# Towards a mathematical foundation for machine learning

Tyrus Berry

George Mason University

Aug. 19, 2021

Supported by NSF-DMS 1854204 and 2006808

# WHY A MATHEMATICAL FOUNDATION?

Learning $f \in \mathcal{C}^s(\mathbb{R}^n, \mathbb{R})$ from $N$ data points

- Fixed data set $\Rightarrow$ engineering problem
- Growing data set $\Rightarrow$ Evolving model $\Rightarrow$ Convergence
- Need to know that our algorithm has a limiting behavior
- Consider the infinite data limit to insure stability
- Ask if the limiting model is the truth
- Mathematical structures provide prior models for truth

# VOLUME GROWS LIKE radius$^{\text{dimension}}$

Learning $f \in \mathcal{C}^s(\mathbb{R}^n, \mathbb{R})$ from $N$ data points $\Rightarrow$ Error $\propto N^{-s/n}$

Many instances:

- Vapnik-Chervonenkis (VC) dimension [1]
- Rademacher complexity [2]
- Kolmogorov width [3]
- Interpolation error in approximation theory [3, 4, 5]
- Bias-variance tradeoff (density estimation/regression) [6, 1]
- Neural networks [7, 8] and sparse grids [9]

Key counterexample: Data $\{x_i\} \subset \mathbb{R}^n$ and feature $y_i = f(||x_i||)$.

# AVOIDING THE CURSE

Learning $f \in \mathcal{C}^s(\mathbb{R}^n, \mathbb{R})$ from $N$ data points $\Rightarrow$ Error $\propto N^{-s/n}$

Coping mechanisms:

- **Smooth it away:** Assume $f$ is very smooth, ie. $s \propto n$
- **Independence:** Assume $Y = f(X)$ is conditionally independent of $X$ given $Z = g(X) \in \mathbb{R}^m$ with $m \ll n$.
- **Redundancy:** Assume $h(X) = 0$ for some $h \in \mathcal{C}^{m+1}(\mathbb{R}^n, \mathbb{R}^{n-m})$.

# SLOW CHANGE REQUIRES FEW NEIGHBORS

All machine learning methods interpolate from neighbors:

- **kNN and Local Linear Regression** ($x_{\mathrm{kNN}}$ is k-th nearest neighbor of $x$):

$$F(x) \approx \frac{1}{k} \sum_{||x-x_j|| \leq ||x-x_{\mathrm{kNN}}||} F(x_j) + a^\top (x - x_j)$$

- **Kernel Regression** ($h$ is bump function, eg. $h(s) = \exp(-s^2)$):

$$F(x) \approx \textstyle\sum_j c_j h((x - x_j)^\top A_j (x - x_j))$$

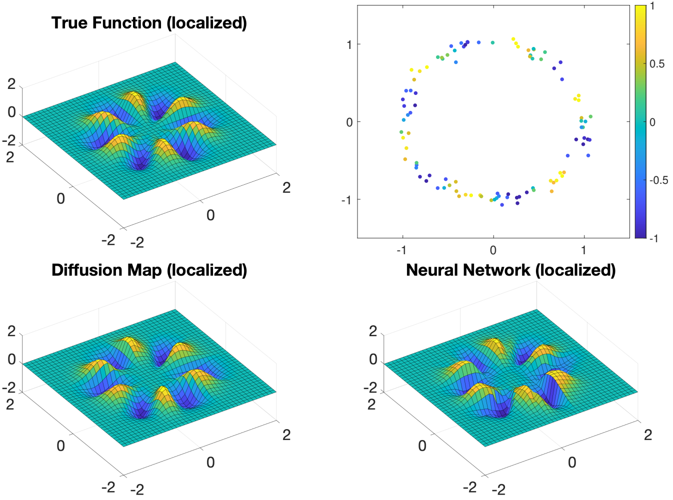
- **Neural Network** ($h$ is typically a sigmoid, but can also be a bump):

$$F(x) \approx \textstyle\sum_j c_j h(a_j^\top x + b_j) = \sum_j c_j h(a_j^\top (x - \tilde{x}_j))$$

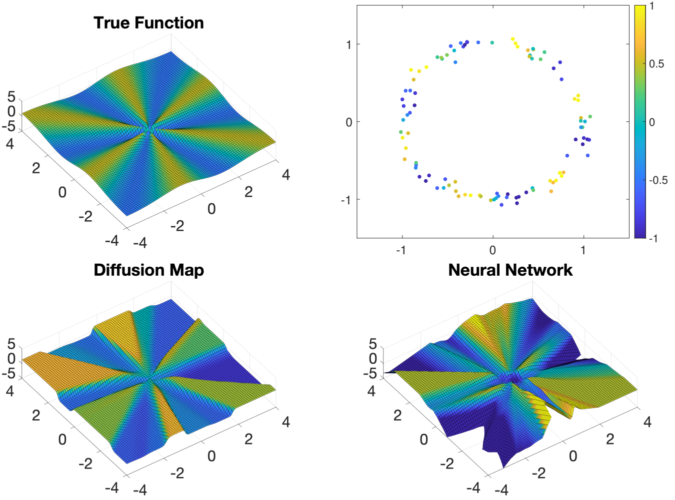
  (where we write $b_j = -a_j^\top \tilde{x}_j$)

