Ryan Vaughn

Research Statement

Summary

I am interested in applying techniques from Riemannian geometry to machine learning and data-driven methods in applied mathematics. Most of my work to date has been in the area of manifold learning, in which one assumes that finite data $\{x_i\}_{i=1}^N \subseteq \mathbb{R}^d$ lies on some unknown Riemannian manifold (M, g). It is often the case that data collected in $\mathbb{R}^d$ is highly correlated, and thus the "data space" $\mathbb{R}^d$ is much higher dimension than the underlying manifold M. Thus, it can be important to attempt to "learn" such an underlying manifold, most prominently to help avoid the so-called "curse of dimensionality".

So far, I have used techniques from manifold learning to help develop a mesh-free numerical scheme for solving diffusion-type PDEs on finite data sampled from manifolds with boundary. A preprint[37] based on these results is in the process of being submitted for publication and several offshoots of this research are planned for future research.

I also have long-term interests in the study of discrete differential forms adapted to data science. While discrete differential forms are well-studied in the finite element community through Discrete Exterior Calculus[26] and Finite Element Exterior Calculus[3][2], less work has been done when given only a sample of finitely many points. A treatment of differential forms on data would provide a connection between topology, geometry, and machine learning.

Through these projects, I hope to provide a more sound understanding of the relationship between topology, geometry, and data.

Introduction and Background: Diffusion Maps

One of the most successful algorithms in manifold learning is *diffusion maps*[12], developed by Coifman and Lafon. Diffusion maps is a data-driven, graph-Laplacian-based method which is related to spectral clustering [29][32] and Laplacian Eigenmaps[5]. It is an unsupervised learning technique used often for clustering, dimensionality reduction, and data-driven forecasting.

An appealing aspect of the diffusion maps graph Laplacian is that it provably converges to the Laplace-Beltrami operator $\Delta : C^\infty(M) \to C^\infty(M)$ even when the data is not sampled uniformly on M. The Laplace-Beltrami operator is a differential operator on M that encodes much information about the geometry of M. In local coordinates, Δ is defined by:

$$\Delta f = -\sum_{i,j=1}^{\dim(M)} \frac{1}{\sqrt{|g|}} \frac{\partial}{\partial x_i} \left(\sqrt{|g|} g^{ij} \frac{\partial f}{\partial x_j} \right).$$

In the case that M is an open subset of $\mathbb{R}^d$, then the Laplace-Beltrami operator simplifies to the standard Laplace operator in $\mathbb{R}^d$

$$\Delta f(x) = -\sum_{i=1}^{d} \frac{\partial^2 f}{\partial x_i^2}(x).$$

The field of *spectral geometry* is devoted to studying the relationship between geometric properties of M and the eigenfunctions of Δ. In this way, diffusion maps provides a connection between machine learning and spectral geometry. Not only that, but eigenfunctions of Δ are a generalization of a Fourier basis on M, providing a useful tool from harmonic analysis.
In the work of Hein et. al.[25][24], it was shown that the diffusion maps graph Laplacian $L_{\varepsilon,N}$ on N data points and with parameter $\varepsilon > 0$ converges in probability to an averaging operator Δ_ε as $N \to \infty$. Assuming M has empty boundary, the original diffusion maps paper of Coifman and Lafon[12] showed that this averaging operator Δ_ε converges uniformly to the Laplace-Beltrami operator as $\varepsilon \to 0$. In this sense, we can see that when M is a manifold without boundary, the diffusion maps Laplacian converges to the Laplace-Beltrami operator as $N \to \infty$ and $\varepsilon \to 0$.

$$\operatorname*{plim}_{N\to\infty} L_{\varepsilon,N} = \Delta_\varepsilon \text{ and } \lim_{\varepsilon\to 0} \Delta_\varepsilon = \Delta.$$

The convergence of Δ_ε to Δ for manifolds with boundary has proved to be more problematic. It was shown in [12] that certain terms that would normally cancel in the case of empty boundary would instead blow up for points near the boundary. It was, however, observed that the output of diffusion maps estimated the Laplace-Beltrami operator applied to functions with Neumann boundary conditions, though this behavior was never fully theoretically justified.

