Diffusion analysis of and on graphs, and high-dimensional data

Mauro Maggioni

Duke Mathematics and Computer Science

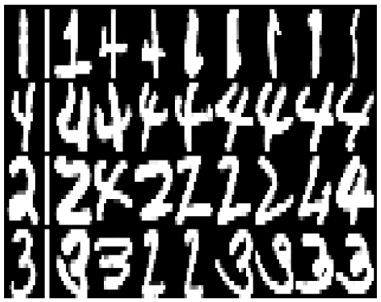
ICIAM 2007, 18/07/07

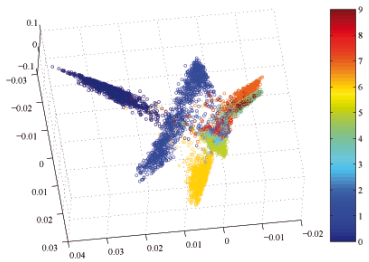
Plan

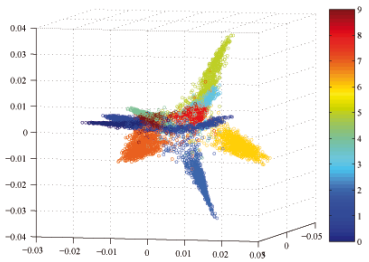
- Setting and Motivation
- Diffusion on Graphs
- Eigenfunction embedding
- Multiscale construction
- Examples and applications
- Conclusion

Handwritten Digits

Data base of about $60,000$ 28×28 gray-scale pictures of handwritten digits, collected by USPS. Goal: automatic recognition. It is a point cloud in 28^2 dimensions. We can think of being given this cloud, and some points are labeled by the digit they correspond to, and we would like to predict the digit corresponding to each point.



