

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

Mauro Maggioni

Mathematics and Computer Science
Duke University

IPAM 2007, 10/03/07

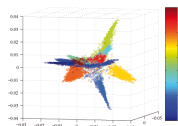
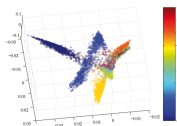
In collaboration with R.R. Coifman, P.W. Jones, R. Schul, A.D. Szlam,
S. Mahadevan...

Funding: NSF-DMS, ONR.

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

Handwritten Digits

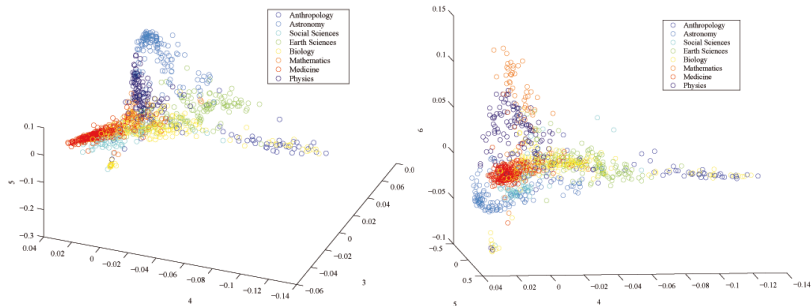
Data base of about 60,000 28×28 gray-scale pictures of handwritten digits, collected by USPS. Goal: automatic recognition. It is a point cloud in 28^2 dimensions. We can think of being given this cloud, and some points are labeled by the digit they correspond to, and we would like to predict the digit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



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

Text documents

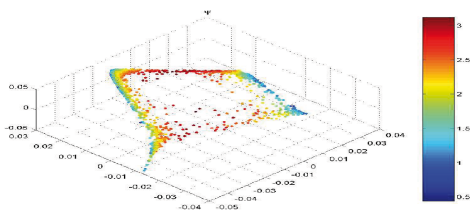
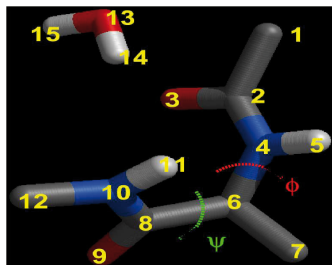
1000 Science News articles, from 8 different categories. We compute about 10000 coordinates, i -th coordinate of document d represents frequency in document d of the i -th word in a fixed dictionary.



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} .

- 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.

- 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

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

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

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

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) \rightarrow eigenfunctions of $T \rightarrow$ Fourier analysis *on* the data, “basis method”
- Look at T for small time (T, T^2, \dots, T^k, k constant) \rightarrow it is diffusion *on* the set \rightarrow “PDE” method, “no basis”
- Look at T at all time scales ($T, T^2, T^4, \dots, T^{2^j}, \dots$) \rightarrow multiscale analysis of both functions and the diffusion process \rightarrow 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) \rightarrow eigenfunctions of $T \rightarrow$ Fourier analysis *on* the data, “basis method”
- Look at T for small time (T, T^2, \dots, T^k, k constant) \rightarrow it is diffusion *on* the set \rightarrow “PDE” method, “no basis”
- Look at T at all time scales ($T, T^2, T^4, \dots, T^{2^j}, \dots$) \rightarrow multiscale analysis of both functions and the diffusion process \rightarrow 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) \rightarrow eigenfunctions of $T \rightarrow$ Fourier analysis *on* the data, “basis method”
- Look at T for small time (T, T^2, \dots, T^k, k constant) \rightarrow it is diffusion *on* the set \rightarrow “PDE” method, “no basis”
- Look at T at all time scales ($T, T^2, T^4, \dots, T^{2^j}, \dots$) \rightarrow multiscale analysis of both functions and the diffusion process \rightarrow 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. 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 be the isometric image of a Euclidean domain. Also, Bérard, Besson and Gallot ('84,'94) use all the eigenfunctions to embed into ℓ^2 .

