Diffusion Maps and Geometric Harmonics

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Stéphane S. Lafon

Dissertation Director: Ronald R. Coifman

May 2004

Abstract

Diffusion Maps and Geometric Harmonics

Stéphane S. Lafon

2004

The purpose of this thesis is twofold. First we investigate the problem of finding meaningful geometric descriptions of data sets. The approach that we propose is based upon diffusion processes. We show that by designing a local geometry that reflects some quantities of interest, it is possible to construct a diffusion operator whose eigendecomposition produces an embedding of the data into $\mathbb{R}^n$ via a *diffusion map*. In this space, the data points are reorganized in such a way that the geometry combines all the local information captured by the diffusion process, and the Euclidean distance defines a diffusion metric that measures the proximity of points in terms of their connectivity. The case of submanifolds of $\mathbb{R}^n$ is the object of greater attention, and we show how to define different kinds of diffusions on these structures in order to recover their Riemannian geometry. General types of anisotropic diffusions are also addressed, and we explain their interest in the study of differential and dynamical systems.

Secondly, we introduce a special set of functions that we term *geometric harmonics*. These functions allow to perform out-of-sample extensions of empirical functions defined on the data set. They can also be employed for embedding the data points in a Euclidean space with a small local Lipschitz distortion. We show that the geometric harmonics, and the corresponding restriction and extension operators are a valuable tool for the study of the relation between the intrinsic and extrinsic geometries of a set. In particular, they allow to define a multiscale extension scheme, in which empirical functions are decomposed into frequency bands, and each band is extended to a certain distance so that it satisfies some version of the Heisenberg principle.

A mes Parents.

Acknowledgments

I would like to express my deepest gratitude to my advisor Professor Ronald R. Coifman for his valuable supervision and for his constant interest in my thesis work. I highly appreciated his trust and his encouragements, but also his patience throughout these 3 years, and there is no day that goes by that I don't think about how fortunate I was to work under him.

I am very grateful to Professor Yves Meyer of ENS Cachan, France, for being the instigator of this adventure when he sent me to Yale in 2000.

I wish to thank Professors Naoki Saito of UC Davis and Steven Zucker of Yale for accepting to serve as readers for this thesis.

I would like to express my appreciation to Professor Jacques Peyrière of Université de Paris Sud, for his advice and interesting discussions.

I am very grateful to INRIA, France and in particular to Dr Evelyne Lutton, director of the COMPLEX team for making this journey possible.

My officemate Mark Tygert substantially contributed to this work through the many conversations we had.

I wish to thank the Gibbs instructors "gang": Ann Lee, Artur Luczak, Mauro Maggioni, Michael Mahoney, Boaz Nadler for fruitful discussions. Many thanks to Per-Gunnar Martinsson for providing me with some Matlab code and to Patrick Huggins for interesting conversations. I guess they all own a piece of the work presented in this thesis.

Many thanks go to Chris Hatchell for his terrific technical assistance.

On the personal side, I would like to thank Amine, Anne, Curt, Julien, Randy, Stéphane and Stéphane for their valuable support.

Contents

Notation

$\mathbb{Z}$ - the set of integers.
$\mathbb{R}$ - the set of real numbers.
dx - the Lebesgue measure on $\mathbb{R}^n$ or the Riemannian volume on a Riemannian manifold.
∇ - gradient in $\mathbb{R}^n$.
Δ - the Laplacian on $\mathbb{R}^n$ or the Laplace-Beltrami operator on a Riemannian manifold. The convention of sign here is such that Δ is a positive semi-definite operator. Its eigenfunctions are noted $\{\phi_j\}$ and its eigenvalues are expressed as squares $\{\nu_j^2\}$:

$$\Delta\phi_j = \nu_j^2\phi_j\,.$$

$p_t(x,y)$ - the Neumann heat kernel on a manifold.
$L^2(\Gamma, d\mu)$ - the space of square integrable functions on Γ with respect to the measure $d\mu$.
$\langle f, g\rangle_\Gamma$ - the inner product for functions in $L^2(\Gamma, d\mu)$:

$$\langle f, g\rangle_\Gamma = \int_\Gamma \overline{f(x)}g(x)d\mu(x)\,.$$

$H_s(\Gamma, d\mu)$ - the Sobolev space of functions whose derivatives of all orders up to s are square integrable on Γ with respect to $d\mu$.
$\|f\|_s$ - the norm in the Hilbert space $H_s(\Gamma, d\mu)$, given by:

$$\left(\sum_{|\alpha|\leq s}\int_\Gamma |\partial^\alpha f(x)|^2 d\mu\right)^{\frac{1}{2}}\,.$$

$\widehat{f}$ - the Fourier transform of f with the frequency convention:

$$\widehat{f}(\xi) = \int_{\mathbb{R}^n} e^{-2i\pi\langle\xi,x\rangle} f(x)dx\,.$$

$f * g$ - the convolution of two functions.
$f = \mathcal{O}(g)$ - means that the ratio of f over g is bounded from above.
$f \asymp g$ - means that the ratio of f and g is bounded from above and below.
$\overline{A}$ - the topological closure of a set.

Chapter 1

Introduction

1.1 The problem of dimensionality reduction

The tremendous growth of available data sources together with the amazing advances in storage capabilities have opened new horizons in science, business and government, and have drastically changed our relationship to the world. Nowadays, we are constantly flooded with information of all sorts and forms, and our strong attachment to induction makes us believe that it is somehow possible to learn from data. Learning is the key concept here, but this term comes in a variety of meanings. For instance, it is sometimes believed that learning should help us make predictions on the future, based on past experiences that are recorded in the data at our disposal. In that sense, learning means memorizing and reproducing information. But learning also means making automatic discoveries, or reinterpreting knowledge. In this case, learning involves comprehending the data as well as the ability to relate different parts of the information.

In any case, in response to the fast growing needs of governments, corporations and engineers to process information, there has recently been a huge effort to develop tools that convert data into useful knowledge, in particular in the following areas:

- biotechnologies: the human genome has now been entirely sequenced, and there remains to understand what each gene does. The analysis and interpretation of DNA microarrays has been a hot topics in molecular biology these last years. Even more challenging is the field of proteomics which aims at determining the structure, function and expression of proteins involved in our metabolism.

- text mining and web search engines: a substantial amount of human communication is produced in the form of text. The obvious consequence is that specific techniques need to be developed for this medium.

- Information Retrieval from meta data in general, with applications ranging from Customer Relationship Management to industrial espionage and foreign intelligence.

The common denominator of data analysis in these fields is that scientists are confronted with large amounts of observations that have high dimensionality. The dimension factor means that the description of each observation in the data or the relationship between observations involves a great number of observable quantities. For instance, in hyperspectral imagery, each observation is a collection of images whose pixels are themselves long vectors of reflectance at different wavelengths, and typically, each point in the data is a $500 \times 500 \times 500$

data cube (see our example on figure 1.1[1]). For this particular application, the data can be represented as points in a vector space, but other kinds of structures, such as graphs, are sometimes more adequate. A good example is provided by DNA sequences: a typical fragment of DNA contains 500 nucleotides, each of which being one out of four possible bases $\{A, C, G, T\}$. A common tool used for reordering the fragments is that of Markov chains of order k for which the nucleotides constitute the different states, and the goal is to estimate the transition probabilities between nucleotides along the DNA molecule. For the homogeneous Markov chain model, this approach is equivalent to viewing the data as a weighted graph with 4^k vertices corresponding to all $k-$words from the alphabet $\{A, C, G, T\}$, and the weight between the vertices being the probability of transitions. Empirical studies have shown that the memoryless Markov chain model ($k = 0$) is unrealistic, and that to obtain accuracy it is necessary to consider higher order Markov chains (see [22]). This means that we have to deal with large graphs having possibly many edges. In short, by high dimensionality, we mean that the representation of the data presents a potentially large number of degrees of freedom.

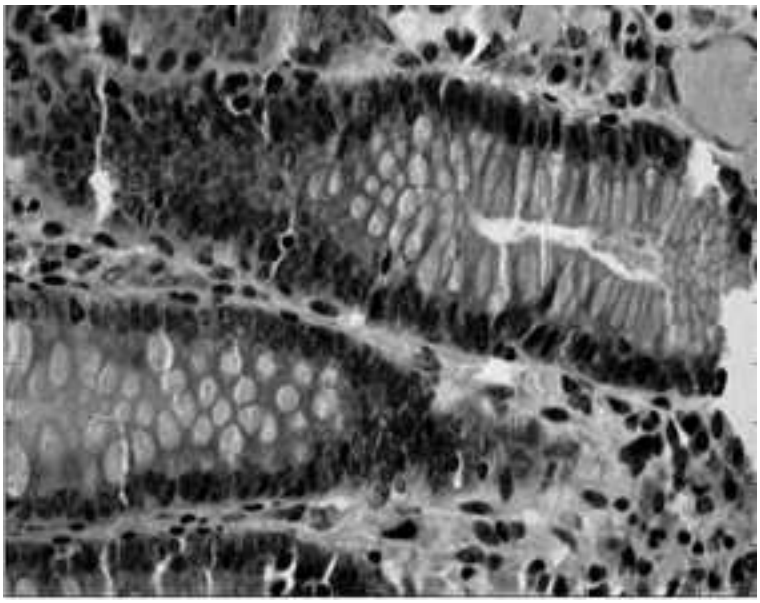

Figure 1.1: A slice of human colon tissue. The actual data is a cube of 652×491 pixels by 128 wavelengths.

The high dimensionality is an obstacle to an efficient processing of the data, and this phenomenon, termed *curse of dimensionality* by Bellman (see [5]), manifests itself in several ways:

- in $\mathbb{R}^d$, all norms are not (numerically) equivalent when d is large. In particular, the same function will have different degrees of smoothness under different norms.
- approximating functions is difficult; grid-based methods require ε^{-d} evaluations of a C^1 function f to approximate it with accuracy ε. Consequently, integration is a hard task too.

[1] Courtesy of Mauro Maggioni, Yale University.

- likewise, density estimation requires an number of sample points that grows exponentially with d, otherwise most bins of any histogram will be empty. This situation high dimension - low sample size is actually quite common. Although it constitutes a challenge for computational efficiency, a large number of observations is in general desirable.

- several algorithms notoriously fast in low dimension become prohibitively slow in high dimension as their complexity (in time, space and sometimes both) scales exponentially with d. For instance, when d is the average number of edges arriving at each node of a graph (its degree), algorithms for nearest neighbors search exhibit a complexity growing like an exponential function of d, making the brute force method the only realistic option... So to speak (the number of vertices is also a limitation).

See [9] for other aspects of this phenomenon and its impact on data analysis. Note that from these few remarks, it is clear that a vector space can be considered to be high-dimensional as soon as it has dimension greater than 10.

Fortunately, in spite of all these difficulties, the situation is not hopeless. Indeed, in many cases, the apparent complexity of the data is an artefact of the choice of their representation and has nothing to do with the actual complexity of the process that generated these data. It is often the case that the variables involved in the representation of the data are correlated through some functional dependence, and as a consequence, although the description of the data is highly multivariate, the number of independent variables that are necessary to efficiently describe the data is often small. This number of free parameters will be referred to as the *intrinsic dimensionality* of the data. Note that this notion applies to points in a vector space as well as vertices of a graph, where it means that some kind of regularity holds (bound on the degree of the vertices, volume growth condition). As an illustration, consider the example of hyperspectral data from the colon tissue (image 1.1) where each pixel is viewed as a point in $\mathbb{R}^{128}$. In the image space, the variability is explained by the local changes of the chemistry, plus some noise with much smaller variance. Therefore we can expect to observe correlations between wavelengths of neighboring pixels, and it is clear that the number of degrees of freedom for hyperspectral data is far less than for arbitrary points in a 128 dimension vector space.

Under the assumption of low intrinsic dimensionality, it seems reasonable to try to transform the representation of the data into a more efficient description by reducing the dimensionality. Not only would this allow to further process the data, but it could also reveal information and provide insight on the process at the origin of the data. Although a major obstruction to overcoming the curse of dimensionality is our lack of understanding of geometry in high dimension (see [8]), various attempts at reducing the dimension of data have been made and this problem now occupies a central position in many fields: in information theory where it is related to compression and coding, in statistics (latent variables), in pattern recognition (feature extraction), statistical learning (manifold learning)...

1.2 A brief review of some dimensionality reduction techniques

To reduce the dimensionality means to find a function Φ that will map our data from the space $\mathcal{X}$ of their original description to a new space $\mathcal{Y}$ where their description is considered to be simpler. In other words, Φ is supposed to discard part of the information in the data. As stated, the problem of dimensionality reduction is ill-posed, and we need to impose some

constraints on Φ. These conditions are defined by what information we are ready to lose, (or equivalently, what we want to preserve), and are dictated by the application we have in mind. Note that the loss of information is not necessarily a negative aspect of the dimension reduction as this information could be totally irrelevant to us (it could be noise for instance).

The mapping Φ will be referred to as an *embedding* of $\mathcal{X}$ into $\mathcal{Y}$.

1.2.1 Model fitting

One way to impose a constraint on the choice of Φ is to assume a model for the data. More precisely, dimension reduction can be achieved by fitting a low-dimensional model, and by using the model to describe the data. This is equivalent to performing some kind of regularization on the data, and therefore the choice or assessment of the model is a critical step prior to further treatment of the samples. The model should reflect our prior knowledge on the data, and at the same time result from a balance between its low complexity and its faithfulness to the data. Indeed, low complexity (measured by the number of parameters, by the Vapnik-Chernovenkis dimension or in terms of Kolmogorov complexity [10]) is desirable to achieve maximum dimensionality reduction and to avoid overfitting, but at the same time, one is inclined to increase the complexity to obtain a good accuracy, at the price of a large variance of estimators. There is no simple answer to that question, also known as a manifestation of the bias-variance tradeoff (see [16] for a description of several strategies).

Linear models: Principal Component Analysis

Assume that the data consist of points $\{x_1, ..., x_N\}$ in $\mathbb{R}^d$. In Principal Component Analysis (PCA), the data are fit to a linear model by computing the best linear approximation in the sense of the quadratic error. If X represents the $N \times d$ matrix of the data where the rows represent the samples, then we need to solve the problem

$$\arg \min_{\overline{x} \in \mathbb{R}^n, \dim H = k} \sum_{i=1}^{N} \|x_i - \overline{x} - P_H x_i\|^2$$

where P_H is the orthogonal projector onto a vector space H. It can be checked that the solution is obtained by the mean of the data

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{M} x_i$$

and

$$H = \text{span}\{u_1, u_2, ..., u_k\}$$

the linear space spanned by the eigenvectors $\{u_1, u_2, ..., u_k\}$ (the principal components) corresponding to the k largest eigenvalues $\{\sigma_1^2, \sigma_2^2, ..., \sigma_k^2\}$ of the matrix

$$X^T (I - M)^T (I - M) X$$

where M is the $N \times N$ matrix with each entry equal to $\frac{1}{N}$. The quadratic deviation from the affine space $\overline{x} + H$ is simply the sum

$$\sum_{i=k+1}^{N} \sigma_i^2 \,.$$

Therefore, the more eigenvalues we retain, the more accurate the model. At the same time, the dimension reduction is directly related to the number k of eigenvectors that we keep. From a statistical point of view, the data set can be thought of as the realizations of N i.i.d. random variables. Then $\frac{1}{N}X^T(I-M)^T(I-M)X$ is the empirical covariance matrix and its k top eigenvectors for the axes of maximum variance.

The linear model is often unadapted as the correlations between the variables of the data are generally nonlinear. Despite this problem, PCA is extremely popular because of its simplicity and the interpretability of its results.

Kernel PCA

Among the different attempts to correct the fact that PCA cannot handle nonlinear data sets, kernel PCA is of particular interest in this thesis because it relies on the diagonalization of a positive semi-definite kernel restricted to a data set $\{x_1, x_2, ..., x_N\}$. More precisely, assume the existence of a function Φ that maps the data points into some vector space V, possibly infinite dimensional (the "feature space"). In the space V, classical PCA can be performed, and the principal components can be employed for classification or regression ends. In particular, if the classification scheme is linear (in V), it will only involve the evaluation of inner products between linear combinations of terms of the form $\Phi(x_j)$ and possibly $\Phi(x)$ where x is a new point. In particular, all computations are based upon the knowledge of the matrix with entries $\langle \Phi(x_i), \Phi(x_j) \rangle$. Consequently, one does not need to know Φ explicitly, but instead one could start from a positive semi-definite kernel $k(x_i, x_j)$ and think of it as the Gram matrix $\langle \Phi(x_i), \Phi(x_j) \rangle$. Now kernel PCA is equivalent to the diagonalization of the matrix $k(x_i, x_j)$, and all operations are realized in terms of the known quantities $k(x, x_j)$ (see [15] for more details).

Popular choices of kernels are $k(x,y) = (\langle x, y \rangle)^p$ (polynomial kernel) and $k(x,y) = \exp(-\frac{\|x-y\|^2}{\sigma^2})$ (Gaussian kernel). Kernel PCA reduces the dimension of the space of functions defined on the data in the sense that all such functions are represented as a linear combination of bumps of the form $k(x_i, .)$, and the dimensionality of this space depends on the decay of the spectrum of this kernel on the data.

Cluster analysis

Suppose that there exists a measure of similarity between the observations. Clustering aims at partitioning the data into groups such that the similarity between points in the same groups is smaller than that between points of different groups. A popular clustering technique is based on mixture models for which the density of the observations is represented as a convex combination of elementary shapes, like Gaussians for instance.

1.2.2 Preservation of mutual distances

Suppose that the data is a collection of points $\Gamma = \{x_0, x_1, ...\}$ in some metric space $(\mathcal{X}, \rho)$. For instance $\mathcal{X}$ can be a Euclidean space with its associated norm, or a graph with ρ being the geodesic distance between vertices. Let $\Phi : (\mathcal{X}, \rho) \to (\mathcal{Y}, \eta)$ be an embedding into another metric space. Define the *expansion* of Φ as

$$M(\Phi) = \sup_{u,v \text{ in } \mathcal{X}} \frac{\eta(\Phi(u), \Phi(v))}{\rho(u,v)}$$

and its *shrinkage* to be

$$m(\Phi) = \sup_{u,v \text{ in } \mathcal{X}} \frac{\rho(u,v)}{\eta(\Phi(u),\Phi(v))} .$$

Therefore we have, for all u and v in $\mathcal{X}$,

$$\frac{1}{m(\Phi)}\rho(u,v) \leq \eta(\Phi(u),\Phi(v)) \leq M(\Phi)\rho(u,v) .$$

The *Lipschitz distortion* of Φ is the following product:

$$\mathrm{dist}(\Phi) = m(\Phi)M(\Phi) .$$

The distortion measures how much stretching or contraction is applied to points in $\mathcal{X}$ when embedded via Φ, and it is a generalization of the condition number of linear functions to nonlinear mappings. In particular,

$$\mathrm{dist}(\Phi) \geq 1$$

with equality if and only if Φ is a transform that changes all distances in the same ratio.

When mutual distances between the data points are meaningful and therefore need to be preserved, we want to obtain an embedding Φ with distortion as close to 1 as possible. But at the same time, it is necessary to reduce the dimension and a trade-off is to be found between these two competing requirements.

