Dimension Reduction, Discretization and Diffusion on Compact Sets

Speaker: S.B. Damelin, GA and IMA, University of Minnesota

University of Leicester, August 2006

This paper and related work can be found on my homepage: http://math.georgiasouthern.edu/∼ damelin

Main references:

- S.B. Damelin and P. Grabner, *Numerical integration, energy and asymptotic equidistribution on the sphere*, Journal of Complexity, 19(2003), pp 231-246.
- S. B. Damelin, T. Devaney and R. Luke,*Paley Wiener Theorems revisited and their applications to inverse scattering problems*, preprint, 2006.
- S. B. Damelin, V. Maymeskul, *On point energies, separation radius, and mesh norm for s-extremal configurations on compact sets in $\mathbb{R}^n$*, Journal of Complexity, Volume 21(6)(2005), pp 845-863.
- Y. Ma, S. B. Damelin, O. Masoud and N. Papanikolopoulos, *Activity Recognition via Classification Constrained Diffusion Maps*, Lecture notes in Computer Science, IEEE 2nd international symposium on visual computing, to appear.

Structure of talk

- Learning data on compact sets: The [DC] and [CD] problems.
- Towards the [DC] problem on compact sets: sensor networks, G invariant kernels and invariance.
- Towards the [CD] problem on rectifiable manifolds: Numerical integration, discretization.
- Discretization on homogenous, reflexive manifolds.
- Riesz configurations, Asymptotic equidistribution for s extremal points.
- Point energies, Mesh norm and Separation for Riesz points on d rectifiable sets.
- Markov Chains on Data, Diffusion Distances and Embeddings onto Spheres.
- Density Invariance on Spheres.

Learning data on compact sets- background and motivation

In this talk, I will be interested in the following two tasks:

Task 1 [DC]: Learning mathematical methods in order to study meaningful descriptions of data sets or a finite number of given discrete objects

Task 2 [CD]: Learning mathematical methods in order to study complicated structure by way of discretization into a finite number of discrete objects

These two questions, as it turns out, are connected mathematically in more ways than you can possibly imagine. Id like to devote this talk to discussing some of my interests in these two tasks.

The [DC] Problem

Nowdays, we are constantly flooded with information of all sorts and forms and a common denominator of data analysis in many emerging fields of current interest are large amounts of observations that have high dimensionality.

EXAMPLE

Suppose that a source produces a high dimensional data set

$$X := \{x_1, ..., x_n\}\,,\, n \geq 1$$

that we wish to analyse. Typically, in a given problem of this kind, one is often faced with a finite data set or a set of finite objects in a measure space with absolutely continuous density μ'.

Examples:

- Each member of X could be the fames of a movie produced by a digital camera.
- Pixels of a hyperspectral image in a computer vision problem, say face recognition.
- Objects from a statistical, computer science or machine learning model which needs to be clustered.

Indeed, the problem of finding mathematical methods in order to study meaningful structures and descriptions of **data sets** for learning tasks is an exploding area of research with applications as diverse as critical infrastructure, complex networks, clustering, imaging neural and sensor networks, wireless communications, financial marketing and dynamic programming

The list is endless and exponentially exploding...........

When dealing with these types of sets X, **high dimensionality** is often an obstacle for any efficient processing of the data. Indeed, many classical data processing algorithms have a computational complexity that grows exponentially with the dimension (the so called "**curse of dimension**").

On the other hand, the source of the data may only enjoy a limited number of degrees of freedom. In this case, the high dimensional representation of the data is a natural function of the source and so the data has actually a low intrinsic dimensionalty.

That is, there is a high (local or global) correlation between many of the variables that describe the members of X.

In addition,

Challenges

(1) The members of X may be nonlinear and so there is a need to learn these nonlinearities.

(2) The members of X may be sampled from a fixed source M but at different rates and so we are interested in the underlying geometry of the source M and not on the the distribution of the points.

(3) The data may be noisy.

Towards the [DC] problem on compact sets: sensor networks, G invariant kernels and invariance.

The first task performed by any data processing system is data acquisition or sampling in which measurements are collected through a number of sensors. Sensor networks are dense, wireless networks of small, low cost sensors nodes which collect and disseminate environmental data locally.

In particular, wireless sensors as deployed frequently by the US army using UAV's on dense terrain on the **surface of the earth** may have a large spectrum on intensity values.

Let us formalize this idea mathematically as follows:

Defn I will call a source M **admissible**, if M is a $d \geq 1$ dimensional, homogeneous space of a compact, reflexive Lie group G embeddded in some Euclidean space of fixed dimension $d + r$ for some $r > 0$.