Current Project: Diffusion Maps for Boundary Value Problems

There has been recent interest in using diffusion maps as a data-driven method for numerically solving elliptic PDEs[18][23]. This approach is potentially appealing, since diffusion maps does not require one to generate a triangulation or mesh of the underlying manifold. However, the required restriction to functions with Neumann boundary conditions is a significant restriction when solving PDEs.
In our forthcoming paper, *Diffusion Maps for Embedded Manifolds with Boundary with Applications to PDEs*[37], we provide a more rigorous treatment of diffusion maps near the boundary. Using the notion of weak solutions to PDEs, we show that the averaging operator Δ_ε converges uniformly to Δ in well-known weak (variational) sense. By using an additional consistent estimator for boundary detection[9], we are able to adapt diffusion maps for use in boundary value problems. This work provides rigorous theoretical justification of convergence for elliptic PDEs related to the Laplace-Beltrami operator with Dirichlet, Neumann, and Robin boundary conditions. In particular, it also explains the previously observed tendency for Diffusion maps to output functions with Neumann boundary conditions. Our work is also shown to be effective experimentally by numerically solving PDEs with Neumann, Dirichlet, and Robin boundary conditions on several data sets.
My contribution to this work was to adapt the original diffusion maps proof using a different set of coordinates called *semigeodesic coordinates*[28]. Semigeodesic coordinates in this case are a set of coordinates well-suited for computations made near the boundary. To show their usefulness, we first look at a crucial lemma from the proof in diffusion maps.
The crux of the argument in the proof of convergence of diffusion maps relies on the following asymptotic expansion:

$$\int_M k\left(\frac{|x-y|^2_{\mathbb{R}^d}}{\varepsilon^2}\right) f(y)dV = m_0 f(x) + \varepsilon^2 \frac{m_2}{2}(\omega(x)f(x) - \Delta f(x)) + \mathcal{O}(\varepsilon^3) \tag{1}$$

where m_0 and m_2 are constants, and $\omega(x)$ is a term depending on the embedding of M. Using a few normalization steps, one can then isolate the Laplace-Beltrami operator on the right-hand side.

This procedure is then used to construct the averaging operator Δ_ε and then to show that such an averaging operator converges uniformly to Δ as $\varepsilon \to 0$.

The proof of this expansion is carried out by first showing that one can localize the integral on the left hand side of (1) to a small ball B in M centered at x. In such a geodesic ball, one can Taylor expand the left-hand term in Riemannian normal coordinates. In Riemannian normal coordinates, the ball B is radially symmetric, and odd-order terms of the Taylor polynomial vanish due to radial symmetry. In this way, a very complicated looking expansion of the left hand side of (1) simplifies to the expression on the right hand side.

When x is a point near the boundary, such a geodesic ball B might intersect the boundary. In this case, the coordinate representation of B can be highly non-symmetric, so problematic terms no longer cancel.

To partially resolve this issue, instead of localizing to a geodesic ball in M, we localize to a semigeodesic coordinate chart, which share many of the useful properties of Riemannian normal coordinates. In this setup, localizing to a semigeodesic coordinate chart is more akin to localizing to a "geodesic hypercylinder" oriented perpendicular to the boundary as opposed to a geodesic ball. Such a hypercylinder is useful because it is still highly-symmetric and yet its intersection with the boundary is well-behaved. The tradeoff for this, however, is that the expansion becomes more complicated, with additional terms related to the curvature of the boundary in M.

By showing that one can always localize to such a chart, and then carrying out an expansion of the left-hand-side of (1), we are able to obtain an analogous, but more complicated expansion as the one in (1). Although most of the order-ε terms in the expansion cancel by symmetry, one term remains which still blows up as x approaches the boundary:

$$\varepsilon m_1^{\partial}(x)\left(\eta_x \cdot \nabla(f)(x) - H(x)f(x)\right)$$

This term includes the mean curvature of the boundary $H(x)$, as well as the inward-facing gradient of f. However, we then show that when viewing the Laplacian in a weak (variational) sense, this error term becomes order ε^2. We use this result to show that the averaging operator Δ_ε converges to Δ uniformly in a weak sense near the boundary.