Multidimensional scaling

The existence and construction of isometric embeddings (distortion equal to 1) into Euclidean spaces are easily obtained by using a technique that goes by the name of (classical) Multidimensional Scaling (MDS). The result is the following: the set Γ can be embedded in a Euclidean space if and only if the kernel

$$k(x,y) = \rho^2(x_0,x) + \rho^2(x_0,y) - \rho^2(x,y)$$

is positive semi-definite (x_0 is any point in the data set). The idea here is that k represents (twice) the Gram matrix of the image of the data set through the embedding Φ. The embedding is given by the column of any matrix a such that $k = a^T a$, for instance a can be obtained from the Cholesky decomposition of k or using its Singular Value Decomposition (SVD). Small singular values represent dimensions that can be numerically neglected, and the number of significant singular values will give the dimension of the Euclidean space necessary to isometrically embed the data to a preset accuracy. This was known to the mathematicians of the 1930's and Schoenberg (see [25]) showed that this characterization of the kernel k could also be transferred to other types of kernels: the data points will be isometrically embeddable in a Euclidean space if and only if the kernel $e^{-\rho^2(x,y)}$ is positive semi-definite on Γ.

It can be verified that when the data already lie in a vector space, classical MDS and PCA are dual in the sense that MDS amounts to diagonalizing $(I-M)XX^T(I-M)^T$ whereas PCA is obtained from the eigenvectors of $X^T(I-M)^T(I-M)X$.

We conclude this section by mentioning that classical MDS has inspired several algorithms for visualization of abstract data or for nonlinear dimensionality reduction, like the recent ISOMAP (see [33]) that uses an MDS embedding on the set of geodesic distances of a graph or a manifold.

Randomized embedding

Another approach consists in randomizing the search for an embedding with small distortion. Randomness is a promising tool for overcoming the difficulties related to the curse of dimensionality, and is already exploited for integration and simulation in high dimension (via Monte-Carlo and quasi Monte-Carlo methods) and in fast matrix computations [13]. Random methods can be used for finding low distortion embeddings as it is shown in the following theorem [18]:

Theorem 1 (Johnson-Lindenstrauss). *Let $\varepsilon > 0$ and suppose that $\Gamma \subset \mathbb{R}^d$ contains n points. If $k \geq \frac{24}{\varepsilon^2}\log(n)$ then there exists a mapping $\Phi : \Gamma \to (\mathbb{R}^k, \|.\|_2)$ such that*

$$dist(\Phi) \leq \sqrt{\frac{1+\varepsilon}{1-\varepsilon}}.$$

Moreover, this map can be found in randomized polynomial time.

In other words, n points of a Euclidean space can always be mapped into a space of dimension $\mathcal{O}\left(\frac{1}{\varepsilon^2}\log n\right)$ with a small distortion. The mapping is found by projecting the points onto random low-dimensional linear spaces. It is remarkable that a solution exists for all set Γ (this is not the case if the Euclidean metric is changed into the L^∞ or L^1 distance function), but it is also a drawback because this technique does not take into account the specific geometry that Γ might have. In particular, the intrinsic dimensionality of Γ plays no role in the theorem above, and the embedding is therefore suboptimal.

1.2.3 Global vs local

One of the major drawbacks of the methods presented so far (except for kernel PCA and cluster analysis) is that they all aim at minimizing some global cost function:

- for PCA, one tries to globally fit the data with a linear model. Most data sets are highly nonlinear, and this method fails at capturing the nonlinear structures in the data.
- To realize the potential inefficiency of classical MDS, consider a curve in $\mathbb{R}^d$ (d large) following each of the d coordinate axes successively . This curve will be mapped to itself through classical MDS (no dimension reduction).
- The Lipschitz distortion as defined above is also a global measure in the sense that under a mapping with reasonable distortion, two close points must be mapped to two close points, and two points far apart must stay as such. The Johnson-Lindenstrauss, as well as ISOMAP, are thus largely suboptimal in situations where large distances do not need to be preserved.

As indicated above, trying to preserve large distances can be quite inefficient. Not only is it a limitation on the final dimensionality, but in many applications it is also irrelevant. The reason for this lies in the fact that, often, the distance used to discriminate between data points is only meaningful for close points. Indeed, suppose that the data points x_i were generated by a low dimensional parameter θ_i via a mapping $\Phi : \mathbb{R}^k \to \mathbb{R}^d$ (d much larger than k). If we measure distances with Euclidean distances in both spaces, how do $\|\theta_i - \theta_j\|$

and $\|\Phi(x_i) - \Phi(x_j)\|$ compare? A related question is how smooth can Φ be ? Consider the following simple example: $\theta \in [0,1]$ and $\Phi : [0,1] \to L^2([0,1])$ defined by

$$\Phi(\theta)(t) = \begin{cases} 1 \text{ if } 0 \leq t \leq \theta \\ 0 \text{ otherwise.} \end{cases}$$

This mapping is a simple but instructive model for edges in images for instance, θ being a location parameter. Then it is immediate that

$$\|\Phi(\alpha) - \Phi(\beta)\| = |\alpha - \beta|^{\frac{1}{2}} .$$

This means that Φ is not very smooth and as a consequence, small variations of the parameter value will entail large variations in the high dimensional space. Therefore, although the Euclidean distance in $L^2([0,1])$ will sharply discriminate between points with very close values of the parameter, it is essentially useless for comparing all other points. Thus it is clear that minimizing a global cost function like Lipschitz distortion makes no sense in this case as large distances are meaningless and need not be preserved.

In recent years, several local methods have emerged from the field of manifold learning to address this issue: Local Linear Embedding (LLE) [24], Laplacian eigenmaps [4], Hessian eigenmaps [11], Local Tangent Space Alignment (LTSA) [34], etc... All these techniques aim at minimizing distortions of the form:

$$Q(f) = \sum_i Q_i(f)$$

under the constraints that $\|f\|^2 = 1$. Here the sum is taken over all points in the data set and $Q_i(f)$ is a positive semi-definite quadratic form in f. In all the methods cited above Q_i is local in the sense that it involves values of f in a neighborhood of x_i, typically $Q_i(f)$ is a measure of the variation of f around x_i: squared norm of the gradient for Laplacian eigenmaps, Frobenius norm of the Hessian matrix for Hessian eigenmaps, deviation from a local representation of the points for LLE and LTSA. The idea behind the special form of Q is that one hopes to derive global information from local overlapping structures. In addition, the problem can be solved efficiently as the matrix of Q is sparse. The solution to the optimization problem is given by the axes of inertia of Q, and the first eigenvectors are used to represent the data. It was noted in [15] that all these methods are subcases of kernel PCA. In the first part of this thesis, we will attempt to explain that they also all have an interpretation in terms of diffusion process.

In his thesis work, Belkin [3] showed that the so-called weighted graph Laplacian allows to reconstruct the Laplace-Beltrami operator on a manifold from points uniformly sampled, and that the eigenfunctions of this operator can be used to perform dimensionality reduction. In our work, we show that this is not true for non-uniform densities, and we improve his result by describing an algorithm that handles general densities.

1.3 Extrinsic and intrinsic geometries

In this thesis, the geometry of a set Γ of objects will be defined as a set of rules that describe the relationship between the elements of Γ. When Γ is a subset of a bigger set Ω, we will say the geometry is *intrinsic* to Γ if the rules can be formulated without reference to Ω and the possible structures already existing on it. If it is the case that Ω possesses its own geometry, then the geometry induced on Γ will be referred to as *extrinsic geometry* of Γ.

Weighted graphs and Riemannian manifolds are two examples of structures of particular interest in the work presented here. A weighted graph (Γ, k) is a collection of points Γ together with a real-valued weight function k defined on $\Gamma \times \Gamma$. In that case, the geometry of the graph is defined by the weight k of the edges that describe the degree of association of the vertices. For a Riemannian manifold Γ, the geometry is contained in the field of metric tensors on the manifold. At each point x on the manifold, the tensor $\{g_{ij}(x)\}$ defines an inner product on the tangent space, and a metric in a neighborhood of x via the exponential map. When Γ is a submanifold of a Euclidean space, the metric on each tangent space is inherited from that of the bigger space.

In this thesis, we explore the following simple observation: much of the geometry of a set Γ can be understood through the analysis of the geometry of the space of functions defined on Γ. In some way, this dual approach to the study of the set Γ allows us to deal with the same kind of objects (functions or operators on those functions) for all different types of sets (graphs, manifolds...). Dualization is a powerful tool that allows to generalize some concepts as it was done in distribution theory for instance, where a function is not defined by the values taken on a set, but rather by its action on a space on test functions. The idea of studying functions defined on a set (and operators on these functions) to gain some insight on the geometry of the set itself is not new. Indeed, it is at the center of many inverse problems and has been extensively studied in inverse scattering, potential theory and spectral geometry. For instance consider the problem of learning the structure of the ground from seismic data, or the reconstruction of 3D models from the scattering of X-rays in medical imaging. In these examples, the geometric structure of some objects is inferred from the action of a set of functions (plane waves) on these objects. Similarly, in potential theory, the singularities of the boundary (corners, cusps) are reflected in the behavior of the solutions of the Dirichlet and Neumann problems. Spectral geometry asks the question of whether the geometry of a Riemannian manifold is determined by the spectrum of the Laplace operator. More generally, the study of spectral asymptotics for partial differential operators relates geometric characteristics of sets Γ to the growth of the eigenvalues for such operators.

In this thesis, we explore the relation between both geometries by investigating the action of certain restriction and extension operators, and investigate the question of how the intrinsic and extrinsic diffusion are related.

1.4 Contribution of this thesis

The principal contribution of this thesis is twofold. First we show the relevance and usefulness of diffusion processes for understanding the geometric structures of data sets, and we explain how it is related to the geometry of spaces of functions defined on these data. We also show that this approach generalizes some concepts that were introduced in areas such as manifold learning and differential calculus. Second, we introduce a simple tool, the geometric harmonics, that allows to perform an out-of-sample extension of an empirical function known on the data. In addition to their obvious potential in applications, the geometric harmonics are shown to be very useful to relate the intrinsic and extrinsic geometries of data sets.

Chapter 1 deals with defining diffusion processes on a data set, and these are used to infer a description of the intrinsic geometry of the data. We explicitly construct a diffusion kernel on the data, and employ its spectral properties, spectrum and eigenfunctions, to define a

diffusion map that embeds the data into a Euclidean space, where the Euclidean distance corresponds to a diffusion metric. The case of submanifolds of $\mathbb{R}^n$ is studied in details, and we define a diffusion process that corrects the defects of the classical tool, namely the Graph Laplacian. Numerical experiments are presented, and we discuss the usefulness of general diffusion processes, in particular for the study of dynamical and differential systems.

Chapter 2 includes the extrinsic geometry in the discussion. We construct a special set of functions, termed geometric harmonics, and show that they allow to extend an empirical function known on the data to new points. We then show two other ways to use these functions. First, we empirically show that they provide a reduction of dimensionality of the data with a small local distortion. And second, they are shown to provide a link between the intrinsic and extrinsic geometries of a set, and they allow to define a multiscale extension scheme, in which empirical functions are decomposed into frequency bands, and each band is extended to a certain distance so that it satisfies some version of the Heisenberg principle.

Chapter 2

Diffusion Maps

2.1 Motivation: from local to global

For a large variety of data, the notion of similarity or distance between points is defined only locally. More precisely, at each point of the set is defined a neighborhood, and within this neighborhood, one has a similarity or distance function between points. This idea is contained in that of a Riemannian manifold: for such a structure the domains of the charts define the neighborhoods and the metric tensor explains how to locally measure the distance between points. The point of view of Riemannian geometry is not the only relevant one, and in the following, it will be useful to think of the data as forming a weighted oriented graph. With this approach, the neighborhood of a point x is defined as the set of points that are connected to x, and the similarity between x and y is given by the weight of the edge (x, y).

To illustrate the idea that the similarity measure only makes sense locally, let's consider a database of digit pictures, as those used for calibrating Optical Character Recognition systems. The set consists of images of handwritten digits, and each picture being composed of $m \times n$ pixels, it is generally viewed as a point sitting in the Euclidean space $\mathbb{R}^{mn}$. Ideally, one would like to identify 10 clusters in the data (corresponding to the digits 0,1,...,9), but the geometry of the set as measured with the Euclidean distance (for instance) is quite complex and does not allow to perform an efficient clustering. The inefficiency of the Euclidean distance is easily explained by the following observation: two points of the set are either close, and then their Euclidean distance brings us some useful information, or they are far from each other and the information of their distance is irrelevant (the points can be considered to be at infinite distance). This situation is simply the expression of the fact that in high dimension (mn), the Euclidean distance is *not* a smooth function of the natural parameters controlling the variability within the data set. For example, if two instances of the same digit, say 1, are almost identical except that one is a rotated version of the other by a small angle α, then the Euclidean distance between these points will be proportional to $\sqrt{\alpha}$. In fact, due to the finite resolution, the situation is even worse: two instances of 1 will be either at very large distance or at distance almost zero, making the Euclidean distance practically meaningless for this example.

As a consequence, several instances of the same digit, that we would like to group in a unique cluster, can differ by a large amount in the Euclidean distance. Therefore this distance can only relate instances of these 1's that are very similar. If the parameters that control the variability between these 1's are sufficiently sampled, then one can hope to agglomerate the local information of the Euclidean distance to infer the global structure of

Figure 2.1: Two instances of the same digit. Their Euclidean distance is roughly proportional to $\sqrt{\alpha}$.

the data set. This is precisely what a random walk (diffusion, Markov process) does, as its trajectories chain the different points according to the local geometry.

In this chapter, we explain how the eigenvalues and eigenfunctions of averaging operators, *i.e.*, operators whose kernel corresponds to transition probabilities of a Markov process, define a natural embedding of the data through a *diffusion map*. In the embedding space, the Euclidean distance gives rise to natural metric on the data. We show that this metric measures the distance in terms of diffusion between the data points and that it provides us with robust information on the *intrinsic geometry* of the data set. Furthermore, the study of the eigenvalues allows us to use the eigenfunctions for dimensionality reduction. We also explain how the Neumann heat kernel can be approximated when Γ is a submanifold of $\mathbb{R}^n$, and we illustrate these ideas by some numerical experiments. We conclude this chapter by a discussion on anisotropic diffusions.

2.2 Definition of the diffusion metric

Let $(\Gamma, \mathcal{A}, \mu)$ be a measure space, where Γ is a set whose points are abstract objects. Γ can have a very general form, but in many practical situations, it will consist of finitely many data points, and μ will be the counting measure in order to represent the distribution of the points in the data set. To simplify, we shall assume that μ is finite. Our goal is to study the *intrinsic* geometry of this set, and to do so, we construct a diffusion kernel and use its spectral properties to analyze the geometry of the data.

2.2.1 Construction of a diffusion kernel

Suppose that the geometry of Γ is defined by a kernel $k(x, y)$, that is, assume that $k(x, y)$ measures the degree of similarity between two points x and y. The kernel k represents our

a priori information on Γ. In this section we show how to construct a diffusion kernel from k. We make the following additional assumptions on k:

- k is symmetric: $k(x,y) = k(y,x)$,
- k is positivity-preserving: for all x and y in Γ, $k(x,y) \geq 0$,
- k is positive semi-definite: for all bounded function f defined on Γ,

$$\int_\Gamma \int_\Gamma k(x,y) f(x) f(y) d\mu(x) d\mu(y) \geq 0\,.$$

In the following, these conditions will be referred to as the admissibility conditions.

Let's make a few remarks on these properties. First, the kernel defines a notion of neighborhood, namely the neighborhood of x corresponds to all points y that interact with x, *i.e.*, such that $k(x,y)$ is numerically significant. In that sense, the kernel defines the local geometry of Γ. Second, as we will show in an example, the symmetry is not really a constraint since we can always consider a symmetrized version of the kernel. Third, the positivity preservation property will allow us to renormalize k into a Markov kernel and to define a random walk on the data. Last, as we will see, the third condition is necessary for imposing the positivity of the diffusion metric. Under some technical conditions on μ and Γ, it has an equivalent discrete formulation (see [6]).

An important class of examples is generated by the situation when Γ is a subset of the Euclidean space $\mathbb{R}^n$. In this case, if x and y belong to Γ, the similarity measure is a function of the Euclidean distance $\|x-y\|$:

$$k(x,y) = \eta(\|x-y\|)\,.$$

To guarantee the positivity of this kernel, η must be chosen as the Fourier transform of a positive measure (Bochner's theorem). This type of example is investigated in more details in section 2.3.

Another situation is provided by graph theory. Let the points of Γ be the vertices of an *oriented* graph, and let $b(x,y)$ be the associated adjacency matrix, that is, $b(x,y) = 1$ if there is an edge going from x to y, and $b(x,y) = 0$ otherwise. The kernel b defines a notion of neighborhood for each point, and also a non-symmetric distance given by $1 - b(x,y)$. Clearly b is not symmetric in general, but we can consider

$$k_1(x,y) = \int_\Gamma b(x,u) b(y,u) d\mu(u)$$

and

$$k_2(x,y) = \int_\Gamma b(u,x) b(u,y) d\mu(u)\,.$$

The kernel $k_1(x,y)$ counts the number of common neighbors to x and y, whereas $k_2(x,y)$ counts the number of points for which x and y are common neighbors, *i.e.*, two kernels are admissible.

The kernel k can be re-normalized to be stochastic (to have sum 1 along its rows): define

$$v^2(x) = \int_\Gamma k(x,y) d\mu(y)\,.$$

This is well-defined as $k(x,y) \geq 0$. Then clearly $\tilde{a}(x,y) = \frac{k(x,y)}{v^2(x)}$ has sum 1 along the y coordinate:

$$\int_\Gamma \tilde{a}(x,y)d\mu(y) = 1\,.$$

Moreover, for all x and y, $\tilde{a}(x,y) \geq 0$. As a consequence, $\tilde{a}$ can be interpreted as the transition matrix of a homogeneous Markov process on Γ. This normalization is very commonly used in spectral graph theory (see [7]) where $I - \tilde{A}$ is known as the normalized weighted graph Laplacian. This procedure shows that to each admissible kernel one can associate a random walk on Γ. Note that from an analysis perspective, the operator

$$\tilde{A}f(x) = \int_\Gamma \tilde{a}(x,y)f(y)d\mu(y)$$

corresponding to this kernel is an averaging operator as it fixes constant functions, and it is also positivity-preserving: if $f \geq 0$ then $\tilde{A}f \geq 0$.

Since we are interested in the spectral properties of this operator, it is preferable to work with a symmetric conjugate to $\tilde{A}$: we conjugate $\tilde{a}$ by v in order to obtain a symmetric form and we consider

$$a(x,y) = \frac{k(x,y)}{v(x)v(y)} = v(x)\tilde{a}(x,y)\frac{1}{v(y)}$$

and

$$Af(x) = \int_\Gamma a(x,y)f(y)d\mu(y)\,.$$

The new kernel is therefore conjugate to the stochastic kernel, and shares the same spectrum, and its eigenfunctions are obtained by conjugation by v. In what follows, we will use A rather than $\tilde{A}$ and we will refer to A as a diffusion operator.