- **Reservoir Computer**: Fix $a_j, b_j$, regression to find $c_j$

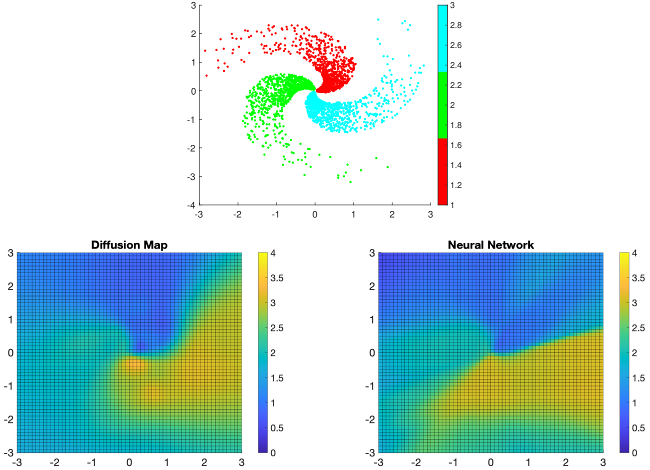
# Nyström vs. Deep net, $(r, \theta) \mapsto sin(6\theta)$



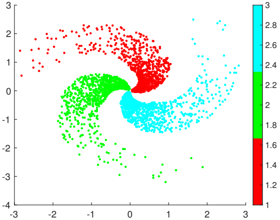
# NYSTRÖM VS. DEEP NET, $(r, \theta) \mapsto \sin(6\theta)$

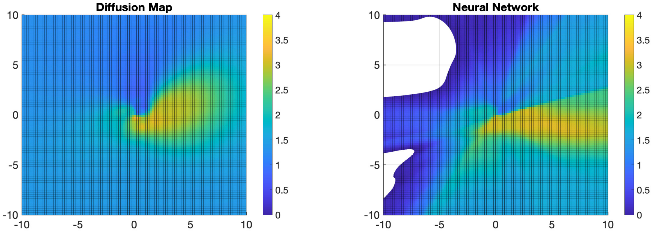


# Nyström vs. Deep net, Extrapolation



# Nyström vs. Deep net, Extrapolation





# INDEPENDENCE

Learning $f \in \mathcal{C}^s(\mathbb{R}^n, \mathbb{R})$ from $N$ data points $\Rightarrow$ Error $\propto N^{-s/n}$

- Want to learn $Y = f(X)$ where $f : \mathbb{R}^n \to \mathbb{R}$
- Assume there is a projection $\beta \in \mathbb{R}^{n \times m}$ such that

$$Y \perp\!\!\!\perp X \,|\, \beta^\top X$$

- Find $\beta$ using Sliced Inverse Regression (SIR) [10, 11]
- Learn $Y = \tilde{f}(\beta^\top X)$ since $\tilde{f} : \mathbb{R}^m \to \mathbb{R}$, $m \ll n$

# INDEPENDENCE

Detect person in crosswalk


Lots of variability, most is irrelevant

# INDEPENDENCE

More generally:

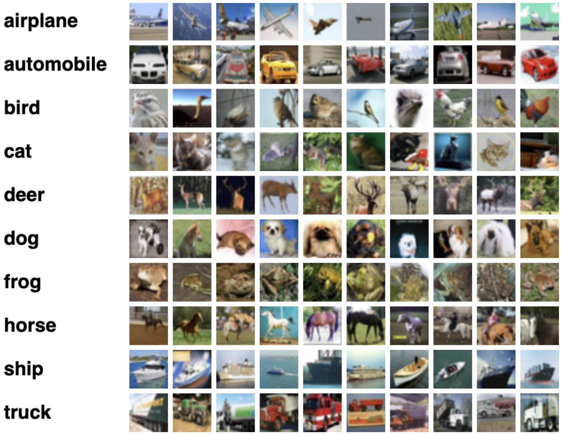
- Want to learn $P(Y \mid X)$
- Assume there is a map $\beta : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$Y \perp\!\!\!\perp X \mid \beta(X)$$

- If we can find $\beta$...
- Learning $P(Y \mid \beta(X))$ may be feasible

# INDEPENDENCE

CIFAR has many irrelevant modes




But they are combined nonlinearly with features

# REDUNDANCY

Unlike smoothness and independence, $f$ is not involved

- Redundancy assumes that most of $X \in \mathbb{R}^n$ is repeats
- E.g. $x_n = a_1 x_1 + \cdots + a_{n-1} x_{n-1}$ is a linear redundancy
- More generally if $AX = 0$ for some $A \in \mathbb{R}^{(n-m)\times n}$
- $X$ appears $n$-dim'l (extrinsic) but is really $m$-dim'l (intrinsic)
- PCA finds $A^{\perp}X \in \mathbb{R}^m$ where $[A\ A^{\perp}]$ is a basis
- The reduction helps learn any $f$