In a similar manner, semigeodesic coordinates are also used to address the problem of sampling on the boundary. In general, since the boundary is of measure zero, one can never expect for sampled points to be boundary points of M. We show that the integral over a small ε-tubular region of the boundary is an asymptotic approximation of an integral over the boundary. This crucially allows us to estimate boundary integrals on data, which are necessary to apply a variational approach. We then use a consistent estimator developed by Berry and Sauer[9] which estimates a distance from the boundary in order to estimate these boundary integrals in practice. Using a variational approach, we are then able to augment the Diffusion maps graph Laplacian for use in several numerical experiments.

Future Research: Fermi Coordinates

Semigeodesic coordinates played a crucial role in our analysis near the boundary. In general, semigeodesic coordinates are particularly amenable for computations around a codimension-one hypersurface of M. Semigeodesic coordinates can be generalized further to so-called *Fermi coordinates*, which are particularly well-adapted to computation of submanifolds of higher codimension.

In [22], many of the appealing properties which we use with semigeodesic coordinates are proven for Fermi coordinates. The motivation for these results is for defining boundary value problems for dirac operators on $\text{Spin}^{\mathbb{C}}$-manifolds of bounded geometry. While the structure of a $\text{Spin}^{\mathbb{C}}$-manifold does not naturally occur in manifold learning, I am interested in seeing if one could adapt this idea to a simpler case. In particular, the boundary value problems treated in [22] are on principal bundles. The paper of Wu and Singer [36] places a general class of techniques related to Diffusion maps as operators on principal bundles. It seems that perhaps that one could use Fermi coordinates to solve PDEs on these related operators.

Future Research: Local Kernels and Fractional PDEs

The theory of local kernels is also a promising area for future research. The radial symmetry of the kernel $k\left(\frac{|x-y|^2_{\mathbb{R}^d}}{\varepsilon^2}\right)$ in diffusion maps is the key to recovering the Laplace-Beltrami operator. In [8] it was shown that by varying properties of a specific class of kernels called *local kernels,* the diffusion maps algorithm would estimate the Laplace-Beltrami operator on M with respect to varying Riemannian metrics on M. This fact was used recently by [18, 23] to solve a more general class of elliptic PDEs on manifolds without meshes. However, the theory developed in [8] is restricted to manifolds without boundary, which is very restrictive for PDE applications. Having generalized the standard diffusion maps result in [37], a more detailed analysis should allow us to apply semigeodesic coordinates near the boundary for general local kernels. This would allow the estimation of the Laplace-Beltrami operator with respect to any Riemannian metric on manifolds with boundary. It would also allow us to solve PDEs with various boundary conditions for the class of elliptic operators in [18, 23].

Diffusion maps has also recently seen use for solving fractional PDEs on manifolds without boundary. In a recent paper [1], the authors were able to adapt diffusion maps to the case of the fractional Laplacian on a compact manifolds without boundary with some additional assumptions. It stands to reason that our treatment of the boundary could also be applied in this case to the fractional case.

Future Research: Discrete Differential Forms for Data

The Hodge Laplacian

$$\Delta^k = -(\delta d + d\delta)$$

is a generalization of the Laplace-Beltrami operator which acts on differential k-forms on M and is formed from the exterior derivative d and its L^2-adjoint δ. The Hodge Decomposition Theorem relates the spectrum of the Hodge k-Laplacian to the De Rham cohomology of a manifold. Specifically, it says that each De Rham cohomology class $[\omega]$ in a compact manifold M has a unique representative form ω for which $\Delta^k\omega = 0$. Thus, one can compute unique cohomology classes by computing the eigenforms of the k-Laplacian with eigenvalue zero.