Lemma 2. *The diffusion operator A with kernel a*

$$Af(x) = \int_\Gamma a(x,y)f(y)d\mu(y)$$

is bounded from $L^2(\Gamma, d\mu)$ into itself. Its norm is

$$\|A\| = 1$$

and is achieved by the eigenfunction v:

$$Av = v\,.$$

Moreover, A is symmetric and positive semi-definite.

Proof. Let $f \in L^2(\Gamma, d\mu)$. We have,

$$\langle Af, f\rangle = \int_\Gamma \int_\Gamma k(x,y)\frac{f(x)}{v(x)}\frac{f(y)}{v(y)}d\mu(x)d\mu(y)\,. \tag{2.1}$$

If we apply the Cauchy-Schwarz inequality:

$$\begin{aligned}\left|\int_\Gamma k(x,y)\frac{f(y)}{v(y)}d\mu(y)\right| &\leq \left(\int_\Gamma k(x,y)d\mu(y)\right)^{\frac{1}{2}}\left(\int_\Gamma k(x,y)\frac{f(y)^2}{v(y)^2}d\mu(y)\right)^{\frac{1}{2}} \\ &\leq v(x)\left(\int_\Gamma k(x,y)\frac{f(y)^2}{v(y)^2}d\mu(y)\right)^{\frac{1}{2}}\,.\end{aligned}$$

Consequently,

$$\langle Af, f\rangle \leq \int_\Gamma |f(x)| \left(\int_\Gamma k(x,y) \frac{f(y)^2}{v(y)^2} d\mu(y) \right)^{\frac{1}{2}} d\mu(x)\,.$$

Let's apply the Cauchy-Schwarz inequality once again:

$$\langle Af, f\rangle \leq \|f\| \left(\int_\Gamma \int_\Gamma k(x,y) \frac{f(y)^2}{v(y)^2} d\mu(y) d\mu(x) \right)^{\frac{1}{2}} = \|f\|^2\,.$$

where we have used the symmetry of the kernel. The positivity results from equation (2.1) and the positivity of k. Last, it is immediate that $Av = v$. □

2.2.2 Spectral decomposition of the diffusion kernel

The operator A being bounded and self-adjoint, the spectral theorem applies:

$$a(x,y) = \sum_{j\geq 0} \lambda_j \phi_j(x) \phi_j(y)$$

where

$$A\phi_j(x) = \lambda_j \phi_j(x)\,.$$

Here we have assumed that A is more than bounded, it is also compact (therefore the spectrum is discrete). The eigenvalues λ_j are non-increasing and non-negative by the positivity of A. In addition, $\lambda_0 = 1$ by lemma 2.

Let $a^{(m)}(x,y)$ denote the kernel of A^m. Then we have

$$a^{(m)}(x,y) = \sum_{j\geq 0} \lambda_j^m \phi_j(x) \phi_j(y)\,. \tag{2.2}$$

There are two possible levels of interpretation for the kernel and its eigenfunctions:

- at the level of the data points, *i.e.*, the elements of Γ, the kernel $a^{(m)}(x,y)d\mu(y)$ has a probabilistic interpretation as (up to a conjugation by v) the probability for a Markov chain with transition matrix a to reach y from x in m steps. Likewise, eigenfunctions can be thought of as coordinates on the data; this idea is explored in the next section.

- the dual point of view is that of the functions defined on the data. The kernel $a^{(m)}$ can be viewed as a bump of scale m, and as the value of m increases, the kernel gets wider on the data points. To relate this scale to the spectrum of A^m, we make the following observation: since $0 \leq \lambda_j \leq 1$, as m increases, only a few terms survive in sum (2.2), namely those for which λ_j^m exceeds a certain threshold. This means that to reconstruct the bump $a^{(m)}(x,y)$ centered at x and of "width" m, a small number of eigenfunctions are needed, and this number gets smaller as the scale m increases. This observation, which corresponds to a version of the Heisenberg principle, shows how the spectral decomposition (2.2) provides a multiscale analysis of the functions defined on the set Γ.

We now make use of the result of lemma 2 to define a mapping of the data into a Euclidean space and we investigate the first point of view presented above.

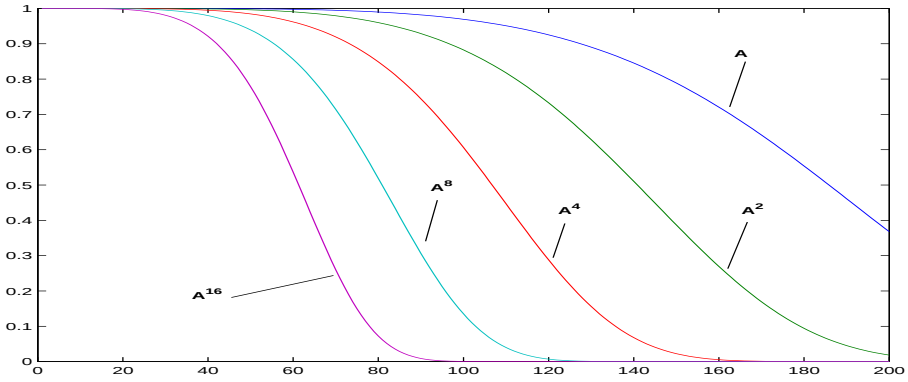

Figure 2.2: Typical spectra of A and some of its iterates.

2.2.3 Nonlinear embedding, diffusion metrics and dimensionality reduction

We introduce the following mapping:

$$\Phi(x) = \begin{pmatrix} \phi_0(x) \\ \phi_1(x) \\ \phi_2(x) \\ \vdots \end{pmatrix}.$$

Φ maps Γ into the Euclidean space $l^2(\mathbb{N})$. Therefore, each eigenfunction is interpreted as a coordinate on the set. This mapping thus takes abstract entities (remember that the data points need not be points in a vector space) and provides a representation of the data as points in a Euclidean space. This seems remarkable, but is it really? In fact there are thousands of ways to achieve this. The relevant question is: what characterizes this mapping?

To be able to answer this question, we also define the family $\{D_m\}_{m\geq 1}$ of metrics on Γ as:

$$D_m^2(x,y) = a^{(m)}(x,x) + a^{(m)}(y,y) - 2a^{(m)}(x,y)$$

which is well-defined because $a^{(m)}$ is a positive semi-definite kernel and it can be checked that

$$D_m^2(x,y) = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} a^{(m)}(x,x) & a^{(m)}(x,y) \\ a^{(m)}(x,y) & a^{(m)}(y,y) \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix}. \tag{2.3}$$

The quantity $D_m(x,y)$ has a set and a functional interpretation. First it can be considered as a diffusion distance between x and y: it measures the rate of connectivity between points of the data set. It will be small if there are a lot of paths of length less than or equal to m between these two points, and it will be large if, on the contrary, the number of connections is small. Unlike the geodesic distance, the diffusion distance is robust to noise and topological short-circuits because it is an average over all paths connecting two points, see Figure 2.3. In this example, the set is composed of points thrown at random on two disjoint disks. Because of the presence of some noise, there is some leakage between the two disks. This entails that the geodesic distance from A to B is not much larger than that between B and C. From the point of view of the diffusion metric, points B and C are

connected by a lot of paths and therefore are close. On the contrary, because of the presence of a bottleneck, points A and B are connected by relatively few paths, making these points very distant from each other. The diffusion distance is therefore able to separate the two disks.

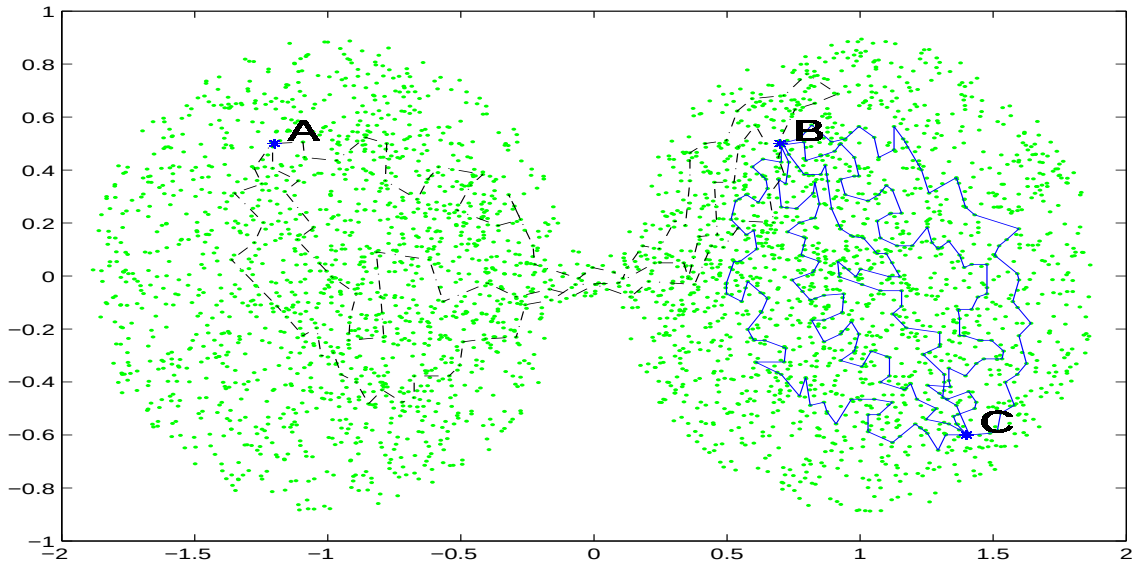

Figure 2.3: Unlike the geodesic distance, the diffusion metric D_m is robust to short circuits. In the example above, points B and C are connected by a lot of paths and therefore are close in the sense of D_m. On the contrary, because of the presence of a bottleneck, points A and B are connected by relatively few paths, making these points very distant from each other.

In addition to being a distance between points of the set, D_m is also a distance between the bumps mentioned in section 2.2.2. Indeed, $D_{2m}(x,y)$ is the Euclidean distance between the columns of indices x and y in the matrix $a^{(m)}$. In other words,

$$D_{2m}^2(x,y) = \int_\Gamma |a^{(m)}(x,z) - a^{(m)}(y,z)|^2 d\mu(z) = \|a^{(m)}(x,\cdot) - a^{(m)}(y,\cdot)\|^2 .$$

A remarkable fact is that this complex quantity can be simply measured in the embedding space $l^2(\mathbb{N})$:

Proposition 3. *We have:*

$$D_m^2(x,y) = \sum_{j\geq 0} \lambda_j^m (\phi_j(x) - \phi_j(y))^2 .$$

In other words, the diffusion metric can be computed as a weighted Euclidean distance in the embedding space, the weights being $\lambda_0^m, \lambda_1^m, ...$ As a corollary, in $l^2(\mathbb{N})$ the diffusion balls are ellipsoids whose axes are parallel to the coordinate axes, with lengths given by the powers of the eigenvalues.

By the weighted Euclidean distance with weights (w_i) we mean that

$$\|(u_i) - (v_i)\|^2 = \sum_i w_i (u_i - v_i)^2 .$$

Proof. The equality is the mere consequence of identity (2.2). □

This proposition gives an answer to the question raised previously: the embedding Φ provides a representation of the data as points of a Euclidean space in such a way that the weighted distance in this space is equal to the diffusion distance on the data. Also, this proposition shows that D_m is a semi-metric in the classical sense (it is symmetric, non-negative and verifies the triangular inequality). Moreover, identity (2.3) shows that when the kernel is strictly definite positive, then D_m is a metric, *i.e.*, $D_m(x, y) = 0 \Rightarrow x = y$.

A corollary of this result is that the embedding generated by the eigenfunctions allows a dimensionality reduction of the data. Indeed, for a given accuracy δ, we retain only the eigenvalues $\lambda_0, ..., \lambda_{p-1}$ that, when raised to the power m, exceed a certain threshold (related to δ), and we use the corresponding eigenfunctions $\phi_0, ..., \phi_{p-1}$ to embed the data points into $\mathbb{R}^p$. The property of this embedding Φ_p is that the Euclidean distance weighted by $\lambda_0^m, ...\lambda_{p-1}^m$ gives the diffusion distance D_m at time m with accuracy δ. As a consequence, the maps Φ and Φ_p will be referred to as *diffusion maps*.

In short: the jumps of the spectrum (spectral gaps) specify the dimension reduction given some accuracy, and the eigenfunctions provide coordinates to implement this reduction. Note that the dimension of the embedding is not necessarily equal to the dimensionality d of the set. Indeed, the dimension of the new representation depends on the diffusion process on the data, and although it is related to d, it is in general greater than this number.

2.3 The case of submanifolds of $\mathbb{R}^n$

We now investigate in further details the case when Γ is a compact differentiable submanifold of $\mathbb{R}^n$. This means that Γ is a Riemannian manifold whose Riemannian metric is given by the Euclidean distance of the ambient vector space $\mathbb{R}^n$.

2.3.1 Framework

We emphasize the fact that the case of Γ being a submanifold is of great practical importance as in many applications, each point of the data set is a collection of numerical measurements. In image processing for instance, an image is a collection of pixels, or each pixel is mapped to its 3×3 neighborhood. In hyperspectral imaging, each point is a sequence of measurements of transmittance at different wavelengths. Another example of interest is that of a system of N particles, where each point is the measurement of $3N$ position coordinates and $3N$ velocity coordinates. These three examples justify the importance of Γ being a subset of $\mathbb{R}^n$. Very often, the data points are locally related by a set of equations that arise from the phenomenon at the origin of the data, and in this case the submanifold model makes sense. In the example of the N particles, the data points are related by the equations arising from the physical laws governing the evolution of the system. For more discussion on the relevance of the submanifold model, see [3] for the general setting, and [17] for the example of computer vision and edge modeling. Furthermore, in a large variety of data, the dimension of the submanifold is much smaller than that of the ambient space. In other words, although the representation of the data points is highly multivariate, the local variability is controlled by a small number of parameters.

As for the notations, let Γ be a C^∞ submanifold of dimension d in $\mathbb{R}^n$, $d < n$, and μ be a measure on Γ. The metric on Γ is that induced by that of the ambient space $\mathbb{R}^n$. We shall assume that μ has a density with respect to the Riemannian measure dx on Γ (*i.e.*,

$d\mu(x) = p(x)dx$). This density $p(x)$ can be thought of the density of the sample points in our data set, thus it does not have to be uniform. From a practical point of view, the following study makes sense if the number of points is sufficiently high, so that discrete sums over the data set can approximate integrals against $d\mu = p(x)dx$.

The Laplace-Beltrami operator on Γ has a simple expression in normal coordinates. We remind the reader that by choosing an orthonormal basis $e_1, ..., e_d$ of the tangent plane T_x to Γ at x, one defines coordinates on T_x. The image of these coordinates by the exponential map exp_x forms a chart around x on Γ, and the corresponding system is called normal coordinates. Normal coordinates are thus merely a local system of coordinates along orthogonal geodesics. From now on, let $(s_1, s_2, ..., s_d)$ denote these coordinates.

If $f \in C^\infty(\Gamma)$, then the Laplace-Beltrami operator acts on f as:

$$\Delta f = -\sum_{j=1}^{d} \frac{\partial^2 f}{\partial s_j^2} \,.$$

From now on, we make the fundamental assumption that the only objects that are observable are defined in terms of the geometry of the ambient space $\mathbb{R}^n$(the so-called extrinsic geometry) and the distribution $d\mu = p(x)dx$ of the points. The idea here is that in practical situations, the sole quantities that we observe from an experiment or a series of measurements are the multidimensional description of the data and their statistical distribution. For instance, we have access to the Euclidean distance between two points, or it makes sense to compute integrals against $d\mu$ (it is a mere summation over the data points), but we do not have the knowledge of the geodesic distances on Γ. Likewise, the action of the Laplace-Beltrami operator cannot be observed as it is an object of the intrinsic geometry of Γ. Our goal is to show that by using the geometry of the ambient space, we can approximate diffusion-related objects (diffusion kernel, infinitesimal generator) whose definitions rely on the intrinsic geometry only.

We restrict our attention to rotation invariant kernels, *i.e.*, of the form

$$k(x, y) = h(\|x - y\|^2) \,.$$

As already mentioned, Bochner's theorem implies that $u \mapsto h(u^2)$ must be chosen as the Fourier transform of a finite positive measure (a popular choice consists in taking $k(x, y) = e^{-\|x-y\|^2}$). This kind of kernel is calculated from the Euclidean distance between the points, which is known to us. We also introduce a scale parameter: let ε be a positive number, the scale will be represented by $\sqrt{\varepsilon}$. In other words, we consider the following family of kernels indexed by ε:

$$k_\varepsilon(x, y) = h\left(\frac{\|x - y\|^2}{\varepsilon}\right) \,.$$

To simplify the proofs, we assume that h has an exponential decay at infinity and that it is infinitely differentiable. The role played by ε is now clear: this parameter specifies the size of the neighborhoods defining the local geometry of the data. Asymptotically, as $\varepsilon \to 0$, this geometry will coincide with that of the manifold.

In the following, we investigate the asymptotic properties, as $\varepsilon \to 0$, of kernels obtained by normalizing k_ε in various ways. More precisely, we study the asymptotic infinitesimal generator resulting from the graph Laplacian normalization and show that in general, as $d\mu$ is not a multiple of the Riemannian measure on Γ (*i.e.*, the density of the points is not uniform on the manifold), this generator is not the Laplace-Beltrami operator. Indeed,

in [3], Belkin shows that the weighted graph Laplacian on points uniformly sampled on a manifold allows to reconstruct the Laplace-Beltrami operator, but as we show, it fails when the density is non-uniform. We then describe a simple modification of this normalization that handles the case of non-uniform density. In other words, we are able to separate the distribution of the points from the intrinsic geometry of Γ.

In the next section, we establish an asymptotic expansion for the following operator:

$$G_\varepsilon f(x) = \frac{1}{\varepsilon^{\frac{d}{2}}} \int_\Gamma k_\varepsilon(x,y) f(y) dy$$

2.3.2 Technical results

This section is dedicated to the proof of Proposition 7 that will be used in the next section to show how the heat kernel on Γ can be approximated via averaging kernels. The result of Proposition 7 is a Taylor expansion of G_ε in terms of powers of $\varepsilon^{\frac{1}{2}}$. The calculations carried out in this section are similar to those that can be found in [3] or in [31].

For $y \in \Gamma$, we consider the orthogonal projection u of y on T_x. Let $(u_1, ..., u_d)$ be the coordinates of u in $e_1, ..., e_d$ and let $(s_1, ..., s_d)$ be the normal coordinates of y. Let $\gamma_1, ..., \gamma_d$ denote the coordinate geodesics on Γ. Without loss of generality we can assume that the origin 0 belongs to Γ and we choose $x = 0$.

The idea here is to express all quantities as functions of the variable u on the (flat) tangent plane T_0. All we have to do is thus a change of variable in the integral defining our operators. To understand the geometric configuration, it is useful to refer to Figure 2.4.

In the first lemma we show a basic result of differential geometry, namely we give the asymptotic form of the Jacobian matrix of the change of variable $(u_1, ..., u_d) \mapsto (s_1, ..., s_d) = y$:

Lemma 4. *Let $y \in \Gamma$ be in a Euclidean ball of radius $C\sqrt{\varepsilon}$ centered at 0 (C is a positive constant). Then if $i \neq j$*

$$\frac{\partial s_i}{\partial u_j} = \mathcal{O}(\varepsilon^{\frac{3}{2}})$$

and

$$\frac{\partial s_i}{\partial u_i} = 1 + 2a_i^2 u_i^2 + \mathcal{O}(\varepsilon^{\frac{3}{2}})$$

where a_i is the curvature of the geodesic γ_i at 0.