Typically, M is an orbit of a compact group G. Our investigation serves two main purposes; to uncover the essential geometry involved and to provide a wider range of interdisciplinary applications.

A natural example to keep in mind is S^d, the d dimensional sphere realized as a subset of $\mathbb{R}^{d+1}$ which is the orbit of any unit vector under the action of $SO(d+1)$, the group of $d+1$ dimensional orthogonal matrices of determinant 1.

Suppose we have also a kernel $k : M \times M \to [0, \infty)$ which is positive definite and G invariant ($k(gx, gy) = k(x, y), x, y \in M, g \in G$).

k will relate members of M.

Theorem [**DLS**] Let M be an admissible source. Then k is zonal, ie it depends only on **distances** between points in $\mathbb{R}^{d+r}$.

We call k a G invariant kernel.

Examples on M

(1) Freud/heat kernels[FK], $k(x, y) = \exp(-||x-y||^s)$, $s > 0$

(2) Riesz kernels[RK], $k(x, y) = \frac{1}{||x-y||^s}$, $x \neq y$, $s > 0$

(1) Vision, bioinformatics; (2) Electrostatic problems.

Question Given an admissible kernel k, does it have the property that one can separate the distribution of data points X with given density from the geometry of the underlying manifold M from which X is sampled.

Yes for [RK] on S^d but...............

Towards the [CD] problem on rectifiable sets

Distribution of points on Spheres: A Motivational Journey

The problem of uniformly distributing points on spheres is an interesting and difficult problem.

In one dimension the problem is easily reduced to uniformly distributing $n \geq 1$ points on a circle and equidistant points or the vertices of the regular n-gon provide an obvious answer.

Carl Friedrich Gauss (1777-1855): Famous Disqvistiones arithmaticae

For $d \geq 2$ the problem becomes much more difficult; in fact, there are numerous criteria for uniformity, resulting in different optimal configurations.

On the one hand it is of some interest on its own to describe a “well distributed” point set of cardinality n and even to define suitable notions of what “well distributed” should mean. On the other hand, as we shall show, numerical integration procedures on the sphere require node sets which are spread evenly with respect to separation and mesh norm.

Discretization of manifolds: Riesz Points and Good Distribution

Lets look at $[-1,1]$. How to discretize this set? Well an obvious choice would be N equally spaced points:

$$x^*_{k,n} = -1 + \frac{2k}{n-1},\ n \geq 2,\ k = 0, ..., n-1.$$

These points also enjoy the property of best packing

$$\mathrm{min}_{i\neq j}|x^*_{i,n} - x^*_{j,n}| = \mathrm{max}_{X\subset[-1,1]}\mathrm{min}_{i\neq j}|x_{j,n} - x_{i,n}|$$

- Runge: Equally spaced points or any asymptotically uniformly distributed point set can be problematic for interpolation by polynomials or for quadrature.
- interpolation operators grow geometrically with n
- Zeroes of classical Jacobi polynomials with ±1 do a better job: Norm $O(logn)$

Now what is so special about zeros of Jacobi polynomials versus equidistant points?

Lets move to the sphere.

Errors of Numerical integration

Here and throughout, we will henceforth denote by $< . >$ the usual inner product on $\mathbb{R}^{d+1}$ and μ will denote d dimensional area measure on S^d.

A set of points X_n on S^d, is said to be **asymptotically equidistributed** if for every spherical cap $C \subseteq S^d$,

$$\lim_{n\to\infty} \frac{\#\{1 \leq j \leq n : x_j \in C\}}{n} = \mu(C).$$

i.e., each intersection of the sphere and half space gets an equal portion of points. By duality, it follows that this is equivalent to

$$\lim_{n\to\infty} R_n(f,\mu) = 0$$

for every continuous function f on S^d.

$$|R_n(f,\mu)| := \left| \int_{S^d} f(x)d\mu(x) - \frac{1}{n}\sum_{j=1}^{n} f(x_j) \right|.$$

A natural measure for the quality of the distribution of a **point cloud** X on the sphere S^d is the spherical cap discrepancy

$$D_{n,d} = \sup_{C \subseteq S^d} \left| \frac{1}{n} \sum_{\substack{k=1 \\ x_k \in C}}^{n} \chi_C(x_k) - \mu(C) \right| ,$$

where the supremum ranges over all spherical caps $C \subseteq S^d$ (intersections of balls and S^d) and where χ_C denotes the indicator function of C. The discrepancy simply measures the maximal deviation between the discrete point distribution X and the normalized surface measure.