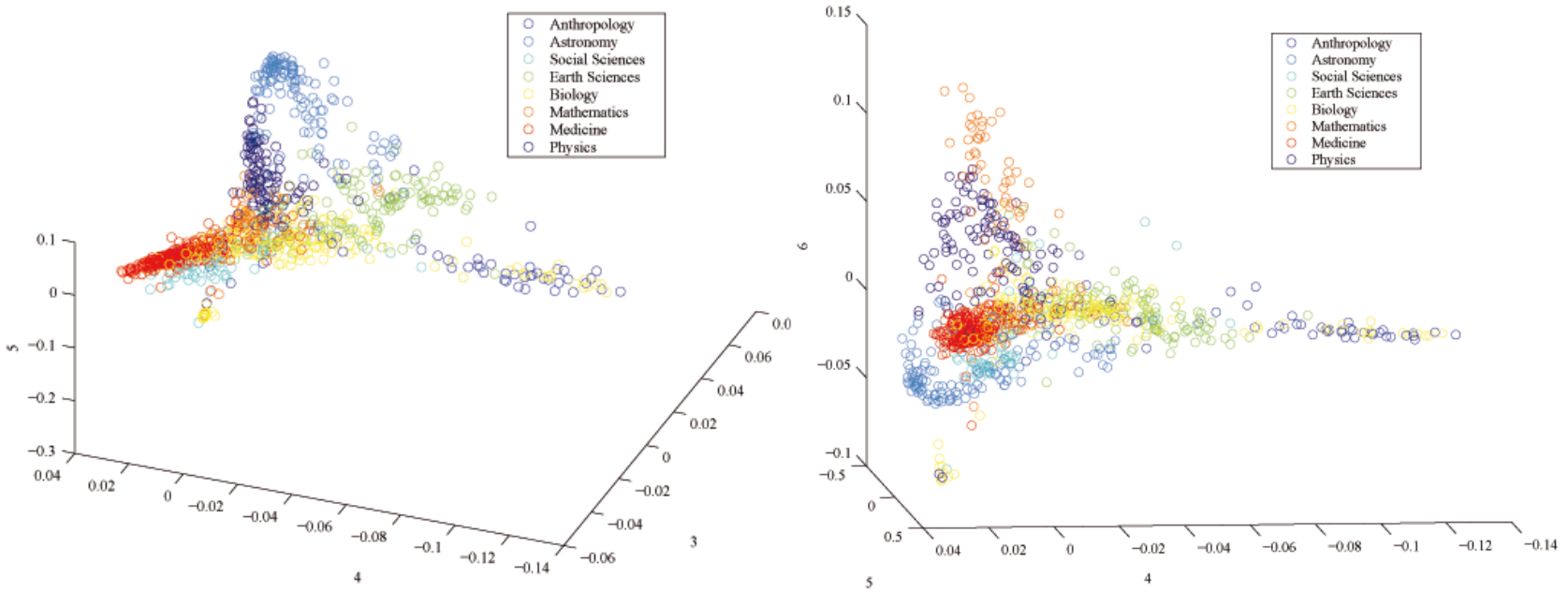




Set of $10,000$ picture (28 by 28 pixels) of 10 handwritten digits. Color represents the label (digit) of each point.

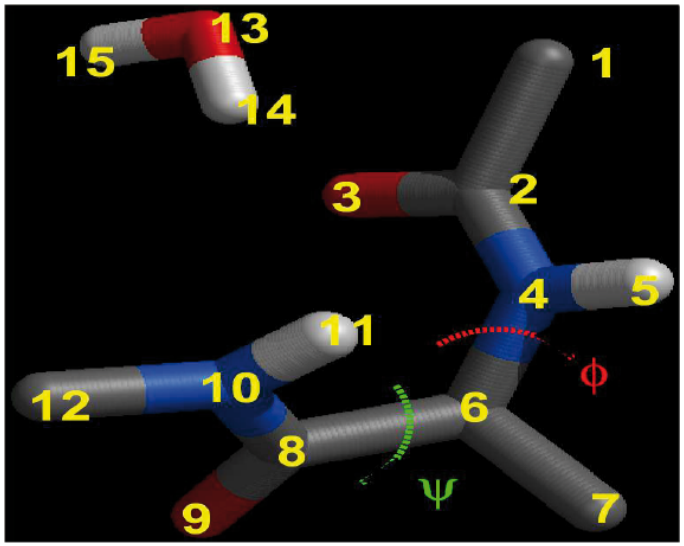
Text documents

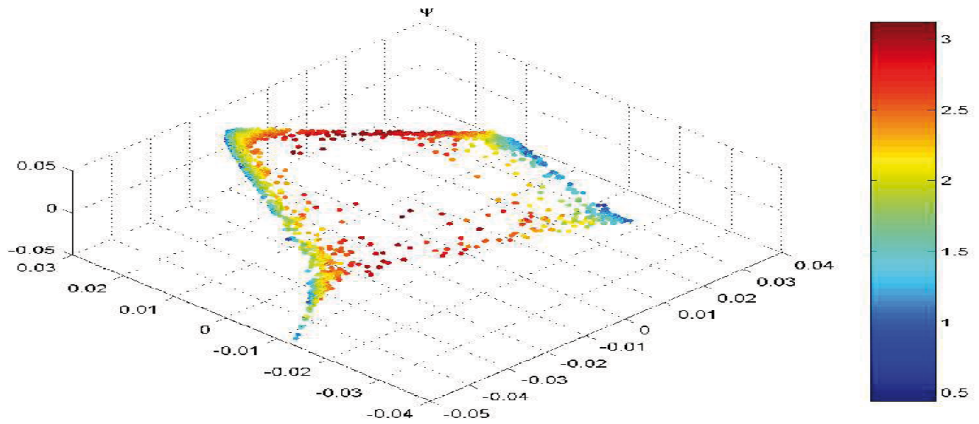
1000 Science News articles, from 8 different categories. We compute about 10000 coordinates, *i*-th coordinate of document *d* represents frequency in document *d* of the *i*-th word in a fixed dictionary.



An example from Molecular Dynamics

The dynamics of a small protein in a bath of water molecules is approximated by a Langevin system of stochastic equations $\dot{x} = -\nabla U(x) + \dot{w}$.