Proof. Let γ be the geodesic between x and y. Since the covariant derivative of the speed vector along γ is zero (by definition of a geodesic), the osculatory plane of γ at 0 is orthogonal to T_0. As a consequence, the deviation of γ from this plane is of order $\varepsilon^{\frac{3}{2}} = (\sqrt{\varepsilon})^3$. Consequently, a small variation du_j of u_j entails a variation of order $\varepsilon^{\frac{3}{2}}$ of s_i. This proves the first equality.

In the osculatory plane, the curve is locally a parabola, and therefore up to the deviation from the plane and terms in $\varepsilon^{\frac{3}{2}}$, we have

$$s_i = \int_0^{u_i} \sqrt{1 + 4a_i^2 v^2} dv + \mathcal{O}(\varepsilon^{\frac{3}{2}})$$

where the $\mathcal{O}$ can be differentiated (the geodesic being C^∞). Eventually, we conclude that

$$\frac{\partial s_i}{\partial u_i} = 1 + 2a_i^2 u_i^2 + \mathcal{O}(\varepsilon^{\frac{3}{2}}) \,.$$

□

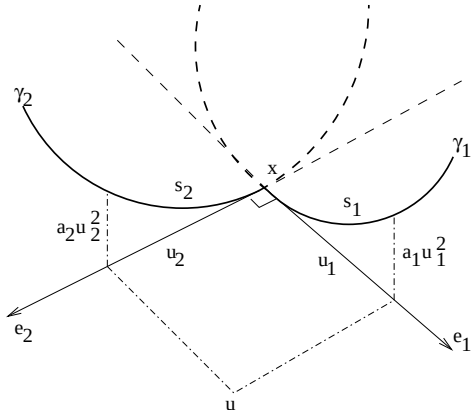

Figure 2.4: Geometric configuration. All terms of order $\varepsilon^{\frac{3}{2}}$ (including the deviation from the osculatory planes) have not been represented.

Corollary 1. *We have*

$$\det(\frac{dy}{du}) = 1 + 2\sum_{i=1}^{d} a_i^2 u_i^2 + \mathcal{O}(\varepsilon^{\frac{3}{2}})\,.$$

The next lemma gives the Taylor expansion of the Euclidean and geodesic distance in the variable u:

Lemma 5. *We have:*

$$\|y(u)\|^2 = \|u\|^2 + \left(\sum_{i=1}^{d} a_i u_i^2\right)^2 + \mathcal{O}(\varepsilon^{\frac{5}{2}})$$

and

$$s_i = u_i + \mathcal{O}(\varepsilon^{\frac{3}{2}})\,.$$

Proof. The first equality is easy to derive from the definition of u and the parabola shape of the γ_i's. The second identity follows from elementary computations: we have seen that

$$s_i = \int_0^{u_i} \sqrt{1 + 4a_i^2 v^2}\, dv + \mathcal{O}(\varepsilon^{\frac{3}{2}})$$

and a Taylor expansion gives the result. □

In the neighborhood of $x = 0$, the Taylor expansion of f is given by

$$f(y) = f(0) + \sum_{i=1}^{d} s_i \frac{\partial f}{\partial s_i}(0) + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} s_i s_j \frac{\partial^2 f}{\partial s_i \partial s_j}(0) + \mathcal{O}(\varepsilon^{\frac{3}{2}})\,.$$

Using the previous lemma we arrive at

Lemma 6.

$$f(y) = f(0) + \sum_{i=1}^{d} u_i \frac{\partial f}{\partial s_i}(0) + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} u_i u_j \frac{\partial^2 f}{\partial s_i \partial s_j}(0) + \mathcal{O}(\varepsilon^{\frac{3}{2}})\,.$$

We now have all the necessary tools to prove the following proposition

Proposition 7. *If $x \in \Gamma \backslash \partial\Gamma$,*

$$G_\varepsilon f(x) = m_0 f(x) + \varepsilon \frac{m_2}{2} \left(E(x) f(x) - \Delta f(x) \right) + \mathcal{O}(\varepsilon^{\frac{3}{2}})$$

where

$$m_0 = \int_{\mathbb{R}^d} h(\|u\|) du$$

$$m_2 = \int_{\mathbb{R}^d} u_i^2 h(\|u\|) du$$

and

$$E(x) = \sum_{i=1}^{d} a_i(x)^2 - \sum_{i=1}^{d} \sum_{j \neq i} a_i(x) a_j(x) \, .$$

Therefore if the density of points is uniform on Γ, the operator G_ε defines an infinitesimal generator of the form Laplace operator + potential, and that this potential is zero when the manifold is a vector subspace of $\mathbb{R}^n$ (in which case the lemma is trivial). Moreover, this proposition shows that the operator G_ε is diagonal up to order ε. We will use this fact in the next section to approximate a particular diagonal operator, namely the Laplace-Beltrami operator. Last, we see that this infinitesimal operator combines information from intrinsic geometry (the Laplace-Beltrami operator) and extrinsic geometry (the curvature potential). Getting rid of the extrinsic information via different normalizations will be one of the goals of the next section.

Proof. The first observation is that due to the exponential decay of h, the domain of integration can be restricted to the intersection of a Euclidean ball of radius $C\sqrt{\varepsilon}$ with Γ. Since $x \notin \partial\Gamma$, and because of lemma 5, the domain of integration can be taken to be the ball $\|u\| < C\sqrt{\varepsilon}$. Thus, up to exponentially small terms,

$$\int_\Gamma h\left(\frac{\|y\|^2}{\varepsilon} \right) f(y) dy \simeq \int_{\|y\| < C\varepsilon} h\left(\frac{\|y\|^2}{\varepsilon} \right) f(y) dy \, .$$

Using lemma 5, we Taylor expand the kernel at $\frac{\|u\|^2}{\varepsilon}$ with respect to the increment $\frac{1}{\varepsilon} \left(\sum_{i=1}^d a_i u_i^2 \right)^2$

$$h\left(\frac{\|y\|^2}{\varepsilon} \right) = h\left(\frac{\|u\|^2}{\varepsilon} \right) + \frac{1}{\varepsilon} \left(\sum_{i=1}^{d} a_i u_i^2 \right)^2 h'\left(\frac{\|u\|^2}{\varepsilon} \right) + \mathcal{O}(\varepsilon^{\frac{3}{2}}) \, .$$

We now invoke corollary 1 and lemmas 5 and 6 to change the variable in the integral defining $G_\varepsilon f(x)$. We obtain:

$$\begin{aligned} \varepsilon^{\frac{d}{2}} G_\varepsilon f(0) &= \int_{\|u\| < C\sqrt{\varepsilon}} \left(h\left(\frac{\|u\|^2}{\varepsilon} \right) + \frac{1}{\varepsilon} \left(\sum_{i=1}^{d} a_i u_i^2 \right)^2 h'\left(\frac{\|u\|^2}{\varepsilon} \right) \right) \\ &\times \left(f(0) + \sum_{i=1}^{d} u_i \frac{\partial f}{\partial s_i}(0) + \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} u_i u_j \frac{\partial^2 f}{\partial s_i \partial s_j}(0) \right) \\ &\times \left(1 + 2 \sum_{i=1}^{d} a_i^2 u_i^2 \right) du + \mathcal{O}(\varepsilon^{\frac{3}{2}}) \, . \end{aligned}$$

The symmetry of k_ε allows to simplify the expression of the leading orders:

- all terms of the kind of $u_i \frac{\partial f}{\partial s_i}(x)$ can be ignored as they are odd and when integrated against even functions they will vanish
- for the same reason, terms like $u_i u_j \frac{\partial^2 f}{\partial s_i \partial s_j}(x)$ can be ignored when $i \neq j$.

We obtain:

$$\begin{aligned}
\varepsilon^{\frac{d}{2}} G_\varepsilon f(0) &= f(0) \int_{\mathbb{R}^d} h\left(\frac{\|u\|^2}{\varepsilon}\right) du - \frac{1}{2}\Delta f(0) \int_{\mathbb{R}^d} h\left(\frac{\|u\|^2}{\varepsilon}\right) u_1^2 du \\
&+ 2f(0) \sum_{i=1}^d a_i^2 \int_{\mathbb{R}^d} h\left(\frac{\|u\|^2}{\varepsilon}\right) u_1^2 du \\
&+ \frac{1}{\varepsilon} f(0) \sum_{i=1}^d \sum_{j=1}^d a_i a_j \int_{\mathbb{R}^d} h'\left(\frac{\|u\|^2}{\varepsilon}\right) u_i^2 u_j^2 du + \mathcal{O}(\varepsilon^{\frac{3}{2}}) \,.
\end{aligned}$$

where the domain of integration was extended to $\mathbb{R}^d$ (exponential decay of h).

$$\begin{aligned}
G_\varepsilon f(0) &= m_0 f(0) - \varepsilon \frac{m_2}{2} \Delta f(0) + 2\varepsilon m_2 f(0) \sum_{i=1}^d a_i^2 \\
&+ \varepsilon f(0) \sum_{i=1}^d \sum_{j=1}^d a_i a_j m_{ij} + \mathcal{O}(\varepsilon^{\frac{3}{2}})
\end{aligned}$$

with

$$m_{ij} = \int_{\mathbb{R}^d} u_i^2 u_j^2 h'\left(\|u\|^2\right) du \,.$$

Now integrations by parts show that $m_{ii} = -\frac{3}{2} m_2$ and if $i \neq j$, $m_{ij} = -\frac{1}{2} m_2$. The proposition results from these identities. □

2.3.3 Asymptotics for the weighted graph Laplacian

We now use the result of Proposition 7 to study asymptotics for the weighted graph Laplacian normalization of the kernel k_ε. In particular, we give the explicit form of the infinitesimal generator, and we show that, in general, it does not coincide with the Laplace-Beltrami operator on the submanifold. The result presented here is a generalization of that of Belkin [3] to non-uniform densities. In particular, we show that the weighted graph Laplacian fails at approximating the Laplace-Beltrami operator in the case of non-uniform densities.

We remind the reader that $p(y)$ is the density function of the measure μ on Γ, *i.e*, $d\mu = p(y)dy$. Let

$$v_\varepsilon^2(x) = \int_\Gamma k_\varepsilon(x,y) p(y) dy \,,$$

and define the averaging operator

$$A_\varepsilon f(x) = \frac{1}{v_\varepsilon^2(x)} \int_\Gamma k_\varepsilon(x,y) f(y) p(y) dy \,.$$

This construction corresponds to viewing the set Γ as a weighted graph with weights of the $k_\varepsilon(x,y)$ and use the graph Laplacian normalization to define A_ε (see section 2.2.1). Note

that the definition of these objects only involves observable quantities. Moreover, $a_\varepsilon(x,y)$ can be put in a symmetric form by considering

$$\widetilde{a}_\varepsilon(x,y) = v_\varepsilon(x) a_\varepsilon(x,y) \frac{1}{v_\varepsilon(y)} .$$

We define the graph Laplacian operator as

$$\Delta_\varepsilon = \frac{I - A_\varepsilon}{\varepsilon} .$$

For $K > 0$, let E_K be the space of all functions $f \in C^\infty(\Gamma)$ such that

- for all multi-index $\alpha = (\alpha_1, ..., \alpha_d)$,

$$\left\| \frac{\partial^{\alpha_1 + ... + \alpha_d} f}{\partial s_1^{\alpha_1} ... \partial s_d^{\alpha_d}} \right\|_2 \leq K^{\alpha_1 + ... + \alpha_d} \|f\|_2$$

- f verifies the Neumann boundary condition: for all $x \in \partial\Gamma$,

$$\frac{\partial f}{\partial \nu}(x) = 0$$

where ν is any tangent vector at x that is normal to $\partial\Gamma$.

Let's explain these two conditions. First, note that for a given $K > 0$, all the estimates given in the previous section are uniform on E_K, that is to say the constants of the $\mathcal{O}$'s are the same for all elements in a ball of E_K. The second condition happens to be the only boundary condition that allows to define a limit operator to Δ_ε. Another useful property of this space is

$$\overline{\bigcup_{K>0} E_K} = L^2(\Gamma) .$$

Proposition 8. *For $f \in E_K$ and $x \in \Gamma \backslash \partial\Gamma$ then*

$$A_\varepsilon f(x) = f(x) + \varepsilon \frac{m_2}{2m_0} \left(\frac{\Delta p(x)}{p(x)} f(x) - \frac{\Delta(pf)(x)}{p(x)} \right) + \mathcal{O}(\varepsilon^{\frac{3}{2}}) .$$

Proof. The idea is to make use of Proposition 7 to obtain asymptotic expansions. We invoke this result to obtain that

$$\int_\Gamma k_\varepsilon(x,y) f(y) p(y) dy = \varepsilon^{\frac{d}{2}} \left(m_0 f(x) p(x) + \varepsilon \frac{m_2}{2} \left(E(x) f(x) p(x) - \Delta(fp)(x) \right) + \mathcal{O}(\varepsilon^{\frac{3}{2}}) \right) .$$

Now plugging-in $f = 1$ yields

$$v_\varepsilon^2(x) = \varepsilon^{\frac{d}{2}} \left(m_0 p(x) + \varepsilon \frac{m_2}{2} \left(E(x) p(x) - \Delta p(x) \right) + \mathcal{O}(\varepsilon^{\frac{3}{2}}) \right) .$$

Taking the ratio gives:

$$A_\varepsilon f(x) = f(x) + \varepsilon \frac{m_2}{2m_0} \left(\frac{\Delta p(x)}{p(x)} f(x) - \frac{\Delta(pf)(x)}{p(x)} \right) + \mathcal{O}(\varepsilon^{\frac{3}{2}}) .$$

□

Corollary 2. *On the space E_K, we have*

$$\lim_{\varepsilon \to 0} \Delta_\varepsilon = H$$

where

$$Hf = \frac{m_2}{2m_0}\left(\frac{\Delta(pf)}{p} - \frac{\Delta p}{p} f\right) = \frac{m_2}{2m_0}\left(\Delta f + 2\left\langle \frac{\nabla p}{p}, \nabla f \right\rangle\right) .$$

Under a conjugation by the density, this operator has the form "Laplacian+potential":

$$pH(\frac{g}{p}) = \frac{m_2}{2m_0}\left(\Delta g - \frac{\Delta p}{p} g\right)$$

where $g = pf$.

Proof. We have to consider two distinct cases, depending on whether x is or is not close to the boundary :

- it can be checked from the proof of Proposition 7 and the first condition imposed on functions of E_K that the result of the previous proposition holds uniformly for all $x \in \Gamma$ at distance from the boundary $\partial\Gamma$ at least equal to $C\sqrt{\varepsilon}$
- if, on the contrary, x is within distance $C\sqrt{\varepsilon}$ from the boundary, then this is where the Neumann condition comes into play:

  $$A_\varepsilon f(x) = \int_\Gamma a_\varepsilon(x,y)(f(y) - f(x))dy + f(x)$$

  since we have an averaging operator, and if $d(.,.)$ is the geodesic distance then

  $$d(y,x) = \mathcal{O}(\sqrt{\varepsilon})$$

  and the Neumann boundary condition implies that

  $$\sup_{d(z,x)<C\sqrt{\varepsilon}} \|\nabla f(z)\| = \mathcal{O}(\sqrt{\varepsilon}) .$$

  We deduce from the mean value theorem that

  $$|f(y) - f(x)| = \mathcal{O}(\varepsilon) .$$

  Here the constants in the $\mathcal{O}$'s do not depend on the point x as Γ is compact. We arrive at

  $$A_\varepsilon f(x) - f(x) = \mathcal{O}(\varepsilon)$$

  if x is at distance less than $C\sqrt{\varepsilon}$ from the boundary.

Combining these two points with the fact that $\mu(\partial\Gamma) = 0$, one can easily conclude that:

$$A_\varepsilon = I - \varepsilon H + \mathcal{O}(\varepsilon^{\frac{3}{2}})$$

on E_K. □

This result proves that when the density is uniform then the limit operator H is equal to a multiple of the Laplace-Beltrami operator on Γ, which was already known (see [3]). From the proof, we see that the normalization allows to get rid of the curvature potential term $E(x)$, but at the price of the introduction of a damping term when the density is not constant. Since

$$Hf = \frac{m_2}{2m_0}\left(\Delta f + 2\left\langle \frac{\nabla p}{p}, \nabla f \right\rangle\right)$$

we see that the damping coefficient is proportional to the relative rate of change of the density. By conjugation with the density, we obtain that H has the form "Laplacian+potential":

$$pH(\frac{g}{p}) = \Delta g - g\frac{\Delta p}{p}$$

where $g = fp$.

Since in most applications, the density is non-uniform, the weighted graph Laplacian method is clearly inappropriate[1] if the goal is to recover the intrinsic geometry of the manifold. We now modify this procedure to handle general densities.

2.3.4 Heat kernel approximation

In the construction of diffusion operators explained in sections 2.2.1 and 2.3.3, the information of the local geometry specified by the kernel k_ε and the distribution of the points on in Γ, given by $d\mu = p(x)dx$, are combined. On the contrary, the Laplace-Beltrami operator is solely defined through the geometry. Therefore, instead of applying the normalization procedure to the kernel $k_\varepsilon(x, y)$, we could rather use the kernel

$$\frac{k_\varepsilon(x, y)}{p(x)p(y)}$$

in order to separate the geometry of Γ from the distribution of the points. In practise, this assumes that p is known, which is often not the case. However, the density can be approximated (up to a multiplication factor) by convolving the kernel with the measure on the set

$$p_\varepsilon(x) = \int_\Gamma k_\varepsilon(x, y)p(y)dy\,.$$

We can now replace k_ε by the kernel

$$\widetilde{k}_\varepsilon(x, y) = \frac{k_\varepsilon(x, y)}{p_\varepsilon(x)p_\varepsilon(y)} \tag{2.4}$$

and proceed as in sections 2.2.1 and 2.3.3 by defining

$$v_\varepsilon^2(x) = \int_\Gamma \widetilde{k}_\varepsilon(x, y)p(y)dy$$

and forming the averaging operator A_ε defined on $L^2(\Gamma)$ by

$$A_\varepsilon f(x) = \frac{1}{v_\varepsilon^2(x)}\int_\Gamma \widetilde{k}_\varepsilon(x, y)f(y)p(y)dy\,.$$

[1] In some situations, it might be desirable to take the distribution of the points into account. Indeed, from a statistical point of view, the information brought by clusters of data points with high density of sample points is more reliable.

Let $a_\varepsilon(x,y)$ be its kernel, and let $a_\varepsilon^{(m)}(x,y)$ be the kernel of A_ε^m. Note that, again, all the quantities involved are observable in the sense given in 2.3.1.

This two step procedure to obtain a diffusion kernel is therefore different from the construction of the graph Laplacian because of the first step that aims at separating the distribution of the data points from the geometry of the underlying manifold. Remark also that dividing by $p_\varepsilon(x)$ in equation (2.4) has no effect in the sense that this factor disappears when one later divides by $v_\varepsilon^2(x)$, however, it has the advantage that the new weight $\widetilde{k}(x,y)$ is symmetric and that consequently, this approach can be cast in the form of a graph Laplacian construction, except that one operates on a modified graph.