In the case that $k = 0$, the Hodge k-Laplacian is simply the Laplace-Beltrami operator acting on functions, and the Hodge theorem says that eigenfunctions of the Laplacian with eigenvalue zero correspond to distinct connected components. This interpretation helps explain why spectral clustering works by projecting data onto eigenvectors of a graph Laplacian corresponding to the set of smallest eigenvalues.

Currently, the higher Hodge k-Laplacians are well-studied in the numerical analysis community in what is called Discrete Exterior Calculus (DEC)[26] and Finite Element Exterior Calculus (FEEC)[3][2]. Both of these theories have seen success in numerical approximation of PDEs involving differential forms. What separates these methods is that they are discretized in such a way that classical theorems such as Stokes' theorem hold exactly. For instance, in such a scheme, a differential form on a Riemannian manifold is realized as a simplicial cochain on a simplicial complex. The exterior derivative is realized as the adjoint of the boundary operator, and the Hodge star is realized as

the Poincare duality map. Since the Hodge Laplacian can be constructed using only the exterior derivative and the Hodge Star, one can formulate the Hodge Laplacian in both the DEC and FEEC schemes.
One of the main obstructions to DEC and FEEC in the case of manifold learning is that they require an explicit triangulation of M into a simplicial complex. Thus, *a priori*, one must already know the homeomorphism class of M, which is an unrealistic assumption in the context of data.
In statistics, there also has been work using a graph Hodge Laplacian. In [27][39], the authors propose the HodgeRank algorithm used for statistical ranking of data that is incomplete or imbalanced. Although this idea is similar to one on manifolds, no convergence result analogous to Diffusion Maps has been done.
There is some evidence that the approximation of higher Hodge Laplacians is possible in manifold learning. Most notably, in the work of Singer and Wu[34][35][36], the authors develop a diffusion maps *graph connection Laplacian* acting on 1-forms. They then prove that it converges to the manifold connection Laplacian in the same sense as in diffusion maps. The manifold connection Laplacian is highly related to the Hodge Laplacian, and the Weitzenbock formula shows that the two operators are equal on a space with zero Ricci curvature. Thus, it seems that a Hodge Laplacian on data is not too far from reach, especially in the case of differential 1-forms.

Publications

Harbir Antil, Tyrus Berry, and John Harlim. Fractional diffusion maps. *arXiv preprint arXiv:1810.03952*, 2018.

Douglas Arnold, Richard Falk, and Ragnar Winther. Finite element exterior calculus: from hodge theory to numerical stability. *Bulletin of the American mathematical society*, 47(2):281–354, 2010.

Douglas N Arnold, Richard S Falk, and Ragnar Winther. Finite element exterior calculus, homological techniques, and applications. *Acta numerica*, 15:1–155, 2006.

Laurent Bartholdi, Thomas Schick, Nat Smale, and Steve Smale. Hodge theory on metric spaces. *Foundations of Computational Mathematics*, 12(1):1–48, 2012.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems*, pages 585–591, 2002.

Mikhail Belkin and Partha Niyogi. Convergence of laplacian eigenmaps. In *Advances in Neural Information Processing Systems*, pages 129–136, 2007.

Tyrus Berry and John Harlim. Variable bandwidth diffusion kernels. *Applied and Computational Harmonic Analysis*, 40(1):68–96, 2016.

Tyrus Berry and Timothy Sauer. Local kernels and the geometric structure of data. *Applied and Computational Harmonic Analysis*, 40(3):439–469, 2016.

Tyrus Berry and Timothy Sauer. Density estimation on manifolds with boundary. *Computational Statistics & Data Analysis*, 107:1–17, 2017.

Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

Bennett Chow, Peng Lu, and Lei Ni. *Hamilton's Ricci flow*, volume 77. American Mathematical Soc., 2006.

Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

Keenan Crane, Fernando De Goes, Mathieu Desbrun, and Peter Schröder. Digital geometry processing with discrete exterior calculus. In *ACM SIGGRAPH 2013 Courses*, page 7. ACM, 2013.