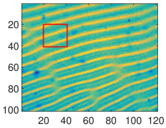
# Redundancy

- More generally assume $h(X) = 0$ for some $h : \mathbb{R}^n \to \mathbb{R}^{n-m}$
- **Sard's lemma:** If $h \in \mathcal{C}^{m+1}(\mathbb{R}^n, \mathbb{R}^{n-m})$ then regular values are dense in $\mathbb{R}^{n-m}$, so either 0 is regular or $\epsilon$ is regular
- **Regular Value Theorem:** The pre-image of a regular value under a smooth map is a manifold of dimension

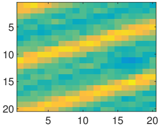
$$\dim(\text{domain}) - \dim(\text{range})$$

- **Upshot:** If $h(X) = 0 \in \mathbb{R}^{n-m}$ are smooth redundancies then $X = h^{-1}(0)$ is a manifold of dimension $m$
- Manifold learning leverages this nonlinear structure

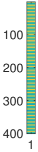
# FINDING HIDDEN STRUCTURE IN DATA




$\Rightarrow$

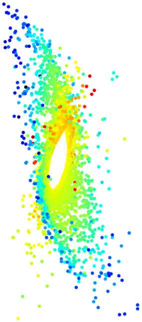


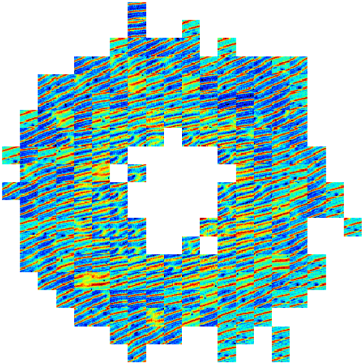

$\Rightarrow$




$\in \mathbb{R}^{400}$

The sub-image geometry:





# ROADMAP

- What is manifold learning? $\Rightarrow$ Estimate Laplacian, $\Delta$
- How to find the Laplacian? $\Rightarrow$ Graph Laplacian, $\mathbf{L}$
- Convergence $\mathbf{L} \to \Delta$ and overcoming limitations
- **Key result:** Extension to non-compact manifolds
- New graph construction based on this extension

# What is Manifold Learning?

- Geometric prior: Data on Riemannian manifold $\mathcal{M} \subset \mathbb{R}^m$
- **Goal**: Represent all the information about a manifold
- A smooth embedded manifold $\mathcal{M} \subset \mathbb{R}^m$ inherits:
  - A metric tensor $g_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$ (inner product)
  - $g$ completely determines the geometry of $\mathcal{M}$
  - A volume form $dV(x) = \sqrt{\det(g_x)}\, dx^1 \wedge \cdots \wedge dx^d$
- Laplace-Beltrami operator, $\Delta$, is equivalent to $g$
  - $\Delta f = \text{div} \circ \nabla = \frac{1}{\sqrt{|g|}} \partial_i g^{ij} \sqrt{|g|} \partial_j f$
  - $g(\nabla f, \nabla h) = \frac{1}{2}(f\Delta h + h\Delta f - \Delta(fh))$

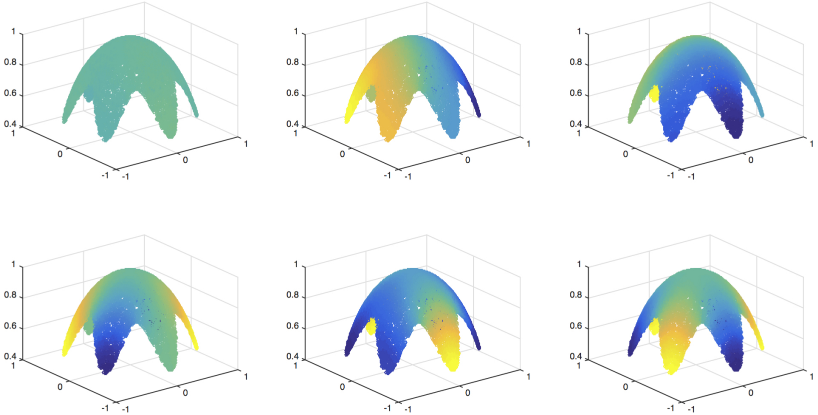
# What is Manifold Learning?

