Diffusion processes on graphs, analysis of high-dimensional data, applications to imaging

Mauro Maggioni
Department of Mathematics
and Computer Science
Duke University

GDA seminar, October 11th, 2006
Overview

1 - Diffusion & analysis on data sets;
2 - Multiscale analysis: bottom-up constructions;
3 - Applications to Semi-Supervised Learning and image denoising;
4 - Applications to Markov Decision Processes;
5 - Open problems, current and future research.
Parametrizations and functions on data sets

A deluge of data: documents, web searching, customer databases, hyper-spectral imagery (satellite, biomedical, etc...), social networks, gene arrays, proteomics data, neurobiological signals, sensor networks, financial transactions, traffic statistics (automobilistic, computer networks)...

Common feature in many of these applications: data is given in “high-dimensional space”, however it has “its own geometry” that is much lower dimensional. It is interesting to: discover and characterize intrinsic properties, such as local dimensionality, local parametrizations. Moreover, in many applications one needs to study functions on the data, and perform tasks such as approximation, smoothing, interpolation, extension to new data. One needs good basis functions on the data, with efficient algorithms for performing these operations.

Regression “on the set”
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 $\mathbb{R}^{36}$, since we have 3 coordinates for each of the 12 atoms. This set is a priori very complicated. However we expect for physical reasons that the constraints on the molecule to force this set to be essentially lower-dimensional. We can explore the space of states by running long simulations, for different initial conditions.

The alanine molecule

The alanine molecule
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.
The Heat Kernel and the Laplacian on Manifolds

Starting point: the heat kernel and diffusion(s). Think of a random walk \( T \) on the data as a way of exploring it.

Ingredient needed: probability of jumping from each point to a neighboring point (e.g. from one document to a similar one, from one state of the molecule to a close-by state, etc...).

Different ways of using this diffusion process:

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

Eigenfunctions

Spectral decomposition:

\[ \Delta \phi_i = \lambda_i \phi_i, \quad H_t(x, y) := e^{-t\Delta}(x, y) = \sum_i e^{-t\lambda_i} \phi_i(x) \phi_i(y). \]

The eigenfunctions \( \phi_i \) of the Laplacian generalize Fourier modes: Fourier analysis on manifolds, global analysis.

Eigenfunctions on a dumbell-shaped manifold, and corresponding diffusion map; pictures courtesy of Stephane Lafon.
Rougher worlds: graph associated with data sets

A deluge of data: documents, web searching, customer databases, hyper-spectral imagery (satellite, biomedical, etc...), social networks, gene arrays, proteomics data, financial transactions, traffic statistics (automobilistic, computer networks)...

Assume we know how to assign local similarities: map data set to weighted graph. Global distances are not to be trusted!

Data often given as points in high-dimension, but constraints (natural, physical...) force it to be intrinsically low-dimensional.

Model the data as a weighted graph $\langle G, E, W \rangle$: vertices represent data points (correspondence could be stochastic), edges connect similar data points, weights represent a similarity measure. Example: have an edge between web pages connected by a link; or between documents with very similar word frequencies.

When points are in high-dimensional Euclidean space, weights may be a function of Euclidean distance, and/or the geometry of the points. How to define the similarity between very similar objects in each category is important but not always easy. That’s the place where field-knowledge goes.
Laplacian on Graphs

Given a weighted graph \((G, W, E)\), the combinatorial Laplacian is defined by 
\[ L = D - W, \]
where \((D)_{ii} = \sum_j W_{ij}\), and the normalized Laplacian is defined by
\[ \mathcal{L} = D^{-\frac{1}{2}} (D - W) D^{-\frac{1}{2}}. \]

These are self-adjoint positive-semi-definite operators, let \(\lambda_i\) and \(\phi_i\) be the eigenvalues and eigenvectors. Fourier analysis on graphs. The heat kernel is of course defined by \(H_t = e^{-t\mathcal{L}}\); the natural random walk is \(D^{-1}W\).

\[ d_{\text{geod.}}(A, B) \sim d_{\text{geod.}}(C, B), \text{ however } d^{(t)}(A, B) \gg d^{(t)}(C, B). \]
Geometrization of Diffusion