Again, we introduce a Laplace operator on Γ by

$$\Delta_\varepsilon = \frac{I - A_\varepsilon}{\varepsilon}$$

acting on the space E_K, where K is a fixed number.

In this section we prove that the operator Δ_ε tends (when acting on E_K) to Δ_0, a multiple of the Laplace-Beltrami operator as $\varepsilon \to 0$. We also show that

$$A_\varepsilon^{\frac{t}{\varepsilon}} = (I - \varepsilon\Delta_\varepsilon)^{\frac{t}{\varepsilon}} \simeq (I - \varepsilon\Delta_0)^{\frac{t}{\varepsilon}} \to e^{-t\Delta_0}\,.$$

Thus, although the family $\{A_\varepsilon\}_{\varepsilon>0}$ does not form a diffusion semigroup, it allows to approximate the heat semigroup of operators $\{e^{-t\Delta}\}_{t>0}$.

Proposition 9. *For $f \in E_K$ and $x \in \Gamma\backslash\partial\Gamma$ then*

$$A_\varepsilon f(x) = f(x) - \varepsilon\frac{m_2}{2m_0}\Delta f(x) + \mathcal{O}(\varepsilon^{\frac{3}{2}})\,.$$

Proof. The approximation of $p(x)$ is defined as

$$p_\varepsilon(x) = \int_\Gamma h\left(\frac{\|x-y\|^2}{\varepsilon}\right) p(y)dy$$

and by Proposition 7,

$$p_\varepsilon(x) = \varepsilon^{\frac{d}{2}} m_0 p(x)\left(1 + \varepsilon\frac{m_2}{2m_0}\left(E(x) - \frac{\Delta p(x)}{p(x)}\right) + \mathcal{O}(\varepsilon^{\frac{3}{2}})\right)\,.$$

Consequently, if

$$\widetilde{G}_\varepsilon f(x) = \int_\Gamma \widetilde{k}_\varepsilon(x,y) f(y) p(y) dy\,,$$

then

$$\begin{aligned}\widetilde{G}_\varepsilon f(x) &= \frac{\varepsilon^{-\frac{d}{2}}}{p_\varepsilon(x)}\int_\Gamma k_\varepsilon(x,y)\frac{f(y)}{m_0}\left(1 - \varepsilon\frac{m_2}{2m_0}\left(E(y) - \frac{\Delta p(y)}{p(y)}\right) + \mathcal{O}(\varepsilon^{\frac{3}{2}})\right)dy \\ &= \frac{1}{p_\varepsilon(x)}\left(f(x) + \varepsilon\frac{m_2}{2m_0}\left(\frac{\Delta p(x)}{p(x)}f(x) - \Delta f(x)\right) + \mathcal{O}(\varepsilon^{\frac{3}{2}})\right)\end{aligned}$$

where we have applied the result of proposition 7. If we plug $f = 1$ in the last equality, we obtain

$$v_\varepsilon^2(x) = \frac{1}{p(x)}\left(1 + \varepsilon\frac{m_2}{2m_0}\frac{\Delta p(x)}{p(x)} + \mathcal{O}(\varepsilon^{\frac{3}{2}})\right)$$

and taking the ratio of $\widetilde{G}_\varepsilon f(x)$ over $v_\varepsilon^2(x)$ yields the result. □

Just like for the graph Laplacian, the operator A_ε can be put in a symmetric form by considering the symmetric kernel

$$\widetilde{a}_\varepsilon(x,y) = \frac{\widetilde{k}(x,y)}{v_\varepsilon(x)v_\varepsilon(y)}\,.$$

Then, it can be checked that

$$a_\varepsilon(x,y) = v_\varepsilon(x)\widetilde{a}_\varepsilon(x,y)\frac{1}{v_\varepsilon(y)}\,.$$

From the previous proposition, we deduce an immediate consequence:

Corollary 3. *On E_K,*

$$\lim_{\varepsilon\to 0}\Delta_\varepsilon = \Delta_0$$

where $\Delta_0 = \frac{2m_0}{m_2}\Delta$.

Proof. The proof is identical to that of corollary 2, when H is replaced by Δ_0. □

We can now prove a result on approximations of the heat kernel:

Proposition 10. *For any $t \in \mathbb{R}$, then on $L^2(\Gamma)$:*

$$\lim_{\varepsilon\to 0} A_\varepsilon^{-\frac{t}{\varepsilon}} = e^{-t\Delta_0}$$

where $e^{-t\Delta}$ is the Neumann heat operator. In other words, the Neumann heat kernel $p_t(x,y)$ on Γ can be approximated by $a_\varepsilon^{\left(\frac{t}{\varepsilon}\right)}(x,y)$.

This result shows that the diffusion of heat on a submanifold can be efficiently computed by properly normalizing a fine Gaussian on the data.

Proof. The idea of the proof is to exploit the fact that the short time heat kernel (for $t = \varepsilon$) on Γ is close to a Gaussian and can therefore be approximated by any bump. Then we use the semi-group property to extend this approximation to large times ($t = \varepsilon + \varepsilon + ... + \varepsilon$).

To simplify we assume that $t = 1$. Observe that it suffices to prove the result on E_K as:

- $$\overline{\bigcup_{K>0} E_K} = L^2(\Gamma)$$
- $(A_\varepsilon)_{\varepsilon>0}$ is uniformly bounded in operator norm by 1 (see lemma 2) on $L^2(\Gamma)$

Consequently, we fix the value of K and we prove the proposition on E_K. In corollary 3, we showed that

$$A_\varepsilon = I - \varepsilon\Delta_0 + \varepsilon^{\frac{3}{2}}R_\varepsilon^{(0)}$$

where $R_\varepsilon^{(0)}$ is bounded on E_K.

If $2^l = \frac{1}{\varepsilon}$, then we need to square A_ε, l times. To do so, we prove by induction that if l is sufficiently large, then if $1 \le m \le l$,

$$A_\varepsilon^{2^m} = (I - \varepsilon\Delta_0)^{2^m} + \varepsilon^{\frac{3}{2}}R_\varepsilon^{(m)} \tag{2.5}$$

with

$$\|R_\varepsilon^{(m)}\| \leq 2^{m+1}\|R_\varepsilon^{(0)}\|\,.$$

Indeed,

$$\begin{aligned} A_\varepsilon^2 &= (I-\varepsilon\Delta_\varepsilon)^2 \\ &= (I-\varepsilon\Delta_0+\varepsilon^{\frac{3}{2}}R_\varepsilon^{(0)})^2 \\ &= (I-\varepsilon\Delta_0)^2+\varepsilon^{\frac{3}{2}}((I-\varepsilon\Delta_0)R_\varepsilon^{(0)}+R_\varepsilon^{(0)}(I-\varepsilon\Delta_0)+\varepsilon^{\frac{3}{2}}R_\varepsilon^{(0)2}) \\ &= (I-\varepsilon\Delta_0)^2+\varepsilon^{\frac{3}{2}}R_\varepsilon^{(1)}\,. \end{aligned}$$

Now observe that by the positivity of Δ_0, if ε is sufficiently small, $\|I-\varepsilon\Delta_0\| \leq 1$ and therefore

$$\|R_\varepsilon^{(1)}\| \leq 2\|R_\varepsilon^{(0)}\|+\varepsilon^{\frac{3}{2}}\|R_\varepsilon^{(0)}\|^2\,.$$

Now if

$$A_\varepsilon^{2^m} = (I-\varepsilon\Delta_0)^{2^m}+\varepsilon^{\frac{3}{2}}R_\varepsilon^{(m)}\,,$$

then

$$A_\varepsilon^{2^{m+1}} = (I-\varepsilon\Delta_0)^{2^{m+1}}+\varepsilon^{\frac{3}{2}}\left((I-\varepsilon\Delta_0)^{2^m}R_\varepsilon^{(m)}+R_\varepsilon^{(m)}(I-\varepsilon\Delta_0)^{2^m}+\varepsilon^{\frac{3}{2}}R_\varepsilon^{(m)2}\right)$$

and the same argument shows that if ε is small enough (independently of m), then

$$\|R_\varepsilon^{(m+1)}\| \leq 2\|R_\varepsilon^{(m)}\|+\varepsilon^{\frac{3}{2}}\|R_\varepsilon^{(m)}\|^2\,.$$

Let $u_m = 2^{-m}\|R_\varepsilon^{(m)}\|$, then

$$u_{m+1} \leq u_m + 2^{m-1-\frac{3}{2}l}u_m^2\,.$$

Suppose that for all $m \leq m_0 \leq l$, $u_m \leq 2u_0$, then by summing the previous inequality, one obtains

$$\begin{aligned} u_{m_0} &\leq u_0 + 2^{-1-\frac{3}{2}l}4u_0^2\sum_{j=0}^{m_0-1}2^j \\ &\leq u_0 + 2^{1-\frac{1}{2}l}2^{m_0-l}u_0^2 \\ &\leq u_0 + 2^{1-\frac{1}{2}l}u_0^2 \\ &\leq 2u_0 \end{aligned}$$

if l is sufficiently large (independently of m_0). We have proved that $u_m \leq 2u_0$ for all $m \leq l$, and equivalently,

$$\|R_\varepsilon^{(m)}\| \leq 2^{m+1}\|R_\varepsilon^{(0)}\|$$

for ε sufficiently small. Noting that $\|R_\varepsilon^{(0)}\|$ remains bounded as $\varepsilon \to 0$, and taking $m = l$ in equation (2.5) yields:

$$A_\varepsilon^{\frac{1}{\varepsilon}} = (I-\varepsilon\Delta_0)^{\frac{1}{\varepsilon}}+2\varepsilon^{\frac{1}{2}}R_\varepsilon^{(0)} \to e^{-\Delta_0}\,.$$

□

The previous proposition shows how to approximate the Neumann heat kernel using fine scale kernels. Since we are interested in using the eigenfunctions and the spectrum of the heat kernel for dimension reduction, we need to know whether the eigenfunctions and eigenvalues of $A_\varepsilon^{\frac{t}{\varepsilon}}$ converge to those of the heat operator. This is indeed the case:

Proposition 11. *The averaging operator A_ε is compact, and one can write*

$$A_{\varepsilon}^{\frac{t}{\varepsilon}} = \sum_{j\geq 0} \lambda_{\varepsilon,j}^{\frac{t}{\varepsilon}} P_{\varepsilon,j}$$

where $P_{\varepsilon,j}$ is the orthogonal projector on the eigenspace corresponding to $\lambda_{\varepsilon,j}$. Furthermore, if the spectral decomposition of the Neumann heat operator is

$$e^{-t\Delta_0} = \sum_{j\geq 0} e^{-t\nu_j^2} P_j$$

where P_j is the orthogonal projector on the eigenspace corresponding to the eigenvalue ν_j^2 of Δ, then we have:

$$\lim_{\varepsilon\to 0} \lambda_{\varepsilon,j}^{\frac{t}{\varepsilon}} = e^{-t\nu_j^2}$$

and

$$\lim_{\varepsilon\to 0} P_{\varepsilon,j} = P_j\,.$$

Proof. It is known that the Neumann heat operator $e^{-t\Delta_0}$ is compact since Γ is a compact manifold (for instance, see a proof in [23]). The same proof can be applied to show the compactness of A_ε. Indeed, the kernel a_ε is C^∞, and therefore so is $A_\varepsilon f$ for $f \in L^2(\Gamma, dx)$. Consequently, Γ being compact, $A_\varepsilon f$ belongs to all Sobolev spaces $H_s(\Gamma, dx)$ for $s \geq 0$. In fact, the derivatives of the kernel being bounded, it follows that A_ε is bounded from $L^2(\Gamma, dx)$ into $H_s(\Gamma, dx)$. Since the injection of $H_s(\Gamma, dx)$ into $L^2(\Gamma, dx)$ is compact when $s > 0$, we conclude that A_ε is a compact operator from $L^2(\Gamma)$ into itself.

A straightforward application of the spectral theorem yields

$$A_{\varepsilon}^{\frac{t}{\varepsilon}} = \sum_{j\geq 0} \lambda_{\varepsilon,j}^{\frac{t}{\varepsilon}} P_{\varepsilon}^{j}\,.$$

Since the heat kernel is compact,

$$e^{-t\Delta_0} = \sum_{j\geq 0} e^{-t\nu_j^2} P_j\,.$$

Now to prove the convergence claims, we refer to classical theorems of spectral perturbation theory. For instance Weyl's theorem asserts that

$$\sup_{j\geq 0} |\lambda_{\varepsilon,j}^{\frac{t}{\varepsilon}} - e^{-t\nu_j^2}| \leq \|A_{\varepsilon}^{\frac{t}{\varepsilon}} - e^{-t\Delta_0}\|\,.$$

A detailed exposition of the main results concerning the perturbation of the singular value decomposition is given in [32]. □

2.3.5 Intrinsic multiscale analysis

Classically, the eigenfunctions of the Laplace-Beltrami operator are viewed as forming a Hilbert basis of $L^2(\Gamma)$, and combining this point of view with simple observations allows us to define an intrinsic multiscale analysis of functions defined on Γ.

The spectral decomposition of the heat kernel is given by

$$p_t(x,y) = \sum_{j\geq 0} e^{-t\nu_j^2} \phi_j(x)\phi_j(y)\,.$$

Since Δ is positive, we have $\nu_j^2 \geq 0$ and in the sum above, the eigenfunctions needed to reconstruct the kernel at time t correspond to the eigenvalues $e^{-t\nu_j^2}$ that are numerically significant. The kernel $p_t(x, y)$, as a function of y, is a bump at the scale $\sqrt{t}$ and centered at x. When t is small, it is close to a very fine Gaussian, and as t increases, the kernel gets coarser and coarser. Thus, in addition to its interpretation as the time, the parameter t also represents a scale, and linear combinations of these bumps at different locations on Γ can be synthesized with only a few eigenfunctions to a good accuracy.

Let $\delta > 0$ be a given accuracy, and let P_t be the orthogonal projector defined by

$$P_t f(x) = \sum_{e^{-t\nu_j^2} > \delta} \langle \phi_j, f \rangle \phi_j(x) \, .$$

The sequence $\{\nu_j^2\}$ having no accumulation point but $+\infty$, when the value of t is doubled, then the number of eigenfunctions defining the projector is roughly divided by 2, whereas the size of the numerical support of $p_t(x, y)$ is approximately doubled. Likewise, the numerical rank of the projector is divided by 2. These observations allow to use the projectors P_t to define a multiresolution analysis of functions of $L^2(\Gamma)$ corresponding to the identity:

$$Id = \sum_{k \in \mathbb{Z}} (P_{2^k t} - P_{2^{k+1} t})$$

and this analysis is only based upon the intrinsic geometry of Γ. In fact, the eigenfunctions $\{\phi_j\}$ constitute the generalization of the Fourier basis to submanifolds, and to each of them is associated a frequency content. Therefore, through these functions it is possible to define a notion of intrinsic frequency and Fourier analysis. The interplay between intrinsic and extrinsic analysis is investigated in the next chapter.

2.3.6 Low-dimensional embedding

So far, we have presented a method for computing the eigenfunctions of the Laplace-Beltrami operator on a submanifold Γ, and we have given these functions the usual interpretation of a Hilbert basis. We now give another point of view on the eigenfunctions, namely we consider the functions as forming a set of coordinates on the submanifold Γ. In addition, we explain how they can be employed for dimensionality reduction.

The spectral decomposition of the Neumann heat kernel on Γ

$$p_t(x, y) = \sum_{j \geq 0} e^{-t\nu_j^2} \phi_j(x) \phi_j(y)$$

allows us to define the distance D_t by

$$D_t^2(x, y) = \sum_{j \geq 0} e^{-t\nu_j^2} (\phi_j(x) - \phi_j(y))^2$$

which becomes small if the amount of heat that has been diffused from x to y at time t is important. Therefore it measures the proximity of points in terms of heat diffusion *assuming the manifold Γ is heat-insulated* (since the eigenfunctions verify the Neumann boundary condition). As previously noticed, this quantity is also equal to the L^2 distance between the bumps $p_t(x, \cdot)$ and $p_t(y, \cdot)$:

$$D_t^2(x, y) = \| p_t(x, \cdot) - p_t(y, \cdot) \|^2 \, .$$

Note that since h is differentiable, for small values of t one has

$$D_t(x,y) \asymp \frac{\|x-y\|}{\sqrt{t}+\|x-y\|}\,.$$

Indeed, a Taylor expansion of the approximation kernel gives $p_t(x,y) \asymp \|x-y\|^2$ when $y \to x$, and $D_t^2(x,y) = p_t(x,x) + p_t(y,y) - 2p_t(x,y) \asymp \|x-y\|^2$. On the other hand, if $\|x-y\| > C\sqrt{t}$, then the diffusion distance is bounded. All constants here depend on the geometry of Γ.

A simple procedure of dimension reduction works as follows: for a given numerical accuracy δ, and a fixed value of t, the dimension m of the embedding should be chosen such that

$$\left| p_t(x,y) - \sum_{0\le j\le m} e^{-t\nu_j^2}\phi_j(x)\phi_j(y) \right| \le \delta\,.$$

We deduce that, just like in Proposition 3, the heat diffusion metric can be efficiently approximated via diffusion maps of the following form:

$$\Phi_m(x) = \begin{pmatrix} \phi_1(x) \\ \phi_2(x) \\ \vdots \\ \phi_m(x) \end{pmatrix}\,.$$

The eigenfunction ϕ_0 is being omitted because it is a constant function if Γ is connected. Therefore, the dimension m of the embedding is such that the diffusion distance can be calculated using a weighted Euclidean metric in this space (in particular, m is greater than or equal to the dimension d of the submanifold). In other words, the description of the data provided by the embedding Φ_m is subject to a global constraint: it approximates the diffusion distance on Γ. But of course, other types of constraints can be imposed on the embedding. In particular, the number m of coordinates can be further reduced if we drop that particular global constraint. For instance, in some situations it might be useful to obtain a piecewise bi-Lipschitz mapping of the data, and possibly with small distortion. Given the importance of the subject (see [18] and [20]), we study this kind of embedding in the next chapter where it can be treated from the more general point of view of positive semi-definite kernels.

2.4 Numerical experiments

In this section, we illustrate the ideas developed so far by numerical examples: we generate sets Γ that are either submanifolds of $\mathbb{R}^n$ or graphs, and we compute the eigenfunctions and eigenvalues of the heat operator using the procedure explained in the previous section. Then we plot the embedding that is obtained from these eigenfunctions. In some cases, we also have compared this embedding with that obtained using the classical weighted graph Laplacian.

These simple experiments underline three advantages in using this method for analyzing the geometry of sets:

- in the simulations, the points of Γ were unordered, and yet they are embedded as a circle, where they are easily reorganized. The eigenfunctions allow to recover the arc

length parametrization of the curve. Although this is not so impressive for curves, things are getting very interesting for submanifolds of higher dimension where the points are naturally reorganized according to the heat flow.

- the technique is completely insensitive to the dimension of the ambient space since the rotation invariant kernels are function of the mutual distances between the points of Γ
- the entire method is fairly robust to noise, as shown in Section 2.4.3.