The set of states of the protein is a noisy set of points in $\mathbb{R}^{36}$.

Goals

- Find parametrizations for the data: manifold learning, dimensionality reduction. Ideally: number of parameters equal to, or comparable with, the intrinsic dimensionality of data (as opposed to the dimensionality of the ambient space), such a parametrization should be at least approximately an isometry with respect to the manifold distance, and finally it should be stable under perturbations of the manifold. In the examples above: variations in the handwritten digits, topics in the documents, angles in molecule...
- Construct useful dictionaries of functions on the data: approximation of functions on the manifold, predictions, learning.

Goals

- Find parametrizations for the data: manifold learning, dimensionality reduction. Ideally: number of parameters equal to, or comparable with, the intrinsic dimensionality of data (as opposed to the dimensionality of the ambient space), such a parametrization should be at least approximately an isometry with respect to the manifold distance, and finally it should be stable under perturbations of the manifold. In the examples above: variations in the handwritten digits, topics in the documents, angles in molecule...
- Construct useful dictionaries of functions on the data: approximation of functions on the manifold, predictions, learning.

Graphs associated with data sets

Assume the data $X = \{x_i\} \subset \mathbb{R}^n$. Assume we can assign local similarities via a kernel function $K(x_i, x_j) \geq 0$. For example $K_\sigma(x_i, x_j) = e^{-||x_i - x_j||^2/\sigma}$.

Model the data as a *weighted graph* (G, E, W): vertices represent data points, edges connect x_i, x_j with weight $W_{ij} := K(x_i, x_j)$, when positive.

Note 1: K typically depends on the type of data.

Note 2: K should be "local", i.e. close to 0 for points not sufficiently close.

Let $D_{ii} = \sum_j W_{ij}$ and

$$\underbrace{\mathcal{L} = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}}}_{\text{normalized Laplacian}}, \quad \underbrace{H = e^{-t\mathcal{L}}}_{\text{Heat kernel}}, \quad \underbrace{P = D^{-1}W}_{\text{random walk}}, \quad \underbrace{T = I - \mathcal{L}}_{\text{symmetrized random wal}}$$

Graphs associated with data sets

Assume the data $X = \{x_i\} \subset \mathbb{R}^n$. Assume we can assign local similarities via a kernel function $K(x_i, x_j) \geq 0$. For example $K_\sigma(x_i, x_j) = e^{-||x_i - x_j||^2/\sigma}$.
Model the data as a *weighted graph* (G, E, W): vertices represent data points, edges connect x_i, x_j with weight $W_{ij} := K(x_i, x_j)$, when positive.
Note 1: K typically depends on the type of data.
Note 2: K should be "local", i.e. close to 0 for points not sufficiently close.
Let $D_{ii} = \sum_j W_{ij}$ and

$$\underbrace{\mathcal{L} = D^{-\frac{1}{2}}(D - W)D^{-\frac{1}{2}}}_{\textit{normalized Laplacian}}, \quad \underbrace{H = e^{-t\mathcal{L}}}_{\textit{Heat kernel}}, \quad \underbrace{P = D^{-1}W}_{\textit{random walk}}, \quad \underbrace{T = I - \mathcal{L}}_{\textit{symmetrized random wal}}$$

The Heat Kernel and the Laplacian on Manifolds

Different ways of using the diffusion process T:

- Look at T for very large time (T^t for t large) $\longrightarrow$ eigenfunctions of T $\longrightarrow$ Fourier analysis *on* the data, "basis method"
- Look at T for small time ($T, T^2, \ldots, T^k$, k constant) $\longrightarrow$ it is diffusion *on* the set $\longrightarrow$ "PDE" method, "no basis"
- Look at T at all time scales ($T, T^2, T^4, \ldots, T^{2^j}, \ldots$) $\longrightarrow$ multiscale analysis of both functions and the diffusion process $\longrightarrow$ wavelets and multiscale dynamical processes.