Extremal configurations

We study numerical integration and discrepancy estimates for configurations X_n on S^d which minimize an energy functional with general kernel k. Important examples of such points are $s \geq 0$ extremal configurations, i.e., points which minimize energies for the Riesz kernel k_R, $|x-y|^{-s}$, $0 < s \leq d$ and logarithmic kernel $-\log|x-y|$, $s=0$.

Set

$$E(k_R, X) = \sum_{\substack{i,j=1 \\ i \neq j}}^{n} k_R(< x_i, x_j >)$$

where $< . >$ denotes inner product in $\mathbb{R}^{d+1}$ and define

$$\mathcal{E}_{k_R}(S^d, n) = \min_{X \in S^d} E_{k_R}(S^d, X).$$

A point set, X_n for which the minimal energy $\mathcal{E}_{k_R}(S^d, n)$ is attained, is called a *k_R-minimal energy point set.* It is clear, that any rotation of a point set of minimal energy again gives a point set of minimal energy; thus such point sets are not unique. Moreover

$$\int_{S^d}\int_{S^d} k_R(< x, y >)d\nu(x)d\nu(y)$$

is minimized by the normalized surface measure μ_d on S^d amoungst all Borel probability measures ν on S^d.

Heuristically, then one expects that a point distribution X_n of minimal g_R energy gives a discrete approximation to the surface measure in the sense that the integral with respect to the surface measure is approximated by a discrete sum over the points.

For the circle, S^1, it is known that minimal energy point sets correspond to the nth roots of unity.

Back to $[-1, 1]$:

- $s = 0$: Counting measures of Fekete polynomials (in this case zeroes of Jacobi polynomials with ± 1) converge weakly to the arcsine distribution which minimizes the energy integral
$$\int \int \log \frac{1}{|x-t|} d\nu(x) d\nu(t)$$
over all Borel probability measures ν supported on $[-1, 1]$.

- $0 < s < 1$: $\frac{c_s}{(1-x^2)^{(1-s)/2}}$: Global effects

- $s \geq 1$: Uniform distribution, $s = \infty$, best packing, e qually spaced points-local effects.

Back to the Sphere S^d:

3 cases to consider: $0 < s < d$, $s = d$, $s > d$:

The measure: d dimensional area measure for all 3 cases.

We expect the counting measures to converge weakly for all 3 cases.

As we move from $s < d$ to $s > d$ we expect the transition from global to local effects to take place.

Discretization of measure: Numerical Integration

Consider:

$$\int_{S^d} f d\mu$$

for continuous $f : \mathbb{R}^{d+1} \to \mathbb{R}$

Approximate by

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

Let

$$\mu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}.$$

Thus we approximate:

$$\int_{S^d} f(d\mu - d\mu_n).$$

Asymptotic equidistribution for $0 < s \leq d$ extremal points

Theorem: DG- Let $d \geq 2$ and $0 < s < d$. Then $0 < s < d$ extremal configurations are asymptotically equidistributed.

This is mainly because the energy integral given by

$$\int\limits_{S^d}\int\limits_{S^d} k_R(<x, y>)\, d\nu(x)\, d\nu(y)$$

is finite in this case with value

$$\frac{\Gamma((d+1)/2)\Gamma(d-s)}{\Gamma(d-s/2))\Gamma(d-s+1)}.$$

For $s \geq d$, the integral diverges for every measure ν which means that the nearest neighbor interactions are dominating.

Theorem DG: k_R energy and discrepancy Let k_R be given, $d \geq 2$, f a polynomial of degree at most $N \geq 1$ on S^d and $0 < \delta \leq \delta_0$. Then

$$|R(f,X)| \leq \|f\|_2 \left(\frac{\frac{1}{n^2} E_{k_R}(S^d, X) - a_0(\delta) + \frac{1}{n} k_R(1-\delta)}{\min_{1 \leq k \leq N} \frac{a_k(\delta)}{Z(d,k)}} \right)^{1/2}$$

with $Z(d,k) = \frac{2k+d-1}{k+d-1}\binom{k+d-1}{d-1}$ counting the linearly independent spherical harmonics of degree k on S^d. Moreover, if $q = q(d)$ is the smallest integer satisfying $2q \geq d+3$, then uniformly for any $m \geq 1$ and $0 < \delta < \delta_0$, there exist positive constants C and C_1 independent of n and X with

$$D_n(X) \leq C \left\{ \frac{C_1}{m} + \left(\frac{\frac{1}{n^2} E_{k_R}(S^d, X) - a_0(\delta) + \frac{1}{n} k_R(1-\delta)}{\min_{1 \leq k \leq N} \frac{a_k(\delta)}{Z(d,k)}} \right)^{1/2} \right\}.$$

Numerical Integration and Zonal Kernels on Compact Homogeneous, Reflexive Sets M

An important motivation for extending the work of Damelin-Grabner relates to being able to discretize (or do quadrature) on tori or other compact sets using energies other than those given by Riesz kernels for example.