- **Manifold learning $\Leftrightarrow$ Estimating Laplace-Beltrami**
- **Hodge theorem:**
  Eigenfunctions $\Delta\varphi_i = \lambda_i \varphi_i$ orthonormal basis for $L^2(\mathcal{M}, g)$
- Smoothest functions: $\varphi_i$ minimizes the functional

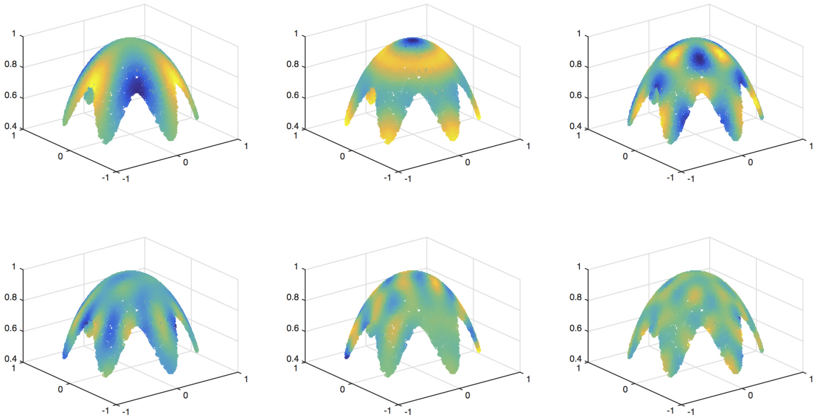
$$\lambda_i = \min_{\substack{f \perp \varphi_k \\ k=1,\ldots,i-1}} \left\{ \frac{\int_{\mathcal{M}} ||\nabla f||^2 \, dV}{\int_{\mathcal{M}} |f|^2 \, dV} \right\}$$

- Eigenfunctions of $\Delta$ are custom Fourier basis
  - Smoothest orthonormal basis for $L^2(\mathcal{M}, g)$
  - Can be used to define wavelet frame
  - Define the Sobolev spaces on $\mathcal{M}$

# Harmonic Analysis on Manifolds

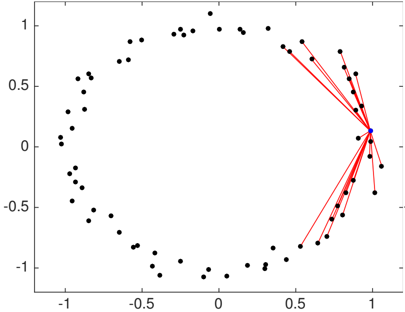


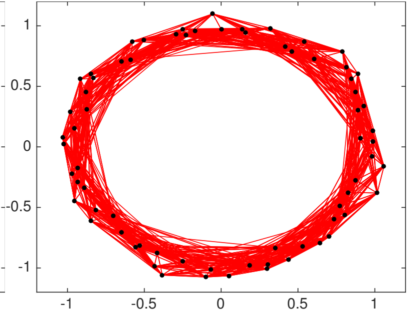
# Harmonic Analysis on Manifolds



# SO HOW DO WE FIND THE LAPLACIAN FROM DATA?

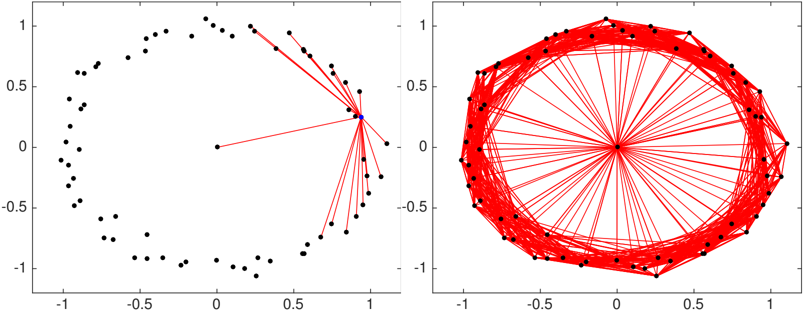
- Assume data lies on (or at least near) a manifold
- Approximate manifold with graph $\Rightarrow$ Connect nearby points



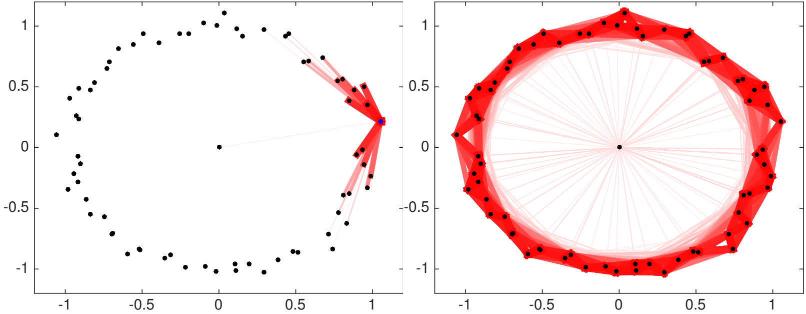
# SO HOW DO WE FIND THE LAPLACIAN FROM DATA?