The Heat Kernel and the Laplacian on Manifolds

Different ways of using the diffusion process T:

- Look at T for very large time (T^t for t large) $\longrightarrow$ eigenfunctions of T $\longrightarrow$ Fourier analysis *on* the data, "basis method"
- Look at T for small time ($T, T^2, \ldots, T^k$, k constant) $\longrightarrow$ it is diffusion *on* the set $\longrightarrow$ "PDE" method, "no basis"
- Look at T at all time scales ($T, T^2, T^4, \ldots, T^{2^j}, \ldots$) $\longrightarrow$ multiscale analysis of both functions and the diffusion process $\longrightarrow$ wavelets and multiscale dynamical processes.

Different ways of using the diffusion process T:

- Look at T for very large time (T^t for t large) $\longrightarrow$ eigenfunctions of T $\longrightarrow$ Fourier analysis *on* the data, "basis method"
- Look at T for small time ($T, T^2, \ldots, T^k$, k constant) $\longrightarrow$ it is diffusion *on* the set $\longrightarrow$ "PDE" method, "no basis"
- Look at T at all time scales ($T, T^2, T^4, \ldots, T^{2^j}, \ldots$) $\longrightarrow$ multiscale analysis of both functions and the diffusion process $\longrightarrow$ wavelets and multiscale dynamical processes.

Eigenfunction Embedding theorems, I

[Joint with P.W. Jones and R. Schul]

We ask whether eigenfunctions of the Laplacian can be used to parametrize Euclidean domains and manifolds, in which generality this may be true, and which conditions such an embedding may satisfy. Originally suggested by Bérard, Besson and Gallot ('84,'94) - however in their case they map to ℓ^2, they require smoothness of the manifold, and the map is not an isometry (or close to it). Other recent proposed techniques include isomap, lle, Hessian eigenmaps, maximum variance embedding; we are aware of proven results only for isomap and Hessian eigenmap, and in both cases the assumptions require the manifold to the isometric image of a Euclidean domain.

Eigenfunction Embedding theorems, I (cont'd)

Independently of the boundary conditions, we will denote by Δ the Laplacian on Ω, For the purpose of this paper (both the Dirichlet and Neumann case) we restrict our study to domains where the spectrum is discrete and the corresponding heat kernel can be written as

$$K_t^{\Omega}(z,w) = \sum \varphi_j(z)\varphi_j(w)e^{-\lambda_j t}\,.$$

where the $\{\varphi_j\}$ form an orthonormal basis for the appropriate Hilbert space with eigenvalues $0 \leq \lambda_0 \leq \cdots \leq \lambda_j \leq \ldots$. We also require

$$\#\{j : \lambda_j \leq T\} \leq C_{Weyl,\Omega}\, T^{\frac{d}{2}}\,|\Omega|\,.$$

Dirichlet case: OK, Neumann: possible problems.

Eigenfunction Embedding theorems, II

[Joint with P.W. Jones and R. Schul]

Theorem (Embedding via Eigenfunctions, for Euclidean domains)

Let Ω be a domain in $\mathbb{R}^d$, with $|\Omega| = 1$, and boundary as above. There are constants $c_1, \ldots, c_6 > 0$ that depend only on d and $C_{Weyl,\Omega}$, such that the following hold. For any $z \in \Omega$, let $R_z \leq \operatorname{dist}(z, \partial\Omega)$. Then there exist $i_1, \ldots, i_d$ and constants $c_6 R_z^{\frac{d}{2}} \leq \gamma_1 = \gamma_1(z), ..., \gamma_d = \gamma_d(z) \leq 1$ such that:

(a) $\Phi : B_{c_1 R_z}(z) \to \mathbb{R}^d$, *defined by*

$$x \mapsto (\gamma_1 \varphi_{i_1}(x), \ldots, \gamma_d \varphi_{i_d}(x))$$

satisfies, for any $x_1, x_2 \in B(z, c_1 R_z)$,

$$\frac{c_2}{R_z} ||x_1 - x_2|| \leq ||\Phi(x_1) - \Phi(x_2)|| \leq \frac{c_3}{R_z} ||x_1 - x_2|| .$$

(b) $c_4 R_z^{-2} \leq \lambda_{i_1}, \ldots, \lambda_{i_d} \leq c_5 R_z^{-2}$.