We are interested in the error of integration for continuous functions f, when the function is given on a **point cloud** $X \subset M$ of cardinality $n \geq 1$. The error in integration is

$$R(f, X) = \int_M f(y) d\mu(y) - \frac{1}{n} \sum_{x \in X} f(x).$$

Here, M carries a unique normalized *surface* (G-invariant) measure which we call μ.

Harmonic analysis on M requires the construction of polynomials on the manifold. If Π_j is the space of all polynomials of total degree j in the ambient space $\mathbb{R}^{d+r}$ then $P_j := \Pi_j|_M$ is the space of degree j polynomials on M. We can also construct sets of *harmonic polynomials* $H_j := P_j \bigcap P_{j-1}^{\perp}$, where the orthogonality is with respect to the inner product $(\cdot, \cdot)$. We will be interested in zonal kernels whose integrals exist on M. The archetype for such kernels is the *Riesz kernel*

$$k_R(x, y) = \|x - y\|^{-s}, \quad s > 0, \quad x, y \in M,$$

where $\| \cdot \|$ is the Euclidean norm in $\mathbb{R}^{d+r}$. We have:

Theorem DLS For a zonal kernel k and probability measure ν, the energy integral

$$\int_M \int_M k(x,y) d\nu(x) d\nu(y)$$

is uniquely minimised by the normalised surface measure μ on M.

Associated with a discrete measure supported in a finite point set $X \subset M$ we have the discrete energy defined formally as

$$E_\kappa(X) = \frac{1}{n^2} \sum_{y,z \in X,\, y \neq z} \kappa(y,z).$$

This allows us to state:

Theorem DLS1 There exists a sequence of zonal kernels k_α converging to the δ distribution (the distribution for which all Fourier coefficients are unity) as $\alpha \to 0$ and satisfying that for any $\alpha < \alpha_0$ for some fixed α_0, (1) κ_α is positive definite and (2) $\kappa_\alpha(x,y) \leq \kappa(x,y)$ for all $x, y \in M$.

Theorems DLS-DLS1, form the basis for an extension of Theorem DG.

[CD] problem and energies on d rectifiable sets

The following work is due to Damelin and Maymeskul for Riesz kernels k_R.

We say that a set A belongs to the class A^d if, for some $d^{'} \geq d$, $A \subset \mathbb{R}^{d'}$,

(1) $H^d(A) > 0$ and

(2) A is a finite union of bi-Lipschitz images of compact sets in $\mathbb{R}^d$, that is

$$A = \bigcup_{i=1}^{m} \phi\left(K_i\right),$$

where each $K_i \subset \mathbb{R}^d$ is compact and $\phi_i : K_i \to \mathbb{R}^{d'}$ is bi-Lipschitz on K_i, $i = 1, \ldots, m$.

Here and in what follows, $H^d(\cdot)$ denotes d-dimensional Hausdorff measure in $\mathbb{R}^{d'}$.

Examples

(i) Compact sets A in $\mathbb{R}^d$ with $\mathcal{H}^d(A) > 0$: with $d' = d$, these sets are bi-Lipschitz images of themselves under the identity map. (One can also consider these sets embedded in $\mathbb{R}^{d'}$ for $d' > d$.) For example, balls, d-dimensional cubes and parallelepipeds, d-dimensional Cantor sets having positive d dimensional Hausdorf measure.

(ii) d-dimensional spheres in $\mathbb{R}^{d+1}$ (more generally, ellipsoids), since a closed hemisphere is a bi-Lipschitz image of a d-dimensional ball under a stereographic projection.

(iii) Quasismooth (chord-arc) curves in $\mathbb{R}^{d'}$. These are Jordan curves $A \subset \mathbb{R}^{d'}$ satisfying the following condition: there exists a constant C such that, for any two points x, $y \in A$, the length of the (shortest) subarc of A with endpoints x and y is bounded by $C|x - y|$. In this case, the bi-Lipschitz mapping is given by a natural parametrization of the curve.

Theorem DM4 Separation λ and Point Energies.

(a) Let $A = S^d$ be the d sphere. For $d \geq 2$ and $s < d-1$,

$$\lambda(X_n, s) \geq cN^{-1/(s+1)}.$$

(b) Let $A = S^d$ be the d sphere. For $d \geq 3$ and $s \leq d-2$. Then

$$\lambda(X_n, s) \geq cN^{-1/(s+2)}$$

which is sharp in s for $s = d - 2$.