Diffusion distance at time $t$ is defined by

$$d^{(t)}(x, y) = \| H_{t/2}(x, \cdot) - H_{t/2}(y, \cdot) \|_{L^2(\mathcal{M})}$$

$$= \sqrt{\langle \delta_x - \delta_y, H^t(\delta_x - \delta_y) \rangle}$$

$$= \sqrt{\sum_i \mu_i^t (\phi_i(x) - \phi_i(y))^2}$$

Surprisingly, the eigenfunctions of the Laplacian also allow to analyze the geometry of the manifold, and provide embeddings of the manifold ("diffusion maps"): for $m = 1, 2, \ldots, t > 0$ and $x \in \mathcal{M}$, define

$$\Phi^{(t)}_m(x) = (\mu_i^{\frac{t}{n}} \phi_i(x))_{i=1,\ldots,m} \in \mathbb{R}^m.$$ 

This map is an approximate isometry (it is an isometry for $m = +\infty$) to Euclidean $\mathbb{R}^m$ from $\mathcal{M}$ with the diffusion metric (not the Riemannian metric!).
Show the “3D” example!
Show the circle+shading example!
MD example (cont’d).
We are able to discover the lower-dimensional set of configurations, parametrize it, estimate its local dimensionality (it is not constant!), consider natural classes of (diffusion) operators on this set, and build Fourier and wavelet bases on it. For example it turns out that some of the parameters discovered by chemists on the basis of experiments and chemical-physical considerations can be discovered empirically as being Fourier-like functions on the set of states!
Several Data Sets Explored

Several applications of spectral kernel methods. For Laplacian eigenfunctions, the following works in particular:

- Regression and classification in the supervised and semi-supervised learning context [M. Belkin, P. Nyogi; RR Coifman, MM, A.D. Slzam]
- fMRI data [F. Meyer, X. Shen]
- Art data [W Goetzmann, PW Jones, MM, J Walden]
- Hyperspectral Imaging in Pathology [MM, GL Davis, F Warner, F. Geshwind, A Coppi, R. DeVerse, RR Coifman]
- Molecular dynamics simulations [RR. Coifman, G.Hummer, I. Kevrekidis, S. Lafon, MM, B. Nadler]
- Text documents classification [RR. Coifman, S. Lafon, A. Lee, B. Nadler; RR Coifman, MM]
**Application to Hyper-spectral Pathology**

For each pixel in a hyper-spectral image we have a whole spectrum (a 128-dimensional vector for example). We view the ensemble of all spectra in a hyper-spectral image as a cloud in $\mathbb{R}^{128}$, induce a Laplacian on the point set and use the eigenfunctions for classification of spectra into different classes, which turn out to be biologically distinct and relevant.

On the left, we have mapped the values of the top 3 eigenfunctions to RGB.
Application to Hyper-spectral Pathology, cont.’d

Semi-supervised classification of tissue types in a dermatology tissue sample. Left: training set, center: classification with nearest neighbors, right: classification by diffusion. Nonlinearity in the structure of the spectra is due to non-uniform lighting and sample preparation conditions: diffusion follows these deformations by “moving on the manifold” of variations.
Examples of nuclei patches classified by the algorithm from three different biopsies. From left to right: a normal, benign (adenoma) and malignant biopsy. The colors have the following meaning: green=classified normal, blue=classified benign (adenoma), red=classified malignant.
Application to text document classification

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. The diffusion map gives the embedding below. Clustering in the range of diffusion map results in good unsupervised performance for document classification.

Embedding \( \Xi_6^{(0)}(x) = (\xi_1(x), \ldots, \xi_6(x)) \): on the left coordinates 3, 4, 5, and on the right coordinates 4, 5, 6.
Summary for the “Fourier part”

- it is useful to consider only local similarities between data points;
- it is possible to organize this local information by diffusion;
- parametrizations can be found by looking at the eigenvectors of a diffusion operator (Fourier modes);
- these eigenvectors yield a nonlinear embedding into low-dimensional Euclidean space;
- the eigenvectors can be used for global Fourier analysis on the set/manifold.

Problem: Either very local information or very global information: in many problems the intermediate scales are very interesting! Would like multiscale information!

Possibility 1: proceed bottom-up: repeatedly cluster together in a multi-scale fashion, in a way that is faithful to the operator: diffusion wavelets.

Possibility 2: proceed top-bottom: cut greedily according to global information, and repeat procedure on the pieces: recursive partitioning, local cosines...