Eigenfunction Embedding theorems, III

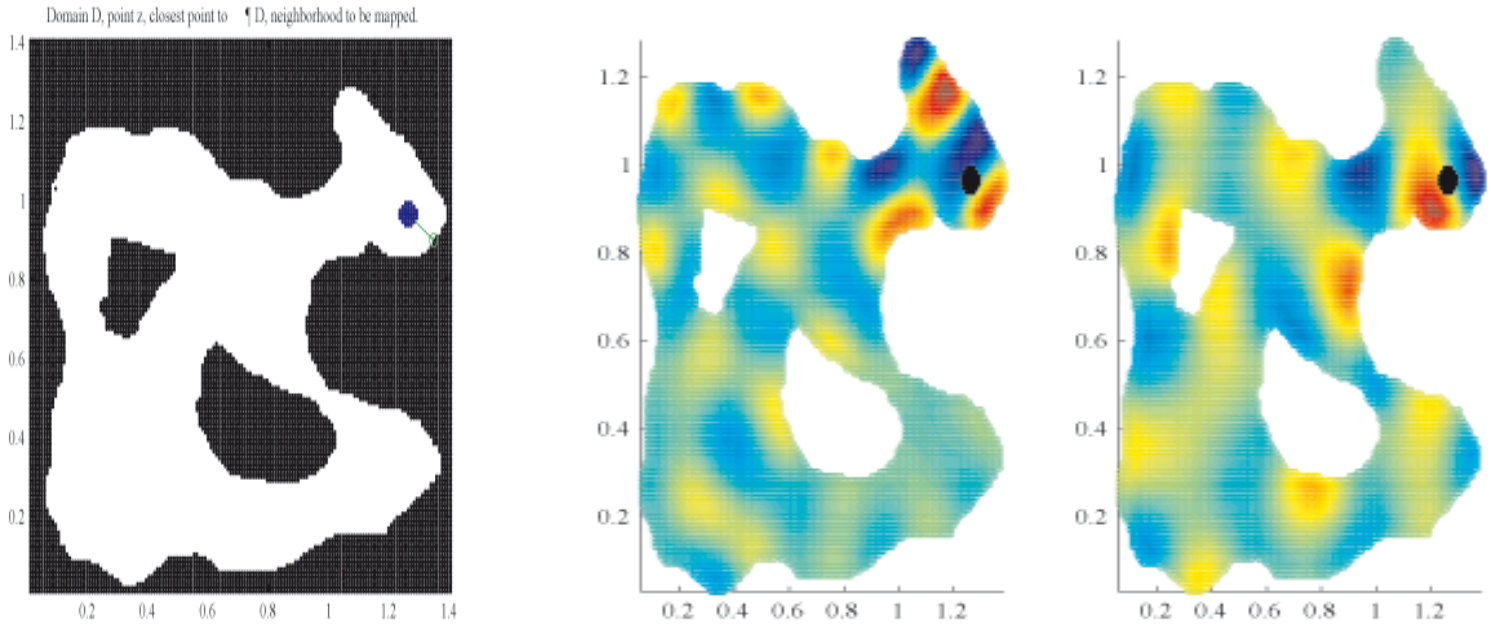


Figure: Top left: a non-simply connected domain in $\mathbb{R}^2$, and the point z with its neighborhood to be mapped. Top right: the image of the neighborhood under the map. Bottom: Two eigenfunctions for mapping.