Concerning the implementation, the heat kernel is approximated using a fine scale Gaussian kernel appropriately normalized as an averaging operator (as explained in Section 2.3.4). Since we are not dealing with continuous data but with finitely many points, all integrals against the empirical measure $p(x)dx$ of the data are computed as discrete sums, *i.e.*,

$$p_\varepsilon(x) = \sum_y k_\varepsilon(x, y)$$

and

$$A_\varepsilon f(x) = \sum_y a_\varepsilon(x, y) f(y)$$

where $a(x, y)$ is obtained from $k(x, y)$ by applying the proper normalization, as described in Section 2.3.4. These summations come down to approximating integrals using the rectangle rule of integration, which does not require to know any kind of ordering of the points. From now on, we choose to use the Gaussian kernel

$$k_\varepsilon(x, y) = e^{\frac{\|x-y\|^2}{\varepsilon}}$$

In matrix notations, the graph Laplacian can be computed as follows:

Weighted graph Laplacian

1. Form the matrix $\mathbf{K}_1$ with entries $\exp(-\frac{\|x_i - x_j\|^2}{\varepsilon})$
2. Set $\mathbf{v} = \text{sqrt}(\mathbf{K}_1 * \mathbf{1})$ where $\mathbf{1} = (1\ 1\ ...\ 1)'$
3. Define $\mathbf{K} = \mathbf{K}_1./(\mathbf{v} * \mathbf{v}')$
4. Diagonalize $\mathbf{K}$ by $[\mathbf{U}, \mathbf{S}, \mathbf{V}] = \text{svd}(\mathbf{K})$
5. The spectrum of the graph Laplacian is that of $\mathbf{K}$ whereas its eigenvectors are given by $\mathbf{U}(:, i)./\mathbf{U}(:, 1)$

We used the Matlab notations where ./ and sqrt are pointwise operations. The choice of the scaling parameter ε is also a matter of concern and we choose ε to be of the order of the average smallest non-zero value of $\|x_i - x_j\|^2$, that is to say, we set

$$\varepsilon = \frac{1}{N} \sum_{i=1}^{N} \min_{j:\, x_j \neq x_i} \|x_i - x_j\|^2$$

The approximation of the eigenfunctions and eigenvalues of the Laplace-Beltrami operator (see Section 2.3.4) are obtained from the diagonalization of the matrix $\mathbf{K}$ constructed as follows:

Approximate Laplace-Beltrami

1. Form the matrix $\mathbf{K}_1$ with entries $\exp(-\frac{\|x_i - x_j\|^2}{\varepsilon})$
2. Set $\mathbf{p} = \mathbf{K}_1 * \mathbf{1}$ where $\mathbf{1} = (1\ 1\ ...\ 1)'$
3. Define $\mathbf{K}_2 = \mathbf{K}_1 ./ (\mathbf{p} * \mathbf{p}')$
4. Set $\mathbf{v} = \mathrm{sqrt}(\mathbf{K}_2 * \mathbf{1})$
5. Define $\mathbf{K} = \mathbf{K}_2 ./ (\mathbf{v} * \mathbf{v}')$
6. Diagonalize $\mathbf{K}$ by $[\mathbf{U}, \mathbf{S}, \mathbf{V}] = \mathrm{svd}(\mathbf{K})$
7. The eigenvalues of Δ are approximated by those of $\mathbf{K}$, and its eigenfunctions ϕ_i are approximated by $\mathbf{U}(:, i) ./ \mathbf{U}(:, 1)$

2.4.1 Curves

Closed curves

We first discuss the case of closed curves in $\mathbb{R}^n$. We assume that Γ is a C^∞ simple curve (it has no double points) of length 1. Since Γ has no boundary, the Neumann heat kernel is merely the heat kernel.

This case is degenerate as from the heat diffusion point of view, all such curves are the same: the amount of heat that has propagated from x to y at a given time t only depends on the initial distribution of temperature and the length of the curve between x and y. Equivalently, every curve is isometric to a circle and the heat kernel is a function of the geodesic distance. As a consequence, all closed simple curves can be identified to a circle of the same length, and for the circle, the eigenfunctions of the Laplace-Beltrami operator are known to be the Fourier basis. For these curves, the heat kernel is

$$p_t(\alpha, \beta) = \frac{1}{\sqrt{4\pi t}} \sum_{j \geq 0} e^{-\frac{(\alpha - \beta + 2\pi j)^2}{4t}} = \sum_{j \in \mathbb{Z}} e^{-j^2 t} e^{2i\pi j(\alpha - \beta)}$$

where α and β are the curvilinear abscissas of two points on Γ. Thus

$$p_t(\alpha, \beta) = 1 + 2 \sum_{j \geq 1} e^{-j^2 t} (\cos(2\pi j \alpha) \cos(2\pi j \beta) + \sin(2\pi j \alpha) \sin(2\pi j \beta))$$

which constitutes the spectral decomposition of this kernel.

This identity shows that for very moderate values of t, only the first terms contribute to this sum, and the heat flow can be accurately computed using the embedding $\alpha \mapsto (\cos(2\pi\alpha), \sin(2\pi\alpha))$. In other words, the curve Γ is mapped onto a circle in the plane. We therefore have shown that the heat metric can be computed on a closed simple curve as the cord length of a circle to any accuracy:

$$e^t D_t^2(x, y) = \sum_{j \geq 1} e^{-(j^2 - 1)t} \left| e^{2i\pi j \alpha} - e^{2i\pi j \beta} \right|^2 = \left| e^{2i\pi\alpha} - e^{2i\pi\beta} \right|^2 \left(1 + e^{-3t} r_t(x, y)\right)$$

where $r_t(x, y)$ is a bounded function.

This is illustrated on Figure 2.5 where two helix curves and a trefoil curve in $\mathbb{R}^3$ are embedded as a circle in $\mathbb{R}^2$. The conclusion of these examples is that no matter how complicated these curves may be, they are immediately re-organized as circles when the kernel is properly normalized. These examples also show that the weighted graph Laplacian embedding is sensitive to the density of the points. In particular, when the density has a steep peak, this embedding tend to map all points around this peak to a single point, creating a corner.

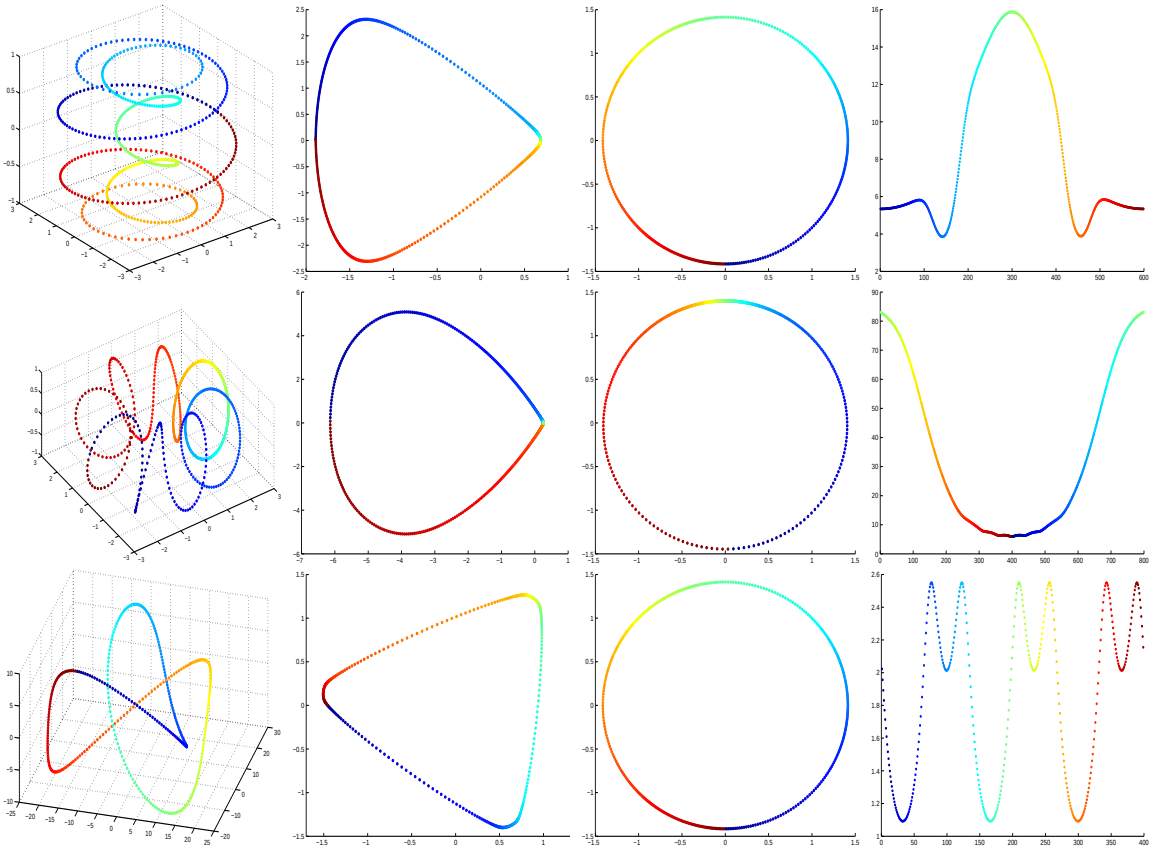

Figure 2.5: Curves in $\mathbb{R}^3$ (two helix curves and the trefoil curve), their embedding using the graph Laplacian (2^{nd} column), their embedding using the correct normalization (3^{rd} column) and the density of points on these curves (last column).

Curves with endpoints

We now consider an example of curve with two endpoints. We studied a sequence of face images from the UMIST Face Database[2], more particularly Γ is a set of 36 pictures of the face of a same person turning his head. Each picture is a pre-cropped 112×92 pixel image in grayscale colormap, and the time sampling rate being sufficiently high (see Figure 2.6), we expect them to be organized along a curve. To recover this point, we proceeded as follows (see Figure 2.7):

[2]Courtesy of Daniel Graham and Nigel Allinson (see [14])

Figure 2.6: In the original file, the pictures are ordered from left to right, and top to bottom

- Initially, the pictures were indexed by the time parameter, or equivalently, by the angle of the head. To illustrate the capabilities of reorganization of the method, we shuffled the set at random so that they appear unordered.
- We computed their mutual distances using the L^2 metric in $\mathbb{R}^{112\times 92}$. Although this metric is generally not adapted to the discrimination of images (see introduction of this chapter), its use yields satisfactory results in this case because the sampling rate in time is sufficiently high and there is very little noise.
- We computed the approximation of the eigenfunctions of the Laplace-Beltrami operator on this structure.

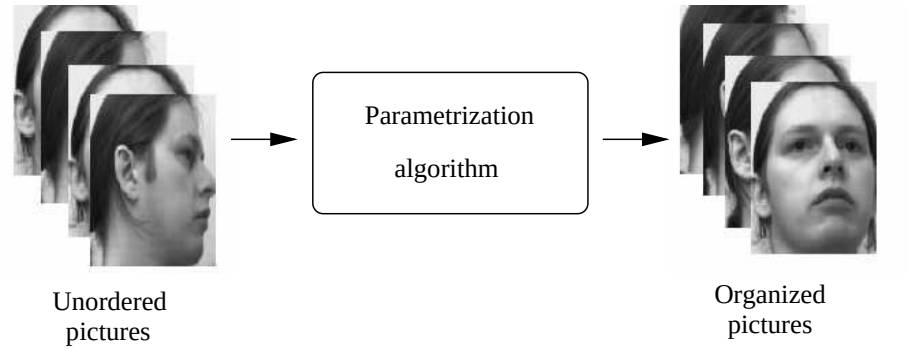

Figure 2.7: The data are completely unordered, and the algorithm reorganizes the sequence of pictures.

From the spectrum, it is clear that, $\lambda_0 = 1$ being ignored, the eigenvalue $\lambda_1 = 0.97$ prevails over the following ones ($\lambda_2 = 0.87$, $\lambda_3 = 0.74$, $\lambda_4 = 0.63$,...) as shown on Figure

2.8. The second eigenfunction ϕ_1 associates a real number to each image, and when this set of numbers is reordered as a non-decreasing sequence and rescaled to have range $[-1, 1]$, we obtain the graph shown on Figure 2.8. It is striking to see how this graph looks like that of half a period of cosine, which is precisely the first non-trivial Neumann eigenfunction of the Laplace-Beltrami on a non-closed curve. Indeed, remember that for a curve with two endpoints and of length L, the Neumann eigenfunctions of Δ are of the form $\cos(j\frac{s}{L})$ where s the arclength variable with $s = 0$ and $s = L$ corresponding to the endpoints.

Therefore the data seem to be approximately lying along a curve in $\mathbb{R}^{112\times 92}$, whose endpoints are easily identified, and ϕ_1 allows to recover the organization of the data with respect to time, or more precisely with respect to the angle of rotation of the face.

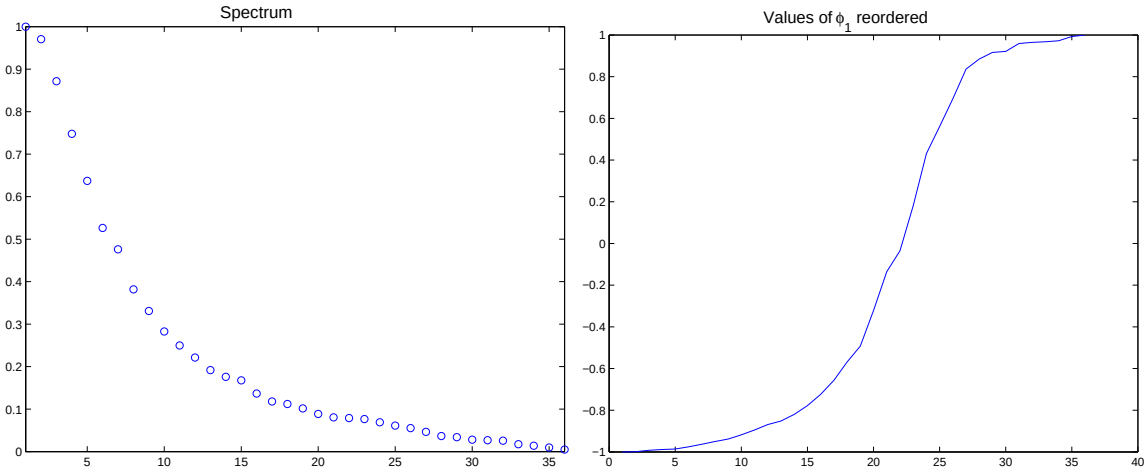

Figure 2.8: Left: spectrum of the Laplace-Beltrami operator. $\lambda_0 = 1$ being ignored, the eigenvalue $\lambda_1 = 0.97$ prevails over the following ones. Numerically, $\lambda_2 = 0.87$, $\lambda_3 = 0.74$, $\lambda_4 = 0.63$. Right: values of ϕ_1 reordered. This graph is very similar to that of a cosine on half a period, which is the second eigenfunction of the Laplace operator on a curve with two endpoints.

2.4.2 Surfaces

We now move on to the case of surfaces. For these submanifolds, and unlike the case of curves, the curvature does play a role in the diffusion of heat. In what follows, we compute the embedding provided by ϕ_1, ϕ_2 and ϕ_3 for different surfaces: an ellipsoid, a torus, a dumbbell shape and a set of images parameterized by two real numbers.

Ellipsoid

The ellipsoid is the simplest closed surface, with a lot of symmetries. Its image via the mapping (ϕ_1, ϕ_2, ϕ_3) of the Laplace-Beltrami operator is an ellipsoid-like shape, although it differs from an actual ellipsoid. For the graph Laplacian, the density on the surface plays a major role. We emphasize this fact on Figure 2.10 where the density is taken to be approximately constant along meridians, but also to be essentially concentrated around the parallel line $\varphi = \pi$. The result is that, just like in the case of curves, the graph Laplacian

Figure 2.9: Left: Set of images randomly permuted. This is the input of the algorithm. Right: output of the algorithm, the sequence is reordered with respect to the angle of rotation of the head (the sequence is to be read from left to right, and top down).

tend to map high density patches into very small patches, creating an edge at the maximum where the density is maximum.

Torus

The torus is an example of non simply connected surface. To implement the embedding, we choose to use ϕ_1, ϕ_2, ϕ_5 and ϕ_6. This choice enables us to represent the torus in cylindrical coordinates:

- eigenfunctions ϕ_1 and ϕ_2 are essentially the cosine and sine of the angle on the big circle of the torus
- eigenfunction ϕ_5 captures the z coordinate
- eigenfunction ϕ_6 essentially computes the distance of the points to the axis of the torus

As is, the embedding $(\phi_1, \phi_2, \phi_5, \phi_6)$ maps the torus as a subset of $\mathbb{R}^4$. A further transformation allows to obtain the cylindrical coordinate representation.

Dumbbell

This time, our set is a dumbbell. It is made up of two large components C_1 and C_2 which are connected by a thin bottleneck. The diffusion between two points of the same component is easy when compared to the heat diffusion between the two components. This is illustrated on Figure 2.12, where the embedding via (ϕ_1, ϕ_2, ϕ_3) tends to separate C_1.

Figure 2.13 shows the eigenfunctions $\phi_1, \phi_2, \phi_3, \phi_4, \phi_5$ and ϕ_6. The second eigenfunction ϕ_1 separates the two sides of the dumbbell, and in fact it is known that it is the solution of the relaxed normalized cut problem for the surface (see [27]).