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 $0 \leq \lambda_0 \leq \dots \leq \lambda_j \leq \dots$. 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

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 $c_1, \dots, c_6 > 0$ that depend only on d and $C_{\text{Weyl}, \Omega}$, such that the following hold. For any $z \in \Omega$, let $R_z \leq \text{dist}(z, \partial\Omega)$. Then there exist i_1, \dots, i_d and constants

$c_6 R_z^{\frac{d}{2}} \leq \gamma_1 = \gamma_1(z), \dots, \gamma_d = \gamma_d(z) \leq 1$ such that:

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

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

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

Eigenfunction Embedding theorems, III

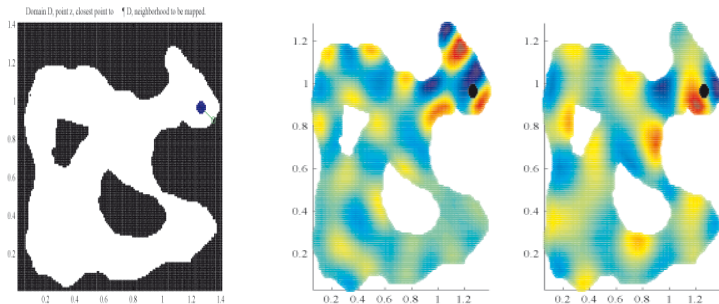


Figure: Top left: a non-simply connected domain in \mathbb{R}^2 , and the point z with its neighborhood to be mapped. Top right: the image of the neighborhood under the map. Bottom: Two eigenfunctions for mapping.

Eigenfunction Embedding theorems, IV

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}^α for some $\alpha > 0$. For any $z_0 \in \mathcal{M}$, let (U, x) be a coordinate chart such that $z_0 \in U$, $g^{ij}(x(z_0)) = \delta^{ij}$ and for any $w \in U$, and any

$$\xi, \nu \in \mathbb{R}^d, c_{\min}(g) \|\xi\|_{\mathbb{R}^d}^2 \leq \sum_{i,j=1}^d g^{ij}(x(w)) \xi_i \xi_j,$$

$$\sum_{i,j=1}^d g^{ij}(x(w)) \xi_i \nu_j \leq c_{\max}(g) \|\xi\|_{\mathbb{R}^d} \|\nu\|_{\mathbb{R}^d}.$$

We let $r_{\mathcal{M}}(z_0) = \sup\{r > 0 : B_r(x(z_0)) \subseteq x(U)\}$.

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

Eigenfunction Embedding theorems, IV

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 $c_1, \dots, c_6 > 0$, depending on $d, c_{\min}, c_{\max}, \|g\|_{\alpha \wedge 1}, \alpha \wedge 1$, and $C_{\text{Weyl}, \Omega}$, such that the following hold. Let $R_z = r_{\mathcal{M}}(z)$. Then there exist i_1, \dots, i_d and constants $c_6 R_z^{\frac{d}{2}} \leq \gamma_1 = \gamma_1(z), \dots, \gamma_d = \gamma_d(z) \leq 1$ such that:

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

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

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

Heat triangulation theorem, V

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, \dots, p_d be d linearly independent directions. There are constants $c_1, \dots, c_5 > 0$, depending on $d, c_{\min}, c_{\max}, \|g\|_{\alpha \wedge 1}, \alpha \wedge 1$, and the smallest and largest eigenvalues of the Gramian matrix $(\langle p_i, p_j \rangle)_{i=1, \dots, d}$, 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 $i = 1, \dots, d$ and let $t_z = c_6 R_z^2$. The map

$$\begin{aligned} \Phi : B_{c_1 R_z}(z) &\rightarrow \mathbb{R}^d \\ x &\mapsto (R_z^d K_{t_z}(x, y_1)), \dots, R_z^d K_{t_z}(x, y_d) \end{aligned}$$

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 \rightarrow pigeon-hole \rightarrow find eigenfunction with gradient of the correct size in a given direction \rightarrow 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

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

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

Then take limits of smooth metrics to the C^α metric. Pretty easy for the heat kernel, some tricks (time-stopping arguments) for the gradient,

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

- Hierarchical 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;
- ...