(c) For any $0 < s < d - 1$, there exists

$$\lim_{n\to\infty} \frac{\max_{1\leq j\leq n} \mathcal{E}_{j,s}(S^d, n)}{\min_{1\leq j\leq n} \mathcal{E}_{j,s}(S^d, n)} = 1.$$

Markov Chains on Data, Diffusion Distances and Embeddings onto Spheres

Let $X = \{x_1, x_2, ..., x_n\}$ be a data set in a metric space of high dimension.

We construct a graph (X, k) where:

- To each point x_i corresponds a node.
- Every two nodes are connected by an edge with a non negative weight/kernel $k(x, y)$.

The quantity $k(x, y)$ should reflect the degree of similarity or interaction between and x and y. The choice of the weight/kernel is crucial and application-driven.

Over the past 5 years, new techniques have emerged for manifold learning

- Isomap [Tenenbaum-DeSilva-Langford 00]
- L.L.E. [Roweis-Saul 01]
- Laplacian eigenmaps [Belkin-Niyogi 01]
- Hessian eigenmaps [Donoho-Grimes 03]
- Diffusion metrics [Coiffman-Lavon-04]

They all aim at finding coordinates on data sets by computing the eigenfunctions of a psd matrix.

Markov Chain on Data

Define the degree of a node x as

$$d(x) := \sum_{z \in X} k(x, z).$$

Form the $n \times n$ matrix P with entries

$$p(x, y) := \frac{k(x, y)}{d(x)}.$$

Because

$$\sum_{y \in X} p(x, y) = 1, \, p(x, y) \geq 0$$

P is the transition matrix of a Markov chain on the graph of the data and $I - P$ (following Chung-97) is the normalized graph Laplacian.

Time parameter t

$p_t(x, y)$ is the probability of transition from x to y in time $t \geq 1$ steps. Therefore, it is close to 1 if y is easily reachable from x in t steps. This happens if there are many paths connecting these two points

t defines the granularity of the analysis. Increasing the value of t is a way to integrate the local geometric information of the data.

In what follows, we will interested in defining a metric on the data set so that points are close in the metric if they are HIGHLY CONNECTED on the graph (X, k).

The metric will be be defined by way of an embedding Ψ_t of the data set X onto a sphere S^d of fixed dimension $d \geq 1$ realized as a subset of the Euclidean space $\mathbb{R}^{d+1}$. Here t is fixed.

In what follows $k(x, y) = k(y, x)$ and will be a function of distances between points x and y in X.

Diffusion distance:

- The diffusion metric measures proximity in terms of connectivity in the graph.
- It is useful to detect clusters
- Robust to noise unlike geodesic distance

Density Invariant Diffusion maps

Consider a perturbed kernel k_R:

$k_{\delta,R}(x,y) := |x-y|^{-s}$, $0 < s \leq d$, $d \geq 1$, $|x-y| \geq \delta$, $x, y \in S^d$

and 0 otherwise.

Suppose that the data set X is sampled from S^d with density μ'.

Question For each fixed δ, does $k_{\delta,R}$ have the property that one can separate the distribution of data points X with given density μ' from the geometry of the underlying submanifold M from which X is sampled.

Fix $\delta > 0$ and set for a scale $\varepsilon > 0$

$$k_{\delta,\varepsilon,R}(x,y) = \frac{\varepsilon k_{\delta,R}(x,y)}{l_\delta(x) l_\delta(y)},\ x,y \in S^d$$

where

$$l_\delta(y) = \sum_{x \in X} k_{\delta,\varepsilon,R}(x,y),\ y \in S^d.$$

If n is the cardinality of X, form the $n \times n$ matrix $P_{\delta,\varepsilon}$ with entries

$$\frac{k_{\delta,\varepsilon,R}(x,y)}{l_\delta(x)},\ x,y \in X$$

$P_{\delta,\varepsilon}$ is the transition matrix of a Markov chain on the graph of the data and $I - P_{\delta,\varepsilon}$ (following Chung-97) is the normalized graph Laplacian.

Theorem D For any fixed $\delta > 0$, there exists a fixed operator Δ_δ with compact support on S^d so that

$$\varepsilon(I - P_\delta) \to \Delta,\ n \to \infty,\ \varepsilon \to 0^+.$$

Many open problems: All preprints and related papers can be found on

http://math.georgiasouthern.edu/$\sim$ damelin

Thankyou for your attention