- **Problem:** Noise and outliers can lead to *bridging*



# SO HOW DO WE FIND THE LAPLACIAN FROM DATA?

- To prevent bridging we weight the edges
- Edges are given weights $K_\delta(x, y) = e^{-\frac{||x-y||^2}{4\delta^2}}$



# So how do we find the Laplacian from data?

- Data set $\Rightarrow$ *weighted graph*
- Edge Weights defined by a kernel function

$$K_\delta(x_i, x_j) = e^{-\frac{||x_i - x_j||^2}{4\delta^2}}$$

- Bandwidth $\delta$ determines localization
- 'Adjacency' matrix: $\mathbf{K}_{ij} = K(x_i, x_j)$
- 'Degree' matrix: $\mathbf{D}_{ii} = \sum_j \mathbf{K}_{ij}$
- Normalized graph Laplacian: $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{K}$

# POINTWISE CONVERGENCE

**Theorem:** (Belkin & Niyogi, 2005, Singer, 2006)
For $\{x_i\}_{i=1}^N \subset \mathcal{M} \subset \mathbb{R}^m$ uniformly sampled on a compact manifold and for $\vec{f}_i = f(x_i)$ where $f \in C^3(\mathcal{M})$

$$\frac{1}{\delta^2}\left(\mathbf{L}\vec{f}\right)_i = \Delta f(x_i) + \mathcal{O}\left(\delta^2, \frac{1}{N^{1/2}\delta^{1+d/2}}\right)$$

$\delta =$ bandwidth
$N =$ number of points

# RESTRICTIONS THAT HAVE BEEN OVERCOME TO DEAL WITH REAL DATA:

- Arbitrary sampling (Coifman & Lafon, 'Diffusion maps', 2006)
- Other kernel functions (Berry & Sauer, 2015)
- Non-compact manifolds (Berry & Harlim, 2015)
- Boundary (Coifman & Lafon, 2006; R. Vaughn Thesis 2020)
- Spectral convergence (von Luxburg et al. 2008, Trillos et al. 2020, Berry & Sauer 2019)

# RESTRICTIONS THAT HAVE BEEN OVERCOME TO DEAL WITH REAL DATA:

- Arbitrary sampling (Coifman & Lafon, 'Diffusion maps', ACHA 2006)
- Other kernel functions (Thesis 2013; Berry & Sauer, ACHA 2015)
- Non-compact manifolds (Berry & Harlim, ACHA 2015)
- Boundary (Coifman & Lafon, ACHA 2006; Berry & Sauer, J. Comp. Stat. 2016)
- Spectral convergence (Luxburg et al., Ann. Stat. 2008, Berry & Sauer, submitted)

# LOCAL KERNELS

- A *local kernel* has exponential decay:

$$K_\delta(x, x + \delta y) < c_1 e^{-c_2||y||^2}$$

- **Theorem:** Symmetric local kernels converge to Laplacians
  - Every local kernel determines a geometry
  - Every geometry accessible by a local kernel
- Explain success of 'kernel methods' in data science:
  - KPCA: Kernel Principal Component Analysis
  - KSVM: Kernel Support Vector Machines
  - KDE: Kernel Density Estimation
  - RKHS: Reproducing Kernel Hilbert Spaces
  - Spectral Clustering (KPCA)

# RESTRICTIONS THAT HAVE BEEN OVERCOME TO DEAL WITH REAL DATA:

- Arbitrary sampling (Coifman & Lafon, 'Diffusion maps', ACHA 2006)
- Other kernel functions (Thesis 2013; Berry & Sauer, ACHA 2015)
- Non-compact manifolds (Berry & Harlim, ACHA 2015)
- Boundary (Coifman & Lafon, ACHA 2006; Berry & Sauer, J. Comp. Stat. 2016)
- Spectral convergence (Luxburg et al., Ann. Stat. 2008, Berry & Sauer, submitted)

# Tangible Manifolds

- Compute ambient distance $||x - y||_{\mathbb{R}^m}$
- Need localization in $d_{\mathcal{I}}(x,y) = \inf_{\gamma} \left\{ \int_0^1 |\gamma'(t)| \, dt \right\}$
- Assume ratio $R(x,y) = \frac{||x-y||_{\mathbb{R}^m}}{d_{\mathcal{I}}(x,y)}$ bounded away from zero
- We will use the exponential map to change variables
- Assume injectivity radius $\text{inj}(x)$ bounded away from zero

**Definition:** A manifold is uniformly tangible if there are lower bounds on $\text{inj}(x)$ and $\inf_{y \in \mathcal{M}} R(x,y)$ independent of $x$

# CONSISTENCY PART 1

- Matrix times vector converges to integral operator:

$$\left(\mathbf{K}\vec{f}\right)_i = \sum_{j=1}^{N} K_\delta(x_i, x_j) f(x_j) \xrightarrow{N\to\infty} \int_{\mathcal{M}} K_\delta(x_i, y) f(y)\, dV(y)$$

- Assume kernel has fast decay: $K_\delta(x, y) < e^{-||x-y||^2/\delta^2}$
- Localize integral, requires $R(x_i, y) = \frac{||x_i - y||}{d_I(x_i, y)} > 0$

$$\left(\mathbf{K}\vec{f}\right)_i \to \int_{\mathcal{M}\cap \exp_{x_i}(B_\delta(0))} K_\delta(x_i, y) f(y)\, dV(y) + \mathcal{O}(\delta^k)$$