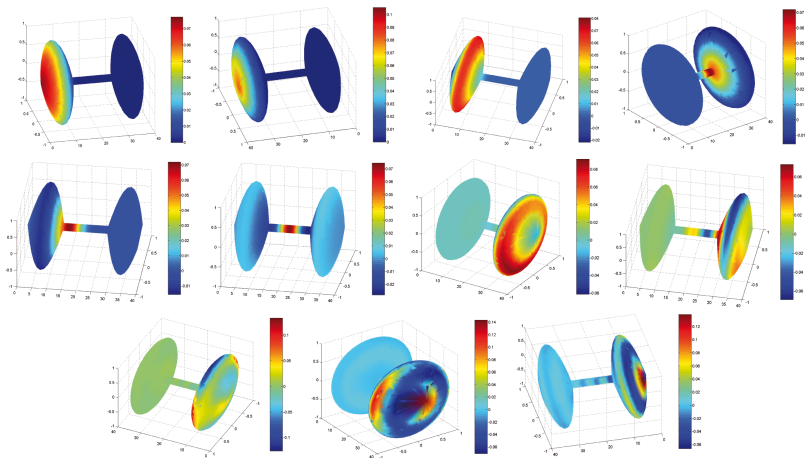
Multiscale Analysis

We construct multiscale analyses associated with a diffusion-like process T on a space X , be it a manifold, a graph, or a point cloud. This gives:

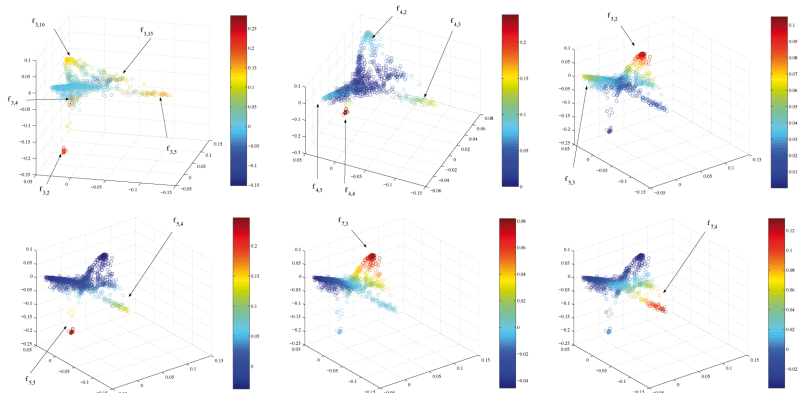
- (i) A coarsening of X at different “geometric” scales, in a chain $X \rightarrow X_1 \rightarrow X_2 \rightarrow \dots \rightarrow X_j \dots$;
- (ii) A coarsening (or compression) of the process T at all time scales $t_j = 2^j$, $\{T_j = [T^{2^j}]_{\Phi_j}^{\Phi_j}\}_j$, each acting on the corresponding X_j ;
- (iii) A set of wavelet-like basis functions for analysis of functions (observables) on the manifold/graph/point cloud/set of states of the system.

All the above come *with guarantees*: the coarsened system X_j and coarsened process T_j behave ϵ -closely as T^{2^j} on X . This comes at the cost of a very careful coarsening: up to $\mathcal{O}(|X|^2)$ operations ($< \mathcal{O}(|X|^3)$!), and only $\mathcal{O}(|X|)$ in certain special classes of problems.

Diffusion Wavelets on Dumbbell manifold



Example: Multiscale text document organization



Scaling functions at different scales represented on the set embedded in \mathbb{R}^3 via $(\xi_3(x), \xi_4(x), \xi_5(x))$. $\phi_{3,4}$ is about Mathematics, but in particular applications to networks, encryption and number theory; $\phi_{3,10}$ is about Astronomy, but in particular papers in X-ray cosmology, black holes, galaxies; $\phi_{3,15}$ is about Earth Sciences, but in particular earthquakes; $\phi_{3,5}$ is about Biology and Anthropology, but in particular about dinosaurs; $\phi_{3,2}$ is about Science and talent awards, inventions and science competitions.