Possibility 3: do both!
A multiscale “network”
Multiscale elements and representation of powers of $T$
Coarsening of Markov chains

We now consider a simple example of a Markov chain on a graph with 8 states.

\[
T = \begin{pmatrix}
0.80 & 0.20 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.20 & 0.79 & 0.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.01 & 0.49 & 0.50 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.50 & 0.499 & 0.001 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.001 & 0.499 & 0.50 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.50 & 0.49 & 0.01 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.01 & 0.49 & 0.50 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.50 & 0.50 \\
\end{pmatrix}
\]

From the matrix it is clear that the states are grouped into four pairs \(\{\nu_1, \nu_2\}\), \(\{\nu_3, \nu_4\}\), \(\{\nu_5, \nu_6\}\), and \(\{\nu_7, \nu_8\}\), with weak interactions between the the pairs.
Some dyadic powers of the Markov chain $T$.

Compressed representations $T_6$, $T_{13}$, and corresponding scaling functions.
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
$$X \rightarrow X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_j \cdots;$$

(ii) A coarsening (or compression) of the process $T$ at all time scales $t_j = 2^j$, $\{T_j\}$, 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, in the sense that the coarsened system $X_j$ and corresponding coarsened process $T_j$ behave exactly as $T^{2^j}$ on $X$. The guarantee come of course at the cost of a very careful coarsening procedure. In general it can take up to $O(|X|^2)$ operations, and only $O(|X|)$ in certain classes of problems (e.g. diffusion on nice manifolds).
Thinking multiscale on graphs...

Investigating other constructions:

- Biorthogonal diffusion wavelets, in which scaling functions are probability densities (useful for multiscale Markov chains)
- Top-bottom constructions: recursive subdivision
- Both...

Applications:

- Document organization and classification
- Nonlinear Analysis of Images
- Semi-supervised learning through diffusion processes on data
- Markov Decision Processes
Application to 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.
Nonlinear image denoising, I

[Joint with A.D. Szlam (Yale→UCLA)]

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! Think of the image as dictating the conductivity, which varies at different locations and different directions. Way better.

Even better: do not work on $[0,1]^2$, but in a space of features of the image. The set of features of an image has image-dependent geometry that dictates preservation of the salient features. Then use the heat propagation dictated by the geometry of the set of features to denoise the image.
Nonlinear image denoising, II

[Joint with A.D. Szlam (Yale→UCLA)]

Some preliminary results
TV (graciously provided by Guy Gilboa); diffusion on $5 \times 5$ patches
Continuation... 1) TV denoising, contrast adjusted manually. 2) Patch graph denoising, contrast adjusted manually.
1) Diffusion made on a graph of 30 realizations of $5 \times 5$ white noise patches; 2) 25 denoisings done with white noise filters as an embedding, and rebuilding the diffusion.
As before, but manually adjusted contrast.
noisy; denoising via diffusion on patches
Top left: image of Barbara, with 4 square $10 \times 10$ pixel regions highlighted. The $5 \times 5$ patches in each region are considered as 25 dimensional vectors, and top right we plot the singular values of their covariance matrix. At the bottom, we project the 25-dimensional points in each region on their top 3 principal components. In region 1, note how the (approximate) periodicity of the texture in region 1 is reflected in the tubular shape of the projection; in region 2, the portions of the image on different sides of the edge are disconnected in the feature space, and note the higher dimensionality, as measured by the singular values; for region 3, note the higher dimensionality (slower decay of the singular values) compared to regions 1 and 4; for region 4 note the very small dimensionality.
Left: image of Lena, with two locations highlighted. Center: row of the diffusion kernel corresponding to the upper-left highlighted area in the image on the left. Right: row of the diffusion kernel corresponding to the bottom-left highlighted area in the image on the left. The diffusion kernel averages according to different patterns in different locations. The averaging pattern on the right is also “non-local”, in the sense that the averaging occurs along well-separated stripes, corresponding to different strikes of hair in the original picture.
Left: image of Barbara, with several locations $p_i$ highlighted. Then from left to right, rows of $K^t(p_i, \cdot)$, for $t = 1, 2, 5$. 
Three eigenfunctions of the diffusion kernel associated with the Lena image, viewed as functions on the square. They capture large-scale cluster structures and could be used for large-scale segmentation.
Run the root script if you have 1.5 minutes!
Towards Nonlinear Analysis of Images

Lift images to graph of features (e.g. patches), and work on the graph: eigenfunctions useful for global segmentation (Shi-Malik), diffusion wavelets for a novel nonlinear multiscale analysis, diffusion processes on graphs associated with the image are excellent for denoising (joint with A.D. Szlam, next slide).