- Change variables to the tangent space $y = \exp_{x_i}(s)$:

$$\left(\mathbf{K}\vec{f}\right)_i \to \int_{B_\delta(0)} K_\delta(x_i, \exp_{x_i}(s)) f(\exp_{x_i}(s))\, ds + \mathcal{O}(\delta^k)$$

- Requires injectivity radius $\mathrm{inj}(x_i) > \delta > 0$

# Consistency Part 2

- Taylor expansion in normal coordinates:

$$f(\exp_x(s)) = f(x) + \nabla f(x) \cdot s + \frac{1}{2} s^\top H(f \circ \exp_x)(0) s$$

- Symmetric kernel $\Rightarrow$ Odd terms integrate to zero

$$\left(\mathbf{K}\vec{f}\right)_i \to \int_{||s||<\delta} \left(K(||s||) + \mathcal{O}(\delta^2 s_i^4) K'(||s||)/||s||\right) \cdot$$
$$\left(f(x_i) + \delta \nabla f(x_i) \cdot s + \frac{\delta^2}{2} s^\top H(f \circ \exp_{x_i})(0) s\right) ds + \mathcal{O}(\delta^4)$$
$$= f(x_i) + m\delta^2 (f(x_i)\omega(x) + \Delta f(x_i)) + \mathcal{O}(\delta^4)$$

- Normalize: $\mathbf{D}^{-1}\mathbf{K}\vec{f} = \frac{\mathbf{K}\vec{f}}{\mathbf{K}\vec{1}} \to \vec{f} + m\delta^2 \overrightarrow{\Delta f} + \mathcal{O}(\delta^4)$
- **Consistency:** $\frac{1}{m\delta^2}(\mathbf{D}^{-1}\mathbf{K} - \mathbf{I})\vec{f} \to \overrightarrow{\Delta f} + \mathcal{O}(\delta^2)$

# Consistency is not enough!

- Extend to arbitrary sampling $x_i \sim q$ (Coifman & Lafon)
- **Variance:** $\mathbb{E}[((L\vec{f})_i - \Delta f(x_i))^2] = \mathcal{O}\left(\frac{q(x_i)^{3-4d}}{N\delta^{2+d}}\right)$
- Negative exponent: $3 - 4d < 0$
- As density $q$ approaches zero the variance blows up!
- **Solution:** Variable bandwidth

Berry and Harlim (ACHA, 2015)

# Variable Bandwidth Kernels

We introduced the variable bandwidth kernel:

$$K_{\delta,\beta}(x,y) = K\left(\frac{||x-y||}{\delta\sqrt{q(x)^{\beta}q(y)^{\beta}}}\right)$$

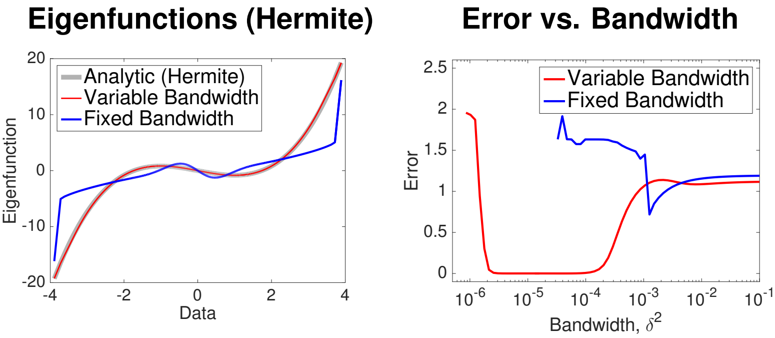
**Theorem** (Berry and Harlim, ACHA, 2015):

$$\mathbf{L}_{\delta,\alpha,\beta}\vec{f} = \Delta f + c_1 \nabla f \cdot \nabla \log q + \mathcal{O}\left(\delta^2, \frac{q^{-c_2}}{\sqrt{N}h^{1+d/2}}\right)$$

- Operator defined by: $c_1 = 2 - 2\alpha + d\beta + 2\beta$
- Variance determined by: $c_2 = 1/2 - 2\alpha + 2d\alpha + d\beta/2 + \beta$

# Example: Variable Bandwidth Kernel

**Gaussian data:** Brownian motion in quadratic potential




Berry and Harlim (ACHA, 2015)

# SUMMARY OF MANIFOLD LEARNING

- Manifold learning $\Leftrightarrow$ Estimating Laplace-Beltrami
- Can estimate Laplace-Beltrami with a graph Laplacian
- For a non-compact manifold:
  - Manifold must be tangible
  - Requires a variable bandwidth kernel
- Other contributions:
  - Access any desired geometry (local kernels)
  - Manifolds with boundary
  - Spectral convergence