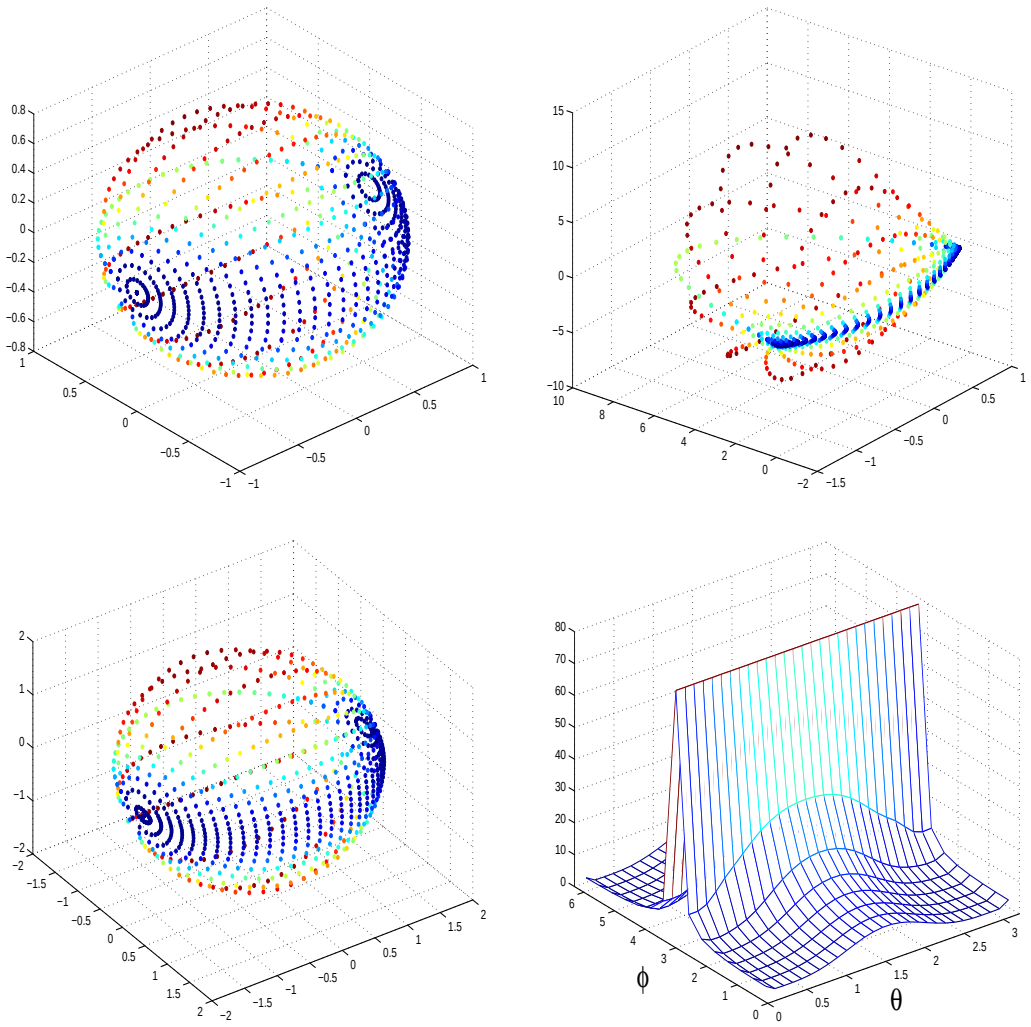

Figure 2.10: Upper left: Original ellipsoid; the colors represent the density. Upper right: Embedding using the graph Laplacian. Lower left: Embedded set using approximate Laplace-Beltrami eigenfunctions. Lower right: density function in the (θ, φ) plane

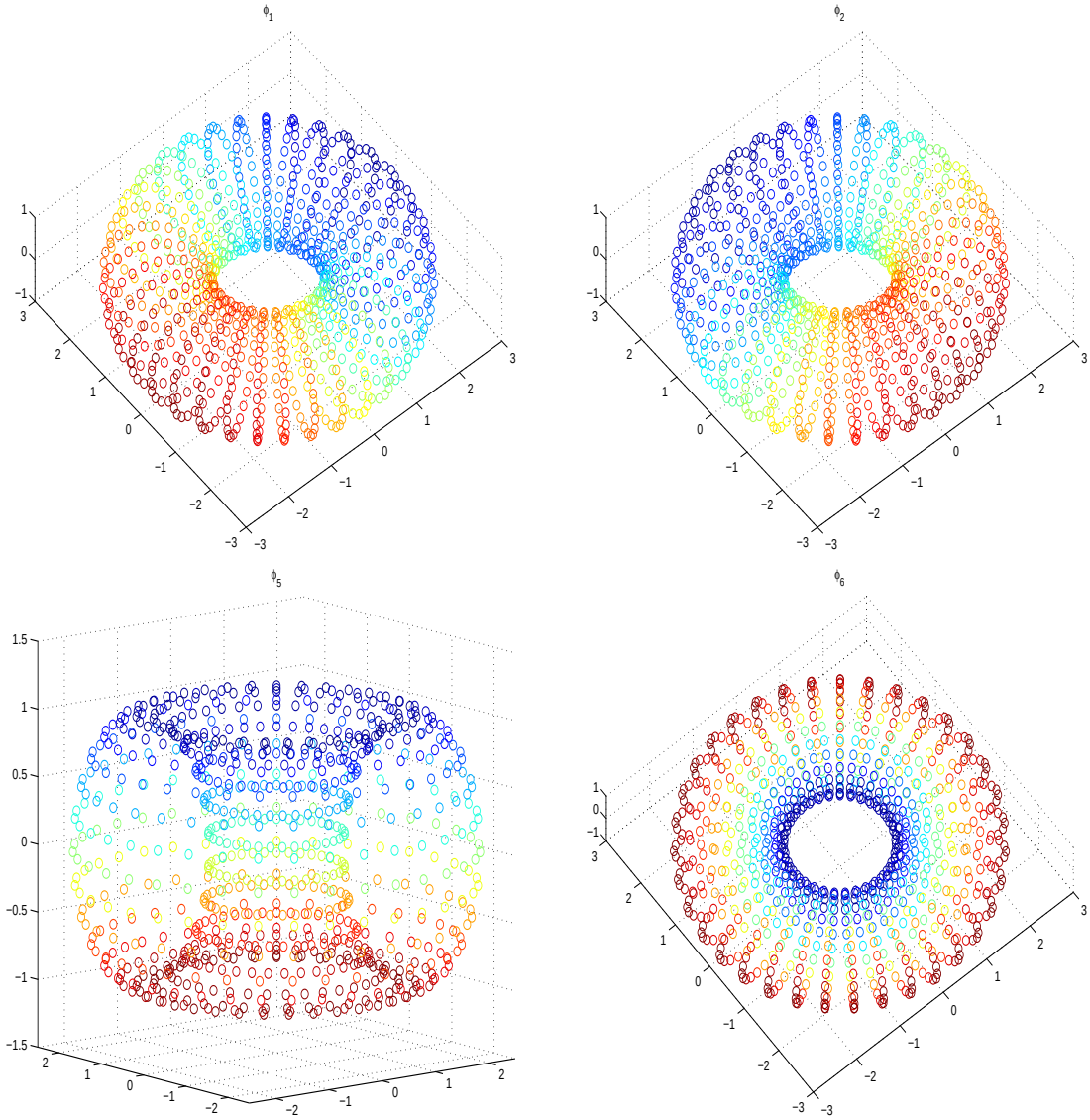

Figure 2.11: Eigenfunctions chosen for the embedding of the torus. ϕ_1 and ϕ_2 (top) measure the angle, ϕ_5 (lower left) computes the z coordinate, and ϕ_6 (lower right) captures the distance to the axis of the torus.

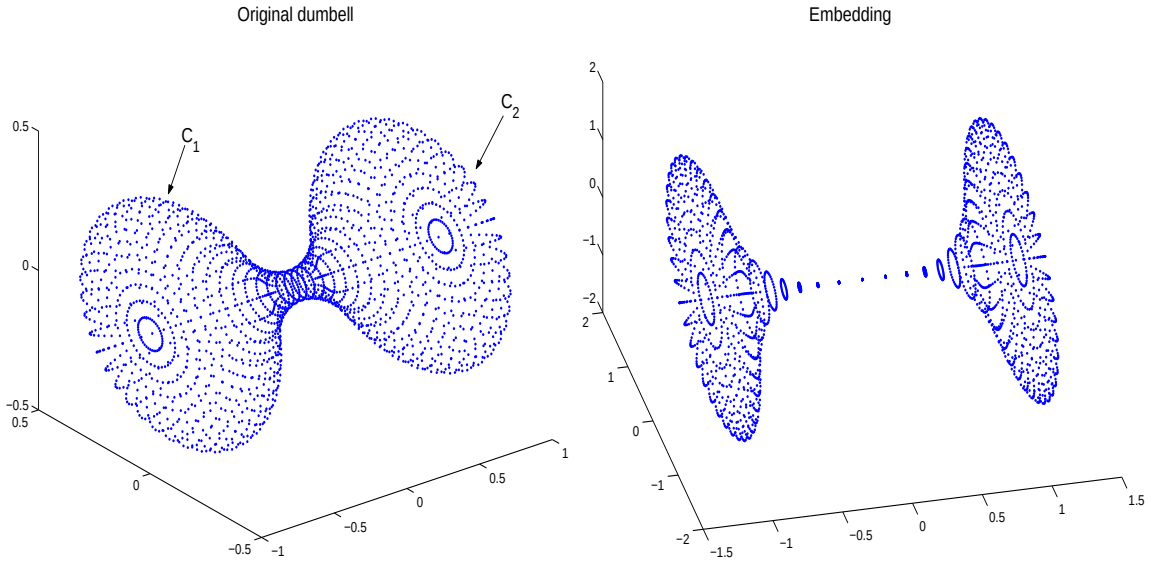

Figure 2.12: Left: dumbbell. Right: Embedded set. The two components are well separated in the embedding space.

Image database

The last example we study for surfaces is that of a database of images parameterized by two real numbers. More precisely, the set Γ is composed of a sequence of 1275 images (75×81 pixels) of the word "3D" viewed under different angles. Each image was generated by a renderer software from a three dimensional model for the two letters "3" and "D", and the object was rotated along the vertical axis (angle α) and horizontal axis (angle β), like shown on Figure 2.14. The data were highly sampled: α was uniformly sampled from -50 to 50 degrees with a step of 2 degrees, whereas β was sampled every 4 degrees from -46 to 50 degrees.

>From the data, we created a graph in which each point is connected with its 8 nearest neighbors. Then each edge (x, y) was assigned the weight $e^{-\frac{\|x-y\|^2}{\varepsilon}}$, and we applied the normalization procedure already described to the resulting kernel. Last we plotted the image of the set by the mapping (ϕ_1, ϕ_2) (Figure 2.14).

The result is that the orientation of the object can be controlled by the two coordinates ϕ_1 and ϕ_2. What appears on Figure 2.14 is that there is a bi-Lipschitz mapping between the angles (α, β) and the couple (ϕ_1, ϕ_2). In other words, the natural parameters of the data set has been recovered by the algorithm.

2.4.3 Robustness to noise

We have already mentioned that diffusion distances exhibit some good robustness to noise perturbation of the data set, and the reason for this being that $D_m(x, y)$ is computed as a sum over all paths joining x and y. This sum has a smoothing effect on small perturbations of the data set.

Concerning the influence of these perturbations on the eigenvalues and eigenfunctions, the answer is provided by classical spectral perturbation theory. In short, the perturbation

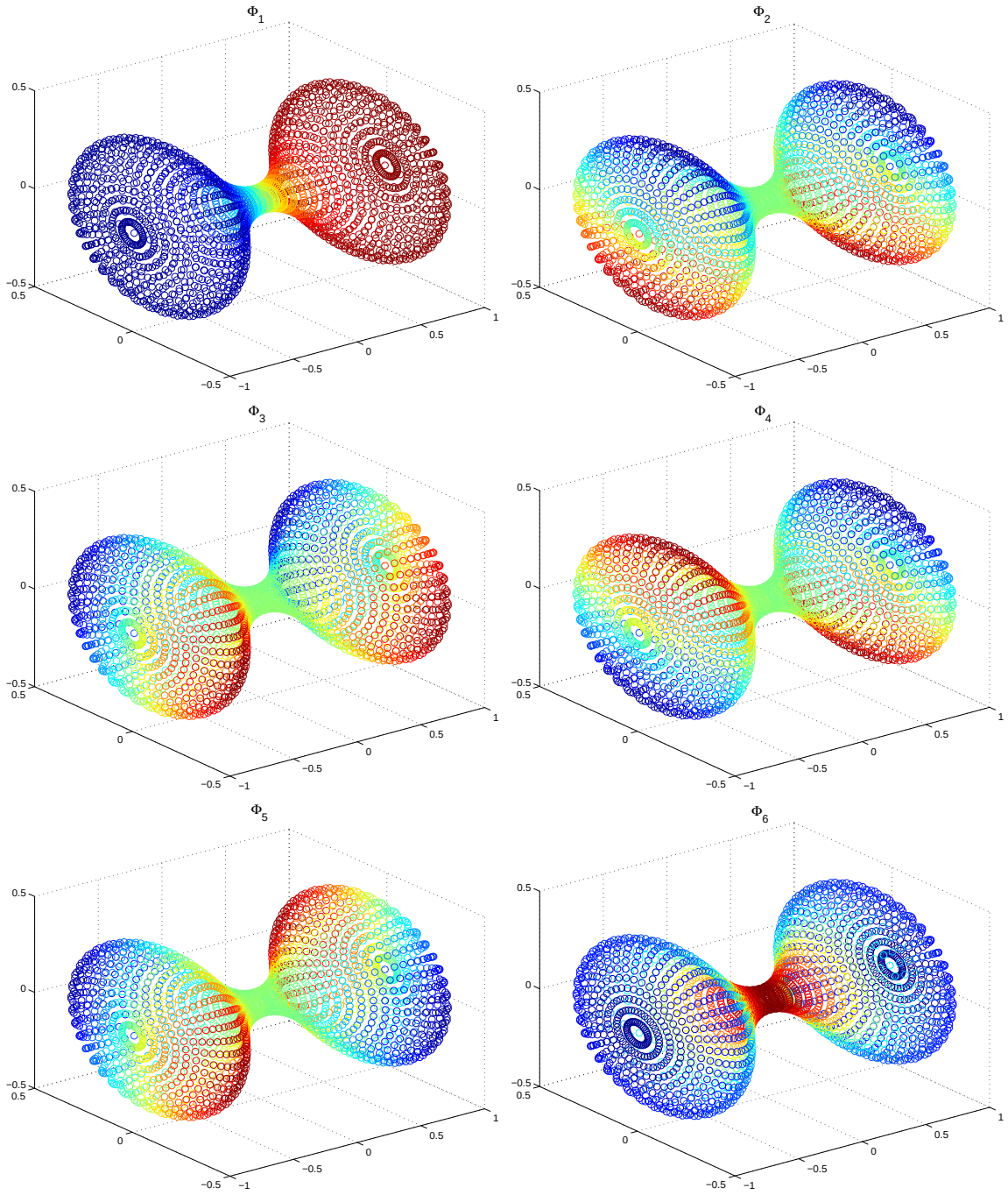

Figure 2.13: Values of the eigenfunctions plotted on the dumbbell. The second eigenfunction ϕ_1 separates the two components C_1 and C_2. It is the solution of the relaxed normalized cut problem. The next four eigenfunctions correspond to a degenerate eigenspace because of the symmetries of the dumbbell.

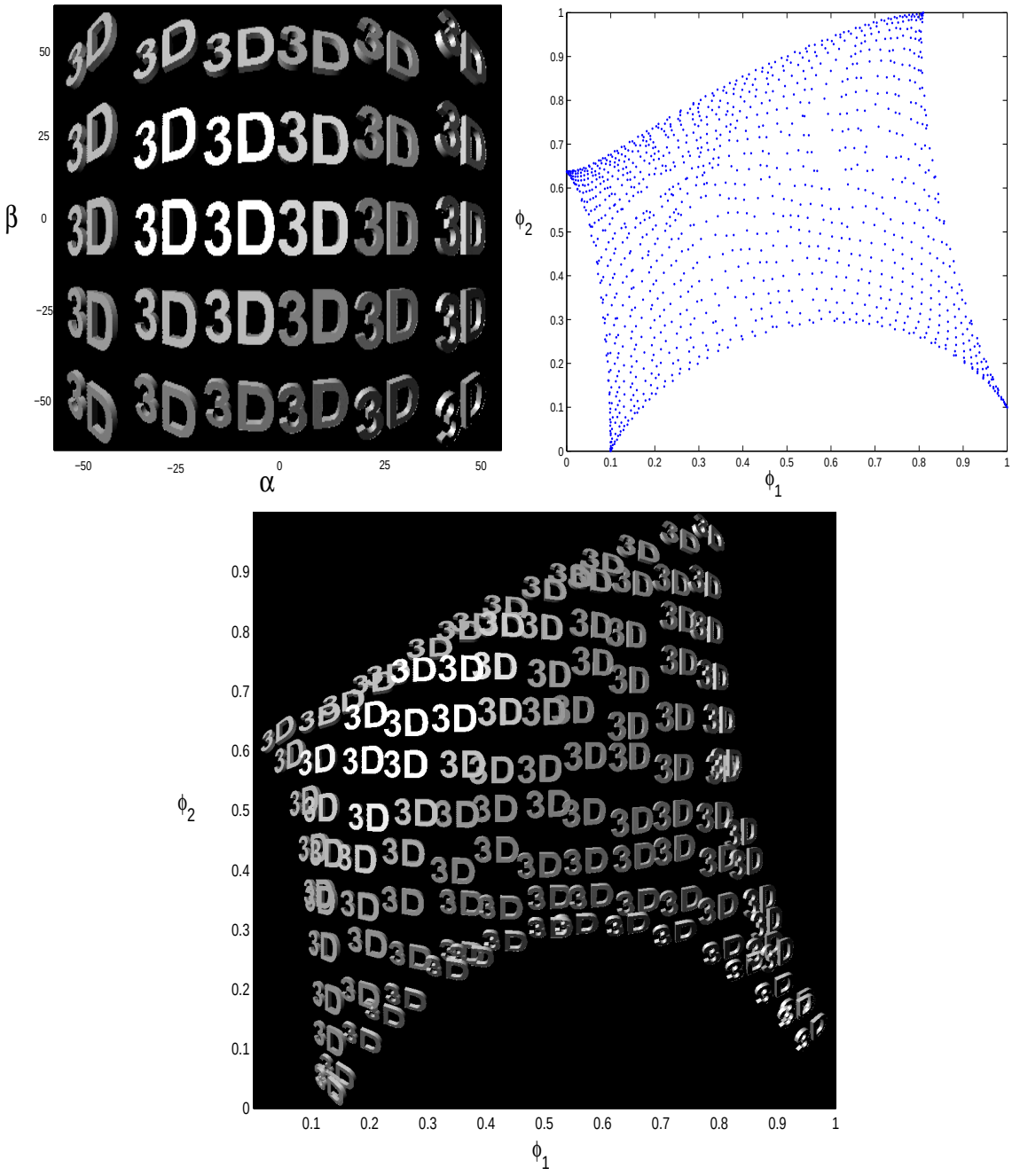

Figure 2.14: Upper left: In the original set, the angle α is discretized 51 times between -50 and 50 degrees, and the angle β is discretized 25 times between -50 and 50 degrees. Upper right: the set is mapped into $\mathbb{R}^2$ via (ϕ_1, ϕ_2). Bottom: some images of Γ are plotted in (ϕ_1, ϕ_2). The natural parameters α and β are recovered.

on the eigenvalues and eigenspaces is controlled by the amplitude of the perturbation on the operators. Remember that Weyl's theorem [32] says that if $\widetilde{A}_\varepsilon$ is a perturbed version of A_ε, with spectrum $\{\widetilde{\lambda}_j\}$ instead of $\{\lambda_j\}$ then

$$\sup_j |\widetilde{\lambda}_j - \lambda_j| \leq \|\widetilde{A}_\varepsilon - A_\varepsilon\|$$

Now if the similarity kernel k_ε is smooth, then a perturbation on the location of the data points can be interpreted as an additive perturbation on the operators by Taylor expanding the kernel and the corresponding operator with respect to the perturbation amplitude. This simple argument shows that the eigenvalues and eigenspaces computation is relatively robust to a perturbation on the data set. However it is to be noted that eigenfunctions corresponding to degenerate eigenvalues are not stable. Nevertheless, in this case, the degeneracy of the eigenspace reveals some symmetry in the data space, and the choice of any set of orthogonal eigenfunction makes sense.

In conclusion, methods of classification or clustering based on the diffusion metric will not be subject to instabilities related to sensitiveness to the particular realization of the data set.

In Figure 2.15, we illustrate our point with the computation of the embedding of a perturbed version of the helix used in Figure 2.5.