Left: image with added Gaussian noise; some diffusion scaling functions as images.
Semi-supervised Learning on Graphs

[Joint with A.D. Szlam]

Given a graph $G$ and $\chi_1, \ldots, \chi_C$ characteristic functions of sets, representing points labeled according to their class (e.g. document topics, digit in a handwritten digit database, functionality of a protein in a protein network...). These labels are known on a small subset $\tilde{G}$ of $G$, and we would like to guess the labels of the non-labeled points. It is an interpolation or smoothing problem.

So far good results (in theory and in practice) with the use of eigenfunctions [Belkin, Niyogi], even better results by using anisotropic diffusions on the graph to smooth the label functions; just started work with diffusion wavelets for this task and other machine learning tasks.

We are applying these techniques to handwritten digits, documents classification, protein functionality prediction in protein networks.
Semi-supervised Learning on Graphs, cont’d

[Joint with A.D. Szlam]

*Eigenfunctions* [Belkin-Niyogi]: Predict unlabeled points by projecting onto a subspace spanned by low-frequency eigenfunctions, restricted to the labelled set $\tilde{G}$. Motivations & assumptions: the label functions are smooth w.r.t. geometry of space, eigenfunctions capture idea of smoothness with respect to the geometry.

*New nonlinear technique*: 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 much more flexible tool than eigenfunctions, for example it is easy to tune time-scales, it is easily tuned to incorporate labeling information, it has a better spectral properties than a spectral projector (e.g. no Gibbs phenomenon), moreover it is very fast to compute!

Experiments on USPS zip code data set show this technique outperforms the previous semi-supervised learning algorithms. We are applying this technique to a problem in prediction of *protein functionality*, where there are more than 40 classes, with encouraging preliminary results.
Show the two moons slide!
Application to Markov Decision Processes

[S. Mahadevan, MM]

A finite Markov decision process (MDP) $M = (S, A, P_{ss'}, R_{ss'})$ is defined as a finite set of states $S$, a finite set of actions $A$, a transition model $P_{ss'}$ specifying the distribution over future states $s'$ when an action $a$ is performed in state $s$, and a corresponding reward model $R_{ss'}$ 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 $V^\pi$ 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 $\pi^*$ 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.
Inverted Pendulum

Left: $Q$-value function for the action “left”, reconstructed from its representation of the diffusion wavelet basis. Right: trajectory of the pendulum in phase space according to the policy learnt.
Top row: trajectory of angle and angle velocity variables. Bottom row: some diffusion wavelets used as basis functions for representation during the learning phase.
Top left: $Q$-value function for the action “left”, reconstructed from its representation of the diffusion wavelet basis. Top right: trajectory of the mountain car in phase space according to the policy learnt (107 steps). Bottom row: some diffusion wavelets used as basis functions for representation during the learning phase.
Current & Future work

• Improving properties of multiscale analysis (largest class of spaces and processes for which it is $O(N)$?)

• Deriving results for approximation of functions in stochastic settings

• Understanding relationships between multiscale geometry of the set, and smoothness of the functions on it, approximation rates, etc...

• Applications: signal processing on manifolds and graphs and its applications (e.g. linear and nonlinear “image” denoising); classification algorithms (e.g. text classification, protein and gene functional classification, target recognition in hyper-spectral imaging); learning; application to Markov decision processes; multiscale structures of complex networks and dynamical systems.
Collaborators

- R.R. Coifman, P.W. Jones (Yale Math) [Diffusion geometry; Diffusion wavelets; Uniformization via eigenfunctions; Multiscale Data Analysis], 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) [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];
- Y. Kevrekidis (Princeton Eng.), S. Lafon (Google), B. Nadler (Weizman) [stochastic dynamics];
- W. Goetzmann (Yale, Harvard Business School), J. Walden (Berkley Business School), P.W. Jones (Yale Math) [Applications to finance]
- H. Mhaskar (Cal State, LA) [polynomial frames of diffusion wavelets, characterization of function spaces];
- J.C. Bremer (Yale) [Diffusion wavelet packets, biorthogonal diffusion wavelets];
- M. Mahoney (Yahoo Research), F. Meyer (UC Boulder), X. Shen (UC Boulder) [Randomized algorithms for hyper-spectral and fMRI imaging]
- Novel collaborations: J. Mattingly, S. Mukherjee (Duke Math,Stat) [stochastic systems and learning]; A. Lin, E. Monson (Duke Physics) [Study of neuron-glia cell interactions in networks]; D. Brady, R. Willett (Duke Engineering) [Compressed sensing and imaging, sensor networks]


Thank you!