# Beyond Manifold Learning

- Data never really lies on a manifold (due to noise)
- A manifold is a measure zero set
- Data is never sampled from a measure zero set
- **Solution 1:** Spectral robustness for bounded noise (Coifman and Lafon), but lose convergence
- **Solution 2:** Manifold + Noise, requires semi-geodesic coordinates, need new algorithms to regain convergence
- **Solution 3:** Generalize beyond manifolds
  - Metric measure spaces
  - Gromov-Hausdorff limits of manifolds

# CONTINUOUS K-NEAREST NEIGHBORS (CKNN)

## Building unweighted graphs from data (TDA)

**CkNN Graph**: Edge $\{x, y\}$ added if $\frac{||x-y||}{\sqrt{||x-x_k||\ ||y-y_k||}} < \delta$

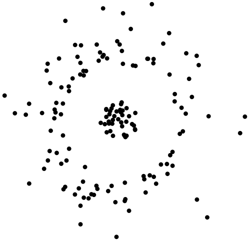
- $x_k = k$-th nearest neighbor of $x$
- Unnormalized graph Laplacian: $\mathbf{L}_{\mathrm{un}} = \mathbf{D} - \mathbf{K}$
- **Corollary:** $\mathbf{L}_{\mathrm{un}}\vec{f} \to \overrightarrow{\Delta_{\tilde{g}} f}$ where $(\tilde{g} = q^{2/d}g, d\tilde{V} = q\,dV)$
- **New result:** Spectral convergence $\mathbf{L}_{\mathrm{un}} \to \Delta_{\tilde{g}}$
- Consistency of CkNN clustering:
  - Conn. comp. of graph $\Leftrightarrow$ Kernel of $L_{\mathrm{un}}$
  - Conn. comp. of $\mathcal{M}$ $\Leftrightarrow$ Kernel of $\Delta_{\tilde{g}}$ (Hodge theorem)

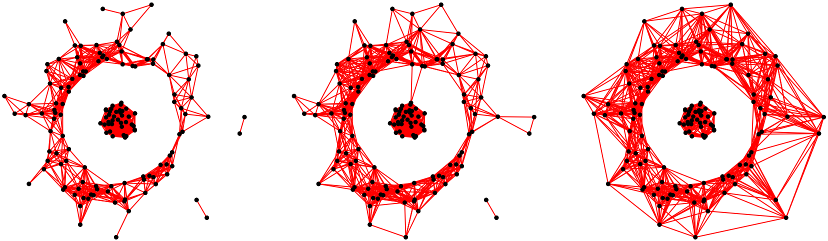
(Berry & Harlim (ACHA, 2015); Berry & Sauer (in review))

# CkNN yields improved graph construction

2D Gaussian with annulus removed:

Persistent vs. consistent homology



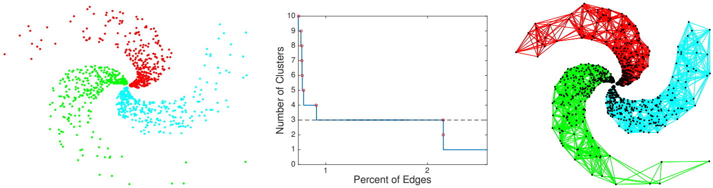


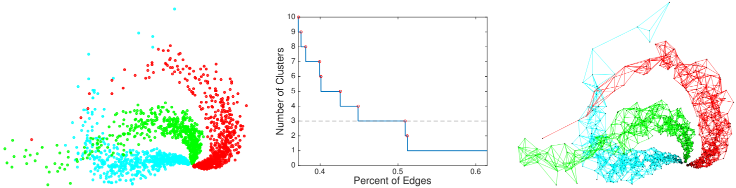
Small bandwidth

Large bandwidth

CkNN

# IMPROVED CLUSTERING USING CKNN



# Conformally Invariant Diffusion Maps (CIDM)

- Data samples $\{x_i\}_{i=1}^N \subset \mathcal{M} \subset \mathbb{R}^n$ of volume $p_{\mathrm{eq}}\, dV$
- Continuous k-Nearest Neighbors (CkNN) dissimilarity:

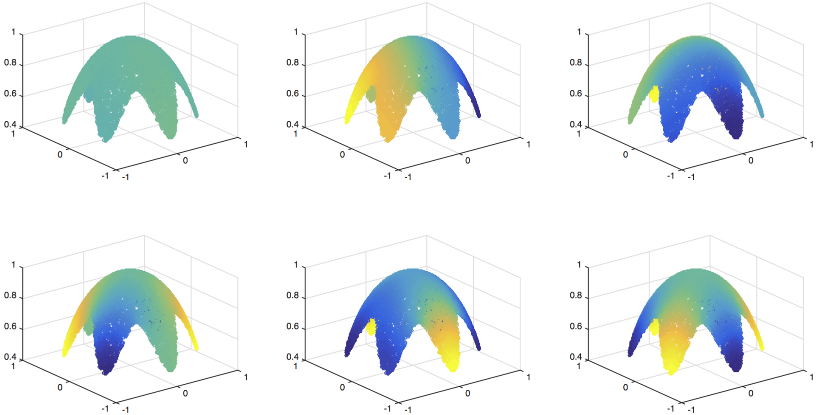
$$d(x_i, x_j) \equiv \frac{||x_i - x_j||}{\sqrt{||x_i - x_{kNN(i)}||\,||x_j - x_{kNN(j)}||}}$$