Doc/Word multiscales

Scaling Fcn	Document Titles	Words
$\varphi_{2,3}$	Acid rain and agricultural pollution Nitrogen's Increasing Impact in agriculture	nitrogen,plant, ecologist,carbon, global
$\varphi_{3,3}$	Racing the Waves Seismologists catch quakes Tsunami! At Lake Tahoe? How a middling quake made a giant tsunami Waves of Death Seabed slide blamed for deadly tsunami Earthquakes: The deadly side of geometry	earthquake,wave, fault,quake, tsunami
$\varphi_{3,5}$	Hunting Prehistoric Hurricanes Extreme weather: Massive hurricanes Clearing the Air About Turbulence New map defines nation's twister risk Southern twisters Oklahoma Tornado Sets Wind Record	tornado,storm, wind,tornadoe, speed

Some example of scaling functions on the documents, with some of the documents in their support, and some of the words most frequent in the documents.

Application to Markov Decision Processes

[joint with S. Mahadevan]

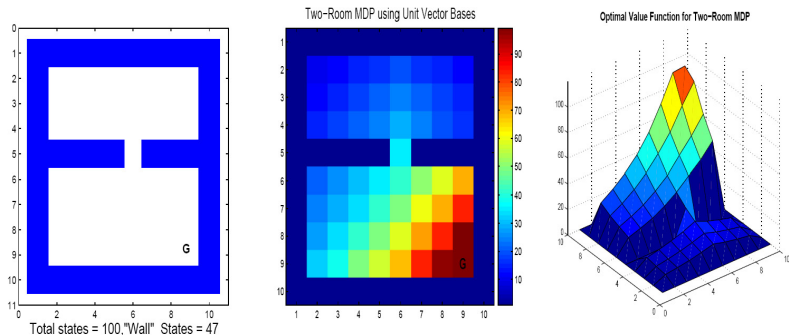
A finite Markov decision process (MDP) $M = (S, A, P_{ss'}^a, R_{ss'}^a)$ is defined as a finite set of states S , a finite set of actions A , a transition model $P_{ss'}^a$ specifying the distribution over future states s' when an action a is performed in state s , and a corresponding reward model $R_{ss'}^a$ specifying a scalar cost or reward. A state value function is a mapping $S \rightarrow \mathcal{R}$ or equivalently a vector in $\mathcal{R}^{|S|}$. Given a policy $\pi : S \rightarrow A$ mapping states to actions, its corresponding value function V^π specifies the expected long-term discounted sum of rewards received by the agent in any given state s when actions are chosen using the policy. Any optimal policy π^* defines the same unique optimal value function V^* which satisfies the nonlinear constraints

$$V^*(s) = \max_a \sum_{s'} P_{ss'}^a (R_{ss'}^a + \gamma V^*(s'))$$

The state spaces of MDPs are often varifolds or graphs; it is crucial to represent certain functions and operators (large-time expectation operators \sim Green's operators) efficiently.