Eigenfunction Embedding theorems, IV

[Joint with P.W. Jones and R. Schul]

Let $\mathcal{M}$ be a smooth, d-dimensional compact manifold, possibly with boundary. Suppose we are given a metric tensor g on $\mathcal{M}$ which is $\mathcal{C}^\alpha$ for some $\alpha > 0$. For any $z_0 \in \mathcal{M}$, let (U, x) be a coordinate chart such that $z_0 \in U$, $g^{il}(x(z_0)) = \delta^{il}$ and for any $w \in U$, and any $\xi, \nu \in \mathbb{R}^d$, $c_{\min}(g)||\xi||^2_{\mathbb{R}^d} \leq \sum_{i,j=1}^d g^{ij}(x(w))\xi_i\xi_j$, $\sum_{i,j=1}^d g^{ij}(x(w))\xi_i\nu_j \leq c_{\max}(g)||\xi||_{\mathbb{R}^d}\,||\nu||_{\mathbb{R}^d}$.
We let $r_{\mathcal{M}}(z_0) = \sup\{r > 0 : B_r(x(z_0)) \subseteq x(U)\}$.

$$\Delta_{\mathcal{M}} f(x) = -\frac{1}{\sqrt{\det g}} \sum_{i,j=1} \partial_j \left(\sqrt{\det g}\, g^{ij}(x) \partial_i f \right)(x)\,.$$

Eigenfunction Embedding theorems, IV

[Joint with P.W. Jones and R. Schul]

Theorem (Embedding via Eigenfunctions, for Manifolds)

Let $(\mathcal{M}, g)$, $z \in \mathcal{M}$ be a d dimensional manifold and (U, x) be a chart as above. Also, assume $|\mathcal{M}| = 1$. There are constants $c_1, \ldots, c_6 > 0$, depending on $d, c_{\min}, c_{\max}, ||g||_{\alpha \wedge 1}, \alpha \wedge 1$, and $C_{Weyl,\Omega}$, such that the following hold. Let $R_z = r_{\mathcal{M}}(z)$. Then there exist $i_1, \ldots, i_d$ and constants $c_6 R_z^{\frac{d}{2}} \leq \gamma_1 = \gamma_1(z) \, , ..., \gamma_d = \gamma_d(z) \leq 1$ such that:

(a) *the map $\Phi : B_{c_1 R_z}(z) \to \mathbb{R}^d$, defined by*

$$x \mapsto (\gamma_1 \varphi_{i_1}(x), \ldots, \gamma_d \varphi_{i_d}(x))$$

such that for any $x_1, x_2 \in B(z, c_1 R_z)$

$$\frac{c_2}{R_z} \, d_{\mathcal{M}}(x_1, x_2) \leq ||\Phi(x_1) - \Phi(x_2)|| \leq \frac{c_3}{R_z} \, d_{\mathcal{M}}(x_1, x_2) \, .$$

(b) $c_4 R_z^{-2} \leq \lambda_{i_1}, \ldots, \lambda_{i_d} \leq c_5 R_z^{-2}$.

Heat triangulation theorem, V

[Joint with P.W. Jones and R. Schul]

Theorem (Heat Triangulation Theorem)

Let $(\mathcal{M}, g)$, $z \in \mathcal{M}$ and (U, x) be as above, where we now allow $|\mathcal{M}| = +\infty$. Let $R_z \leq \min\{1, r_{\mathcal{M}}(z)\}$. Let $p_1, ..., p_d$ be d linearly independent directions. There are constants $c_1, \ldots, c_5 > 0$, depending on $d, c_{\min}, c_{\max}, ||g||_{\alpha \wedge 1}, \alpha \wedge 1$, and the smallest and largest eigenvalues of the Gramian matrix $(\langle p_i, p_j \rangle)_{i=1,...,d}$, such that the following holds. Let y_i be so that $y_i - z$ is in the direction p_i, with $c_4 R_z \leq d_{\mathcal{M}}(y_i, z) \leq c_5 R_z$ for each $i = 1, \ldots, d$ and let $t_z = c_6 R_z^2$. The map

$$\Phi : B_{c_1 R_z}(z) \to \mathbb{R}^d$$
$$x \mapsto (R_z^d K_{t_z}(x, y_1)), \ldots, R_z^d K_{t_z}(x, y_d))$$

satisfies, for any $x_1, x_2 \in B_{c_1 R_z}(z)$,

$$\frac{c_2}{R_z} d_{\mathcal{M}}(x_1, x_2) \leq ||\Phi(x_1) - \Phi(x_2)|| \leq \frac{c_3}{R_z} d_{\mathcal{M}}(x_1, x_2) .$$