- Variable bandwidth kernel, $K_{ij} = \exp\left(\frac{-d(x_i,x_j)^2}{\delta^2}\right)$
- Degree matrix $D_{ii} = \sum_j K_{ij}$ (diagonal)
- Graph Laplacian, $L = \frac{D-K}{\delta^{d+2}}$
- **Theorem:** $L\vec{f} = \Delta_{\hat{g}} f + \mathcal{O}\left(\delta^2, N^{-1/2}\delta^{-1-d/2}\right),\ \hat{g} = p_{\mathrm{eq}}^{2/d} g$
- **Solve:** $(I - D^{-1/2}KD^{-1/2})\vec{v} = \lambda\vec{v}$, set $\vec{\varphi} = D^{-1/2}\vec{v}$

# Harmonic Analysis on Manifolds/Data Sets

- Manifolds with boundary, (R. Vaughn)

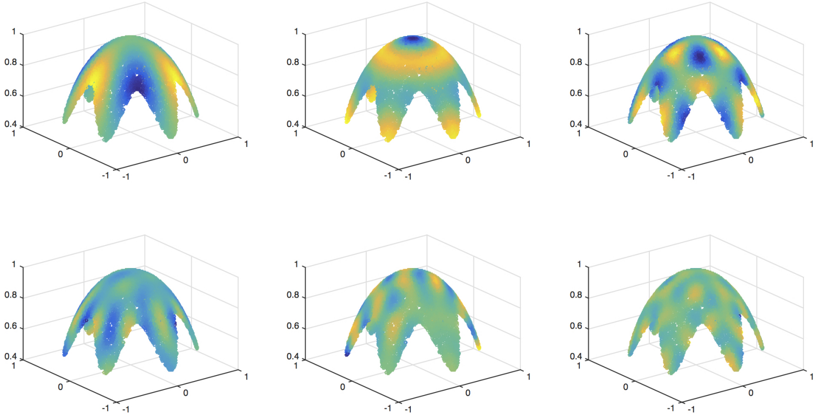
$$\vec{h}^{\top} L \vec{f} \to \int (\nabla h \cdot \nabla f)\, p_{\mathrm{eq}}\, dV$$



# Harmonic Analysis on Manifolds/Data Sets

- Manifolds with boundary, (R. Vaughn)

$$\vec{h}^{\top} L \vec{f} \rightarrow \langle\langle \nabla_{\hat{g}} h, \nabla_{\hat{g}} f \rangle\rangle_{\hat{g}} = \int \hat{g}(\nabla_{\hat{g}} h, \nabla_{\hat{g}} f)\, dV_{\hat{g}}$$



Code and papers available at:

http://math.gmu.edu/˜berry/

**Manifold Learning Papers Discussed**

- B. and Giannakis, *Spectral Exterior Calclulus.*
- R. Vaughn *Diffusion Maps for Manifolds with Boundary.*
- B. and Sauer, *Consistent Manifold Representation for Topological Data Analysis.*
- Coifman and Lafon, *Diffusion maps.*
- B. and Harlim, *Variable Bandwidth Diffusion Kernels.*
- B. and Sauer, *Local Kernels and Geometric Structure of Data.*

# References

[1] V. Vapnik, The nature of statistical learning theory. Springer (2000).

[2] M. Mohri, A. Rostamizadeh, and A. Talwalkar, Foundations of machine learning. MIT press (2018).

[3] A. Pinkus, N-widths in Approximation Theory. Vol. 7. Springer Science & Business Media, (2012).

[4] R. A. DeVore and G. G. Lorentz. Constructive approximation. Vol. 303. Springer Science & Business Media, (1993).

[5] R. A. DeVore, Nonlinear approximation. Acta numerica 7, 51-150 (1998).

[6] D. W. Scott and S. R. Sain. Multidimensional density estimation. Handbook of statistics 24, 229-261 (2005).

[7] A. R. Barron, Universal approximation bounds for superpositions of a sigmoidal function. IEEE Transactions on Information theory 39, no. 3, 930-945 (1993).

[8] A. R. Barron, Approximation and estimation bounds for artificial neural networks. Machine learning 14, no. 1, 115-133 (1994).

[9] H. J. Bungartz and M. Griebel, Sparse grids. Acta numerica, 13, 147-269 (2004).

[10] K. Li, Sliced Inverse Regression for Dimension Reduction. Journal of the American Statistical Association, 86(414), (1991).

[11] Y. Li, L. Zhu, Asymptotics for Sliced Average Variance Estimation. The Annals of Statistics, 35(1), 41-69 (2007).