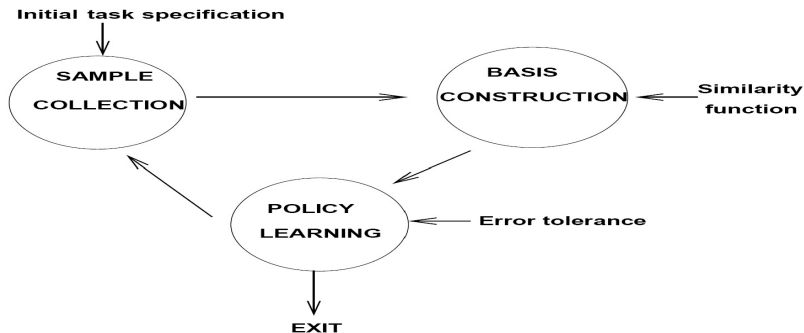


Rooms with bottleneck



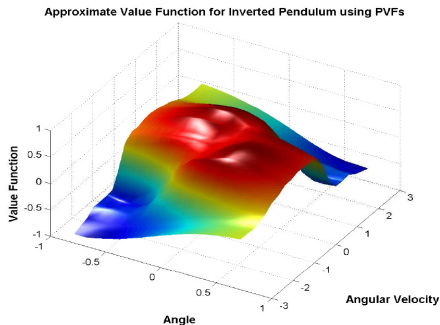
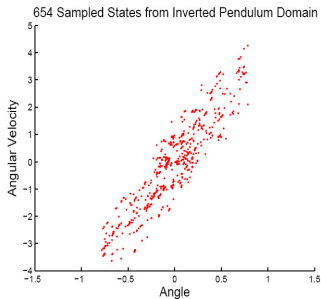
Left: Two rooms connected only by a door, the Goal is in a corner.
Middle and right: corresponding optimal value function.

Representation Policy Iteration



Policy learning could be done with any of the standard algorithms. In the examples that follow we will always use Policy Iteration: the current policy is evaluated (by solving Bellman's equation), it is updated greedily, by choosing at every location the action that leads to a state with higher expected return, and so on iteratively.

Inverted Pendulum

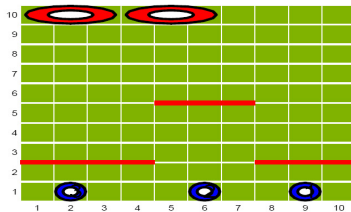
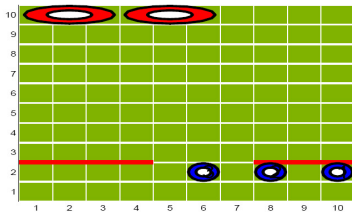


Left: samples collected from several trials with a random policy;
right: value function obtained with RPI.

Blockers

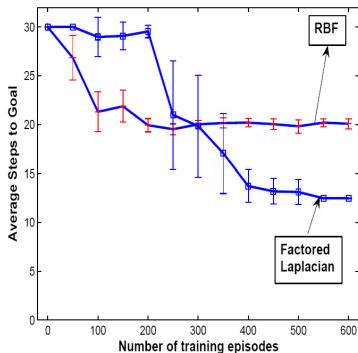
[First proposed by Sallans and Hinton (2004)]

This task is a cooperative multiagent problem where a group of blue agents try to reach the top row of a grid, but are prevented in doing so by “blocker” red agents who move horizontally on the top row. If any agent reaches the top row, the entire team is rewarded by +1; otherwise, each agent receives a negative reward of -1 on each step. The agents always start randomly placed on the bottom row of the grid, and the blockers are randomly placed on the top row. The blockers remain restricted to the top row, executing a fixed strategy.

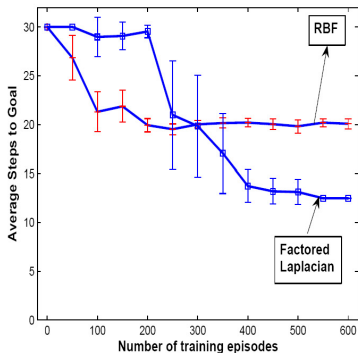


Blocker performance

Results on 10x10 Blocker Domain with Middle and Side Walls



Results on 10x10 Blocker Domain with Middle and Side Walls



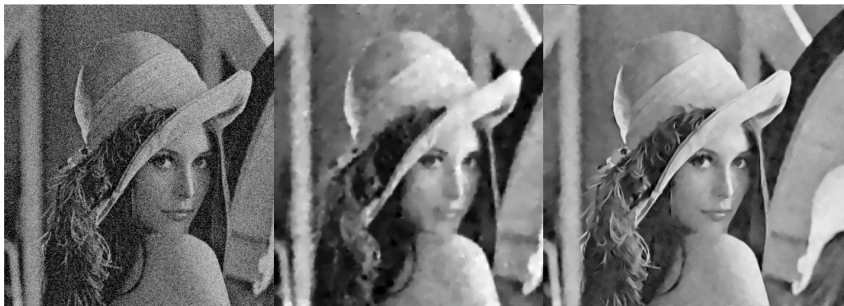
An image is sometimes modeled as a function on $[0, 1]^2$ (or $[0, 1]^3$ if hyperspectral). Well understood that spatial relationships are important, but not the only thing: there are interesting features (e.g. edges, textures), at different scales. Naive idea: apply heat propagation on $[0, 1]$ to the image: denoise by convolution with Gaussian: noise goes away, sharp features as well!