Mathieu Desbrun, Anil N Hirani, Melvin Leok, and Jerrold E Marsden. Discrete exterior calculus. *arXiv preprint math/0508341*, 2005.

Jaap Eldering. *Normally hyperbolic invariant manifolds: the noncompact case*, volume 2. Springer, 2013.

Tingran Gao. The diffusion geometry of fibre bundles: Horizontal diffusion maps. *Applied and Computational Harmonic Analysis*, 2019.

Tingran Gao, Jacek Brodzki, and Sayan Mukherjee. The geometry of synchronization problems and learning group actions. *Discrete & Computational Geometry*, pages 1–62, 2016.

Faheem Gilani and John Harlim. Approximating solutions of linear elliptic pde's on a smooth manifold using local kernel. *Journal of Computational Physics*, 395:563 – 582, 2019.

Alfred Gray. *Tubes*, volume 221. Birkhäuser, 2012.

Alexander Grigor'yan. Heat kernels and function theory on metric measure spaces. *Contemporary Mathematics*, 338:143–172, 2003.

Alexander Grigor'yan and Takashi Kumagai. On the dichotomy in the heat kernel two sided estimates. In *Analysis on Graphs and its Applications (P. Exner et al.(eds.)), Proc. of Symposia in Pure Math*, volume 77, pages 199–210, 2008.

Nadine Große and Cornelia Schneider. Sobolev spaces on riemannian manifolds with bounded geometry: general coordinates and traces. *Mathematische Nachrichten*, 286(16):1586–1613, 2013.

John Harlim, Daniel Sanz-Alonso, and Ruiyi Yang. Kernel methods for bayesian elliptic inverse problems on manifolds. *arXiv preprint arXiv:1910.10669*, 2019.

Matthias Hein, Jean-Yves Audibert, and Ulrike von Luxburg. Graph laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, 8(Jun):1325–1368, 2007.

Matthias Hein, Jean-Yves Audibert, and Ulrike Von Luxburg. From graphs to manifolds–weak and strong pointwise consistency of graph laplacians. In *International Conference on Computational Learning Theory*, pages 470–485. Springer, 2005.

Anil Nirmal Hirani. *Discrete exterior calculus*. PhD thesis, California Institute of Technology, 2003.

Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial hodge theory. *Mathematical Programming*, 127(1):203–244, 2011.

John M Lee. *Introduction to Riemannian manifolds*, volume 176. Springer, 2018.

Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

Thomas Schick. *Analysis on d-manifolds of bounded geometry, Hodge-de Rham isomorphism and L 2-index theorem*. PhD thesis, Johannes Gutenberg Universitat Mainz, 1996.

Thomas Schick. Manifolds with boundary and of bounded geometry. *Mathematische Nachrichten*, 223(1):103–120, 2001.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *Departmental Papers (CIS)*, page 107, 2000.

Amit Singer. From graph to manifold laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006.

Amit Singer and H-T Wu. Vector diffusion maps and the connection laplacian. *Communications on pure and applied mathematics*, 65(8):1067–1144, 2012.

Amit Singer and Hau-tieng Wu. Orientability and diffusion maps. *Applied and computational harmonic analysis*, 31(1):44–58, 2011.

Amit Singer and Hau-Tieng Wu. Spectral convergence of the connection laplacian from random samples. *Information and Inference: A Journal of the IMA*, 6(1):58–123, 2016.

Ryan Vaughn, Tyrus Berry, and Harbir Antil. Diffusion maps for embedded manifolds with boundary with applications to pdes. Preprint: http://math.gmu.edu/ rvaughn5/Publications/VaughnBerryAntil.pdf, 2019.

FW Warner. Extension of the rauch comparison theorem to submanifolds. *Transactions of the American Mathematical Society*, 122(2):341–356, 1966.

Qianqian Xu, Qingming Huang, Tingting Jiang, Bowei Yan, Weisi Lin, and Yuan Yao. Hodgerank on random graphs for subjective video quality assessment. *IEEE Transactions on Multimedia*, 14(3):844–857, 2012.