \left(f(x) \sum_i e^{\frac{-\|x-x_i\|^2}{4t}} - \sum_i e^{\frac{-\|x-x_i\|^2}{4t}} f(x_i) \right) \\ &= \frac{1}{t (4\pi t)^{k/2}} \underline{\mathrm{L}}_n^t f(x). \end{aligned}$$

It is relatively straightforward to extend this to a convergence result for $\mathbb{R}^k$. To extend it to a convergence result for arbitrary manifolds $\mathcal{M}$, two issues arise:

- With very few exceptions, we don't know the exact form of the heat kernel $H^t_{\mathcal{M}}(x,y)$. (It has the nice form of a Gaussian for $\mathcal{M} = \mathbb{R}^k$.)
- Even asymptotic forms of the heat kernel requires knowing the geodesic distance between points in the point cloud, but we can only observe distance in the ambient space.

See their paper for how they deal with these two issues; this involves methods from differential geometry that are very nice but that are not directly relevant to what we are doing.

Next, what about sampling with respect to nonuniform probability distributions? Using the above proof, we can establish that we converge to a weighted Laplacian. If this is not of interest, then once can instead normalize differently and get one of two results.

- The weighted scaling factors can be removed by using a different normalization of the weights of the point cloud. This different normalization basically amounts to considering the normalized Laplacian. See below.
- With yet a different normalization, we can recover the Laplace-Beltrami operator on the manifold. The significance of this is that it is possible to separate geometric aspects of the manifold from the probability distribution on it. This is of interest to harmonic analysts, and it underlies extension of the Diffusion Maps beyond the Laplacian Eigenmaps.

As for the first point, if we have a compact Riemannian manifold $\mathcal{M}$ and a probability distribution $P : \mathcal{M} \to \mathbb{R}^+$ according to which points are drawn in an i.i.d. fashion. Assume that $a \leq P(x) \leq b$, for all $x \in \mathcal{M}$. Then, define the point cloud Laplacian operator as

$$\underline{\mathrm{L}}^t_n f(x) = \frac{1}{n} \sum_{i=1}^{n} W(x_i, x_j) \left(f(x) - f(x_i) \right)$$

If $W(x, x_i) = e^{\frac{\|x - x_i\|}{4t}}$, then this corresponds to the operator we described above. In order to normalized the weights, let

$$W(x, x_i) = \frac{1}{t} \frac{G_t(x, x_i)}{\sqrt{\hat{d}_t(x)} \sqrt{\hat{d}_t(x_i)}},$$

where

$$\begin{aligned} G_t(x, x_i) &= \frac{1}{(4\pi t)^{k/2}} e^{-\frac{\|x - x_i\|^2}{4t}}, \\ \hat{d}_t(x) &= \frac{1}{n} \sum_{j \neq i} G_t(x, x_j), \quad \text{and} \\ \hat{d}_t(x_i) &= \frac{1}{n-1} \sum_{j \neq i} G_t(x_i, x_j), \end{aligned}$$

where the latter two quantities are empirical estimates of the degree function $d_t(x)$, where

$$d_t(x) = \int_{\mathcal{M}} G_t(x, y) P(y) \mathrm{Vol}(y).$$

Note that we get a degree function—which is a continuous function defined on $\mathcal{M}$. This function bears some resemblance to the diagonal degree matrix of a graph, and it can be thought of as

a multiplication operator, but it has very different properties than an integral operator like the heat kernel. We will see this same function next time, and this will be important for when we get consistency with normalized versus unnormalized spectral clustering.