Better idea: do not use simple linear smoothing, but anisotropic/nonlinear smoothing, in order to preserve important structures (mainly edges). Process is image-dependent!

We propose: do not work on $[0, 1]^2$, but in a space of features of the image. Map

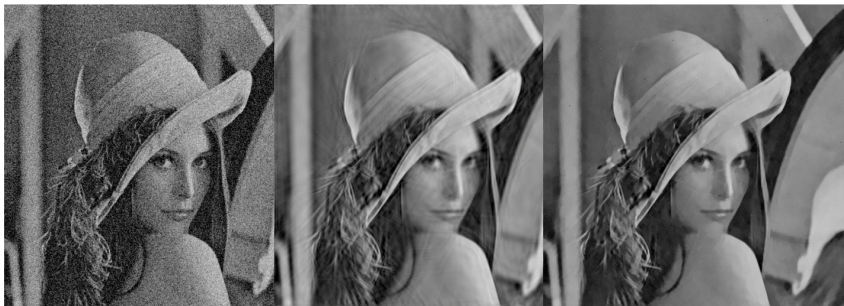
$\Psi(x, y) \in Q \rightarrow (x, y, (I * g_1)(x, y), \dots, (I * g_m)(x, y)) \subset \mathbb{R}^{m+2}$,
and denoise I as a function on $\Psi(Q)$, with the heat kernel on $\Psi(Q)$.

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 7×7 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



Left to right: 1) Barbara corrupted by Gaussian noise $40\mathcal{N}(0, 1)$, from 0 to 255. 2) Denoising using a diffusion built on the graph of 7×7 patches, with a constrained search.

Nonlinear image denoising, IV



Figure: Poisson corrupted image (left) denoised using the patch graph (right).

Semi-supervised Learning on Graphs

[Joint with 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. Let $\chi_i(x) = 1$ if x belongs to class i , 0 otherwise. i ranges of the number of classes. Observe $\chi_i = 0$ outside \tilde{G} .

We use diffusion process built from the geometry of G to smooth χ_i , from \tilde{G} to functions $\hat{\chi}_i$ on G . Now for each point x we have the original coordinates, as well as the vector $(\hat{\chi}_i(x))_i$: we combine this information to construct a new diffusion on G , and start anew by applying this to the initial labels.

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

Semi-supervised Learning on Graph (cont'd)

	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. The training set is 1/15 of the whole set.

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];
- P.W. Jones (Yale Math), R. Schul (UCLA) [Uniformization via eigenfunctions; nonhomogenous Brownian motion];
- S. Mahadevan (U.Mass CS) [Markov decision processes];
- A.D. Szlam (UCLA) [Diffusion wavelet packets, top-bottom multiscale analysis, linear and nonlinear image denoising, classification algorithms based on diffusion];
- 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];
- 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 Phys.) [Neuron-glia cell modeling]; D. Brady, R. Willett (Duke EE) [Compressed sensing and imaging]

Funding: NSF, ONR.

Thank you!

www.math.duke.edu/~mauro

Multiscale Analysis, I

[joint with R.R. Coifman]

We would like to construct multiscale bases, generalizing classical wavelets, on manifolds, graphs, point clouds. Here the space is in general highly *non-symmetric*, not invariant under "natural" geometric transformation, and moreover it is "*noisy*".

Idea: use *diffusion and the heat kernel as dilations*, acting on functions on the space, to generate multiple scales.