Idea of proof

When $t \sim \lambda^{-1} \sim R_z^2$, prove heat kernel resembles Euclidean Dirichlet heat kernel in a ball, in terms of its size *and gradient*, from above and *below*, with constants independent of the smoothness of the manifold. This will give the heat triangulation theorem, since the heat kernel has the correct gradient estimates. For the eigenfunction theorem, look at the spectral expansion of the heat kernel, and observe that the main contribution to that series comes from frequencies in the correct range. So not all eigenfunctions in that range have small gradient $\longrightarrow$ pigeon-hole $\longrightarrow$ find eigenfunction with gradient of the correct size in a given direction $\longrightarrow$ repeat over directions, each orthogonal to the span of the gradients of the previously chosen eigenfunctions.
How to prove smoothness-independent heat kernel estimates? Start with manifold with smooth metric, use probability:

Theorem

Let $x, y \in B_{\delta_0 R_z}(z)$ be such that $||x - y|| < \delta_0 R_z$, $\delta_0 < \frac{1}{4}$. Let τ_n's be the return times in $B_{\frac{3}{2}\delta_0 R_z}(x)$ after exiting $B_{2\delta_0 R_z}(x)$, and $x_n(\omega) = \omega(\tau_n(\omega))$. Then

$$K_s(x,y) = K_s^{Dir(B_{2\delta_0 R_z}(x))}(x,y) + \sum_{n=1}^{+\infty} \mathbb{E}_\omega \left[K_{s-\tau_n(\omega)}^{Dir(B_{2\delta_0 R_z}(x))}(x_n(\omega), y) \chi_{\{\tau_n(\omega) < s\}}(\omega) \right] P(\tau_n < s) \,. \quad (1)$$

Moreover there exists an $M = M(c_{max})$ such that $P(\tau_n < s) \lesssim_{d, M, c_{\min}, c_{\max}} e^{-n \frac{\left(\frac{\delta_0 R_z}{2}\right)^2}{2Ms}}$.

Then take limits of smooth metrics to the $\mathcal{C}^\alpha$ metric. Pretty easy for the heat kernel, some tricks (time-stopping arguments) for the gradient,

To be done:

- Generalize to “manifolds” with varying dimensionality, graphs (what is the boundary, inradius?), data with noise.
- Efficient algorithms, in particular we are working on the heat triangulation theorem;

Fast tour through the rest of the story..

Fourier analysis on data: use eigenfunctions for function approximation.
Fourier summability kernels: in analogy with summability kernels in Euclidean spaces (or the sphere), such kernels can be constructed on rather general metric spaces, modeling data, and yield multiscale approximation schemes with better approximation properties for functions with non-homogeneous smoothness (joint with H.N. Mhaskar).
Wavelets: multiscale wavelets can be constructed on data sets by using the diffusion operator and its power. Original construction is one year old, and novel constructions with better approximation properties, localization and faster algorithms are being developed.
The *diffusion semigroup* itself on the data can be used as a smoothing kernel. We recently obtained very promising results in image denoising and semisupervised learning.

Applications

- Hierarchial organization of data and of Markov chains (e.g. documents, regions of state space of dynamical systems, etc...);
- Distributed agent control, Markov decision processes (e.g.: compression of state space and space of relevant value functions);
- Machine Learning (e.g. nonlinear feature selection, semisupervised learning through diffusion, multiscale graphical models);
- Approximation, learning and denoising of functions on graphs (e.g.: machine learning, regression, etc...)
- Sensor networks: compression of measurements collected from the network (e.g. wavelet compression on scattered sensors);
- Multiscale modeling of dynamical systems (e.g.: nonlinear and multiscale PODs);
- Compressing data and functions on the data;
- Data representation, visualization, interaction;
- ...