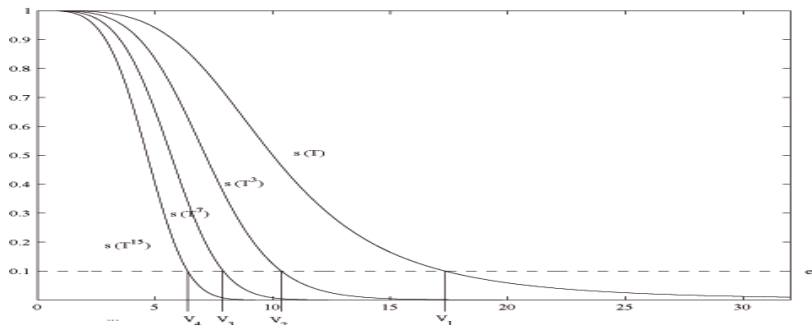
This is connected with the work on diffusion or Markov semigroups, and Littlewood-Paley theory of such semigroups (à la Stein).

We would like to have *constructive* methods for efficiently computing the multiscale decompositions and the wavelet bases.

Multiscale Analysis, II

Suppose for simplicity we have a weighted graph (G, E, W) , with corresponding Laplacian \mathcal{L} and random walk P . Let us renormalize, if necessary, P so it has norm 1 as an operator on L^2 : let T be this operator. Assume T is self-adjoint, and high powers of T are low-rank (T is a diffusion).

A “typical” spectrum for the powers of T would look like this:



Classical Multi-Resolution Analysis

A Multi-Resolution Analysis for $\mathbb{L}^2(\mathbb{R})$ is a sequence of subspaces $\{V_j\}_{j \in \mathbb{Z}}$, $V_{j-1} \subseteq V_j$, with $\bigcap V_j = \{0\}$, $\overline{\bigcup V_j} = \mathbb{L}^2(\mathbb{R})$, with an orthonormal basis $\{\varphi_{j,k}\}_{k \in \mathbb{Z}} := \{2^{\frac{j}{2}} \varphi(2^j \cdot -k)\}_{k \in \mathbb{Z}}$ for V_j . Then there exist ψ such that $\{\psi_{j,k}\}_{k \in \mathbb{Z}} := \{2^{\frac{j}{2}} \psi(2^j \cdot -k)\}_{k \in \mathbb{Z}}$ spans W_j , the orthogonal complement of V_{j-1} in V_j .

$\hat{\varphi}_{j,k}$ is essentially supported in $\{|\xi| \leq 2^j\}$, and $\hat{\psi}_{j,k}$ is essentially supported in the L.P.-annulus $2^{j-1} \leq |\xi| \leq 2^j$.

Because $V_{j-1} \subseteq V_j$, $\varphi_{j-1,0} = \sum_{k'} \alpha_{k'} \varphi_{j,k'}$: refinement eqn.s, FWT.

We would like to generalize this construction to graphs. The frequency domain is the spectrum of $e^{-\mathcal{L}}$. Let

$V_j := \langle \{\phi_i : \lambda_i^{2^j} \geq \epsilon\} \rangle$. Would like o.n. basis of well-localized functions for V_j , and to derive refinement equations and downsampling rules in this context.

Construction of Diffusion Wavelets

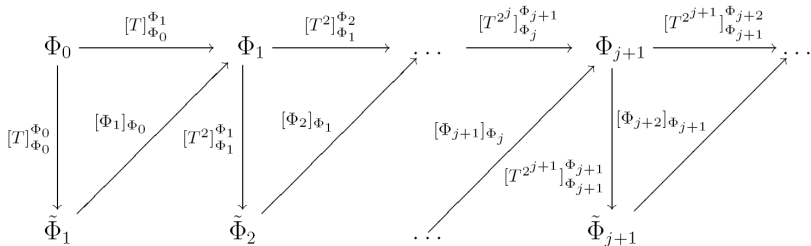


Diagram for downsampling, orthogonalization and operator compression. (All triangles are ϵ -commutative by construction)

Properties of Diffusion Wavelets

- Compact support and estimates on support sizes (not as good as one really would like!);
- Vanishing moments (w.r.t. low-frequency eigenfunctions);
- Bounds on the sizes of the approximation spaces (depend on the spectrum of T , which in turn depends on geometry);
- Approximation and stability guarantees of the construction (tested in practice).

One can also construct diffusion wavelet packets, and therefore quickly-searchable libraries of waveforms.