Nonlinear image denoising, I

Left to right: 1) a clean image, with range from 0 to 255. 2) A noisy image obtained by adding Gaussian noise $40\mathcal{N}(0, 1)$. 3) TV denoising kindly provided by Guy Gilboa. 4) Denoising using a diffusion built on the graph of 5×5 patches, with a constrained search.

Nonlinear image denoising, II

1) Lena with Gaussian noise added. 2) Denoising using a 7x7 patch graph. 3) Denoising using hard thresholding of curvelet coefficients. The image is a sum over 9 denoisings with different grid shifts. 4) Denoising with a diffusion built from the 9 curvelet denoisings.

Nonlinear image denoising, III

Semi-supervised Learning on Graphs

[Joint with R.R. Coifman and A.D.Szlam]

Given: many data points with similarity function, yielding a graph G, of which only a very small subset $\tilde{G}$, are labeled. We use diffusion process to smooth the label functions from $\tilde{G}$ to functions on G. Each point has now a vector of probabilities of belonging to different classes: use this extra information to design a better, anisotropic diffusion on G, and start anew by applying this to the initial labels. Motivations: the diffusion process is a very flexible tool, it is easy to tune time-scales, it is easily tuned to incorporate labeling information, it is very fast to compute.

Experiments on standard data sets show this technique outperforms the previous semi-supervised learning algorithms.

Semi-supervised Learning on Graph (cont'd)

[Joint with R.R. Coifman and A.D.Szlam]

	FAKS	FAHC	FAEF	Best of other methods
digit1	2.0	2.1	**1.9**	2.5 (LapEig)
USPS	4.0	3.9	**3.3**	4.7 (LapRLS, Disc. Reg.)
BCI	45.5	45.3	47.8	**31.4** (LapRLS)
g241c	19.8	21.5	**18.0**	22.0 (NoSub)
COIL	12.0	11.1	15.1	**9.6** (Disc. Reg.)
gc241n	11.0	12.0	9.2	**5.0** (ClusterKernel)
text	**22.3**	**22.3**	22.8	23.6 (LapSVM)

In the first column we chose, for each data set, the best performing method with model selection, among all those discussed in Chapelle's book. In each of the remaining columns we report the performance of each of our methods with model selection, but with the best settings of parameters for constructing the nearest neighbor graph, among those considered in other tables. The aim of this rather unfair comparison is to highlight the potential of the methods on the different data sets.

Acknowledgements

- R.R. Coifman, [Diffusion geometry; Diffusion wavelets; Uniformization via eigenfunctions; Multiscale Data Analysis], P.W. Jones (Yale Math), S.W. Zucker (Yale CS) [Diffusion geometry];
- G.L. Davis (Yale Pathology), R.R. Coifman, F.J. Warner (Yale Math), F.B. Geshwind , A. Coppi, R. DeVerse (Plain Sight Systems) [Hyperspectral Pathology];
- S. Mahadevan (U.Mass CS) [Markov decision processes];
- R. Schul (UCLA), P.W. Jones (Yale Math) [Uniformization via eigenfunctions; nonhomogenous Brownian motion];
- A.D. Szlam (Yale) [Diffusion wavelet packets, top-bottom multiscale analysis, linear and nonlinear image denoising, classification algorithms based on diffusion];
- H. Mhaskar (Cal State, LA) [polynomial frames of diffusion wavelets];
- J.C. Bremer (Yale) [Diffusion wavelet packets, biorthogonal diffusion wavelets];
- M. Mahoney, P. Drineas (Yahoo Research) [Randomized algorithms for hyper-spectral imaging]
- J. Mattingly, S. Mukherjee and Q. Wu (Duke Math,Stat,ISDS) [stochastic systems and learning]; A. Lin, E. Monson (Duke Physics) [Neuron-glia cell modeling]; D. Brady, R. Willett (Duke Engineering) [Compressed sensing and imaging]

Funding: NSF, ONR.

`www.math.duke.edu/~mauro`

Thank you!