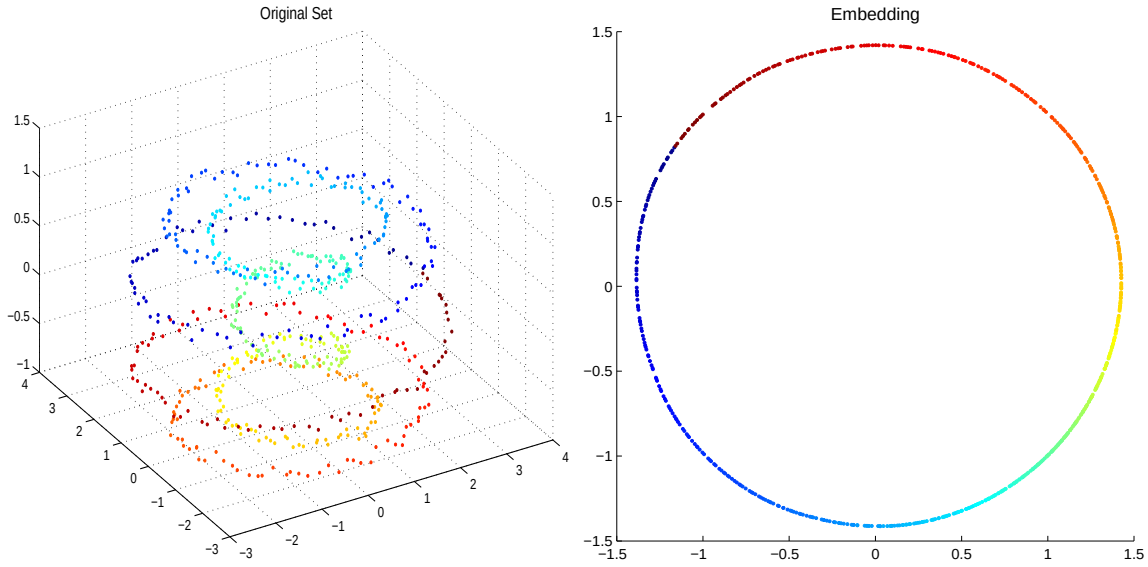

Figure 2.15: Left: the helix of Figure 2.5 was perturbed by an additive Gaussian white noise. Right: the data is still approximately mapped onto a circle.

2.5 Doubling of manifolds and Dirichlet heat kernel

As shown in section 2.3.4, when Γ is a submanifold of the Euclidean space, the Neumann heat kernel can be approximated by properly normalizing a fine Gaussian on the data. An interesting feature of this method is that it is completely blind to the location of the boundary: we do not need to know which point are in a neighborhood of the boundary. In

the case when the boundary $\partial\Gamma$ is known to us, then a simple operation can be performed to obtain the *Dirichlet* heat kernel in addition to the Neumann one, namely we double the submanifold into a closed manifold that no longer stands in $\mathbb{R}^n$, but has the advantage to have no boundary. The way this can be done is by generating a copy Γ_- of $\Gamma = \Gamma_+$, and by identifying points on $\partial\Gamma_-$ with those on $\partial\Gamma_+$. To complete our construction, we need to

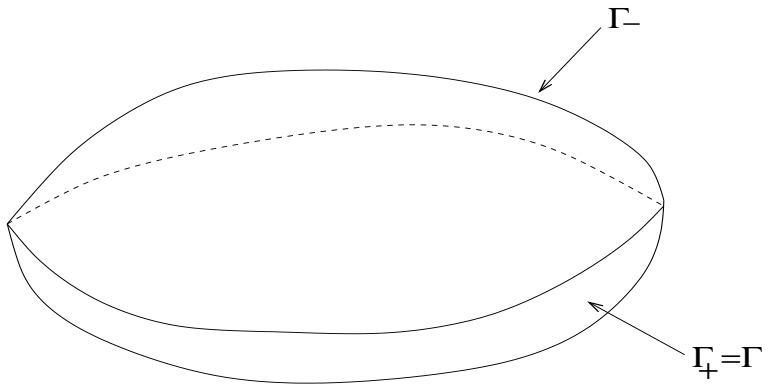

Figure 2.16: The original manifold $\Gamma = \Gamma_+$ is doubled by generating a copy Γ_- of it and by identifying points on the boundary of each copy.

specify a metric on $\Gamma_+ \cup \Gamma_-$. The simplest way is probably the following one: for x and y on $\Gamma_+ \cup \Gamma_-$, we define the distance $d(x, y)$ between x and y to be the length of the shortest path between these points. Thus if x and y belong to the same component, then $d(x, y)$ is the usual geodesic distance. If on the contrary, they are on opposite components, then $d(x, y)$ can be defined as the infimum[3]:

$$d(x, y) = \inf_{z \in \partial\Gamma_+} (d(x, z) + d(z, y))$$

We can now form the kernel

$$k_\varepsilon(x, y) = e^{-\frac{d(x,y)^2}{\varepsilon}}$$

and normalize it as usual to obtain the diffusion kernel a_ε. Because of the obvious symmetry of $\Gamma_+ \cup \Gamma_-$, the eigenspaces corresponding to non-constant functions are degenerate: each of them is the direct sum of the space of even eigenfunctions (with respect to the boundary) with that of odd eigenfunctions. As $\varepsilon \to 0$, the even eigenfunctions tend to the Neumann heat eigenfunctions, whereas *the odd eigenfunctions tend to the Dirichlet heat eigenfunctions.*

The method described here is thus a natural extension of that presented in the previous sections, and if the boundary is known to us, it constitutes a simple way to obtain all eigenfunctions of the heat operator on Γ, regardless of the nature of the boundary condition (Dirichlet/Neumann).

2.6 Anisotropic diffusions

So far, we have mainly employed rotation invariant kernels, but in several situations, it is desirable to define other types of diffusions. Rotation invariant kernels generate isotropic

[3]Note that the computation of $d(x, y)$ is to be carried out only for x and y sufficiently close as the kernel k_ε is localized.

diffusions (like the homogeneous heat diffusion) for which all directions are equal, and in particular all variables in the data play the same role. But of course, it can sometimes be useful to design anisotropic diffusions

- to deal with ambiguous geometric configurations, like crosses, forks, and other bifurcations,
- to take advantage of some a priori knowledge on the data, like the fact that locally, some variables might be more important than others.

The main ingredient in order to construct an anisotropic diffusion is to be able to locally separate variables.

We focus on 2 possible applications of anisotropic diffusions: dealing with crosses and forks, and the analysis of differential systems.

2.6.1 Incomplete data and ambiguous geometries

It is not uncommon for a data set to exhibit structures like forks, and the reason for this is that in several cases, the variables used in the multidimensional representation of the sample do not describe entirely the data because of some loss of information inherent to the data acquisition process. In other words, in their high-dimensional representation, samples are already the result of some projection and consequently, crosses appear in the data (see Figure 2.17). An isotropic random walker along a curve arriving at a fork will choose equally between the several branches, and this behavior is not faithful to the original geometry of the data (before the projection). However, there are several ways to redefine the local geometry of Γ in order to force the diffusion to recognize such a situation and to favor one of the possible paths. A very basic idea is to modify the Euclidean geometry by adding other features than just the coordinates of the points. An example of of modified distance that allows to deal with crosses is to take the new distance to be the sum of the Euclidean distance and a distance between tangent planes.

For instance, at each point of the set, one can perform a local Principal Component Analysis and use the singular values and principal axes to rescale the local metric. More precisely, suppose that Γ is the union of the two coordinate axes in the plane $\mathbb{R}^2$. This set has a cross at the origin. Now, for a given $\varepsilon > 0$, and for all $x \in \Gamma$, compute the PCA of all points in the Euclidean ball of center x and radius $\sqrt{\varepsilon}$. The principal axes are parallel to the coordinate axes and let σ_1^2 and σ_2^2 denote the singular values along the horizontal and vertical axes. Suppose that $x = (u, v)$ and $y = (w, z)$, and let the new distance between x and y be

$$d(x,y)^2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}(u-w)^2 + \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}(v-z)^2$$

Define the kernel

$$k_\varepsilon(x,y) = e^{-\frac{d(x,y)^2}{\varepsilon}}$$

When $|u| > \sqrt{\varepsilon}$, nothing changes compared to a classical Gaussian kernel (with the Euclidean distance). On the contrary, as x gets closer to the origin, this kernel gives higher probability to jumps to points on the same axis as x than the orthogonal axis. For more general sets, this kernel will encourage connections between points that are locally aligned, the reason for this being that the distance $d(x, y)$ is in fact the sum of a distance between the points and a distance between the tangent planes attached to these points.

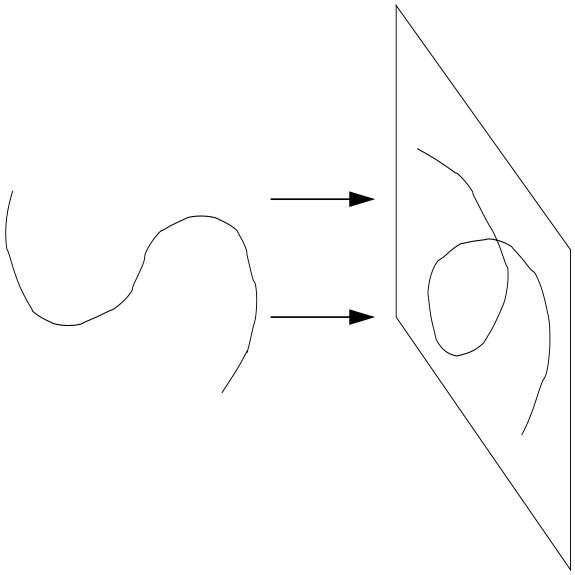

Figure 2.17: Projections create loops.

A different way to proceed is as follows. Define $d_\Gamma(x)$ to be the distance of x to the set Γ. Assuming this is well defined, this is a local computation. For $\varepsilon > 0$, consider the kernel

$$k_\varepsilon(x,y) = \exp\left(-\frac{\|x-y\|^2}{\varepsilon} - \frac{d_\Gamma\left(\frac{x+y}{2}\right)^2}{\varepsilon^2}\right)$$

Just like the previous one, this kernel will favor transitions between points locally aligned.

2.6.2 Differential systems

Another important application of anisotropic diffusions is the study of differential systems. Suppose that the gradient ∇f of a function $f : \Omega \to \mathbb{R}$ is known at all points, and that the domain Ω is compact. To recover f up to a constant, we need to integrate a differential system. Another approach consists in designing a diffusion adapted to this problem, namely by considering the kernel

$$k_\varepsilon(x,y) = \exp\left(-\frac{\|x-y\|^2}{\varepsilon} - \frac{(\langle \nabla_x f, x-y\rangle)^2}{\varepsilon^2}\right)$$

Proposition 12. *Suppose that $f : \Omega \to \mathbb{R}$ is C^2 and has no critical point, i.e., at all point, $\nabla f \neq 0$. Then if g is C^2, then at all point x not on the boundary of Ω,*

$$A_\varepsilon^{\frac{t}{\varepsilon}} g(x) \to e^{-\alpha t \Delta_c} g(x)$$

where Δ_c is the Laplace-Beltrami operator on the curve

$$\Gamma_c = \{y : f(y) = c = f(x)\}$$

and α is some constant independent of x.

The diffusion defined by the kernel k_ε properly renormalized identifies the level sets of f. By iterating the kernel, we allow it to expand by steps of $\sqrt{\varepsilon}$ along the level set and ε in the orthogonal direction. Doing it $\frac{1}{\varepsilon}$ times will force the random walks to remain at distance $\sqrt{\varepsilon}$ from this level set.

Proof. All we need to prove is that as $\varepsilon \to 0$, the operator

$$G_\varepsilon g(x) = \iint_\Omega e^{-\frac{\|x-y\|^2}{\varepsilon}} e^{-\frac{\langle \nabla f, x-y\rangle^2}{\varepsilon^2}} g(y) dy$$

acts like the operator

$$G^c_\varepsilon g(x) = \int_{\Gamma_c} e^{-\frac{\|x-y\|^2}{\varepsilon}} g(y) d\sigma(y)$$

where $d\sigma(y)$ is the Lebesgue measure on the curve Γ_c.

The key point is to locally separate variables. To simplify, suppose that $x = (0,0)$ and $y = (y_1, y_2)$ where the axes of coordinates are taken to be respectively orthogonal and parallel to ∇f at x. In other words, $\nabla f = (0, \|\nabla f\|)$. Up to exponentially small terms, we have:

$$\begin{aligned} G_\varepsilon g(x) &= \int_{|y_2|<C\varepsilon} e^{-\frac{y_2^2 \|\nabla f\|^2}{\varepsilon^2}} \int_{|y_1|<C\sqrt{\varepsilon}} e^{-\frac{y_1^2+y_2^2}{\varepsilon}} g(y_1, y_2) dy_1 dy_2 \\ &= \frac{\varepsilon}{\|\nabla f\|} \left(m_0 \int_{|y_1|<C\sqrt{\varepsilon}} e^{-\frac{y_1^2}{\varepsilon}} g(y_1, 0) dy_1 + \mathcal{O}(\varepsilon^2) \right) \end{aligned}$$

where we have applied the trivial version of lemma 7 along the (flat) lines $y_1 = \textit{constant}$. Now up to terms of order $\varepsilon^{\frac{3}{2}}$, integrating on the segment $\{(y_1, 0) : |y_1| < C\sqrt{\varepsilon}\}$ is identical to integrating on the portion of Γ_c defined as $\{(y_1, y_2) : f(y_1, y_2) = f(x), |y_1| < C\sqrt{\varepsilon}\}$, in other words,

$$G_\varepsilon g(x) = \frac{\varepsilon}{\|\nabla f(x)\|} \left(m_0 \int_{\Gamma_c} e^{-\frac{\|x-y\|^2}{\varepsilon}} g(y) d\sigma(y) + \mathcal{O}(\varepsilon^{\frac{3}{2}}) \right)$$

where we have extended the integration to the whole curve Γ_c by neglecting exponentially small terms. This means that using k_ε on the domain Ω is approximately equivalent to using the isotropic Gaussian kernel $e^{-\frac{\|x-y\|^2}{\varepsilon}}$ with the function g restricted to Γ_c. This approximation holds up to a term of order $\varepsilon^{\frac{3}{2}}$. We can conclude that all the normalizations described before and applied to k_ε will yield the corresponding normalized isotropic Gaussian kernels restricted to the curve Γ_c. In particular, if we apply the normalization for approximating the heat kernel to k_ε, we will obtain that

$$A_\varepsilon g(x) = g(x) + \alpha \varepsilon \Delta_c g(x) + \mathcal{O}(\varepsilon^{\frac{3}{2}})$$

□

This simple example shows that by defining an appropriate local metric, we are able to construct a diffusion that

1. integrates a vector field,
2. allows to compare the trajectories.

The diffusion framework therefore seems to have a great potential for addressing differential equations and dynamical systems.

Chapter 3

Geometric Harmonics

In the previous chapter, we showed how positive semi-definite kernels could be used to analyze the *intrinsic geometry* of a set Γ. We obtained a set of functions that we interpreted as coordinates on the set, as well as a basis of expansion for functions defined on Γ. Although we focused on the case $\Gamma \subset \mathbb{R}^n$, the technique works for abstract sets that are not necessarily subsets of an ambient space.

However, in some situations like those occurring in statistical learning, such an ambient space does exist, and not only is it important to learn the geometry of Γ (viewed as the training set), but it is also essential to be able to extend functions defined on Γ to a neighborhood of this set. This kind of situation is illustrated by the paradigm training-prediction dear to statistical regression theory: imagine that one needs to predict some quantity of interest $f(x)$ associated with a sample point x. One usually proceeds as follows: one first adjusts the parameters of a regression model for f using the information at one's disposal (the training set Γ and the training values for f), this step is termed training or calibration, and then one uses this model to predict the value $f(x)$ for any new point $x \notin \Gamma$. The model chosen for f corresponds to a certain number of constraints, and is essentially arbitrary, unless one has a priori knowledge on the data. We do not address the problem of model selection here, our goal being to explain how one can naturally extend such a function f to a neighborhood of Γ when we are given a model in the form of a functional class, and how the *extrinsic geometry* of the set imposes constraints on the feasibility of this extension.

In what follows, the functional model for f will be represented by a kernel $k(x, y)$, namely f will be a function of the form

$$\int_\Gamma k(x, y)g(y)d\mu(y)$$

for an appropriate g. This means that f is restricted to belong to the space of linear combinations of the kernel k. When k is a semi-definite positive kernel, the Nyström method will allow us to extend f outside the set Γ using a special set of functions that we term *geometric harmonics*.

This chapter is organized as follows: we start by reviewing the notion of semi-definite positive kernels and their interpretation as projectors on a reproducing kernel Hilbert space. We then give the definition of the geometric harmonics and list two of their main properties. We use these properties to design a simple extension algorithm for empirical functions defined on the data set Γ and we illustrate this technique with some examples. We then investigate the subject of bi-Lipschitz parametrization of data sets and how geometric harmonics offer a simple approximate solution to this problem. Last, we describe the interplay

of the intrinsic and extrinsic geometry and we show that the geometric harmonics allow to perform multiscale extensions of empirical functions on the data set.

Throughout this chapter, Γ is a compact subset of $\mathbb{R}^n$, and is endowed with a finite positive measure μ. Capital letters are reserved for functions of one variable defined on Ω and constants, while lower case letters denote functions defined on Γ.

3.1 Positive kernels and associated reproducing kernel Hilbert spaces

This section quickly reviews some very basic notions about reproducing kernel spaces, and a more detailed description of the subject can be found in [2]. Let Ω be a subset of $\mathbb{R}^n$ containing Γ and let k be a positive semi-definite kernel defined on $\Omega \times \Omega$. Remember that this means that for any $m \geq 1$ and any choice of real numbers $\alpha_1, ..., \alpha_m$ and points $x_1, ..., x_m$ in Ω,

$$\sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) \geq 0 \,.$$

As shown in [6], this condition can be replaced by

$$\int_\Omega \int_\Omega \alpha(x)\alpha(y)k(x,y)dxdy \geq 0 \,.$$

It is a classical result (see [2]) that one can associate to k a Hilbert space $\mathcal{H}$ of functions defined on Ω, in which k defines the inner product:

- for dx-almost every $x \in \Omega$, $k(x, \cdot)$ belongs to $\mathcal{H}$,
- for dx-almost every $x \in \Omega$, $\langle f, k(x, \cdot)\rangle_\Omega = f(x)$ where $\langle \cdot, \cdot \rangle_\Omega$ defines the inner product on $\mathcal{H}$.

The construction of the space $\mathcal{H}$ is described in [2]. In short, one has to consider finite linear combinations of the kernels and take the completion. $\mathcal{H}$ is called a reproducing kernel Hilbert space, and k is said to be a reproducing kernel satisfying the identity

$$\langle k(x, \cdot), k(y, \cdot)\rangle_\Omega = k(x, y) \,.$$

Conversely, it is easy to see that any reproducing kernel is positive semi-definite. Therefore the two notions are identical.

For example, let $k(x, y) = h(x - y)$ and $\Omega = \mathbb{R}^n$. Then by Bochner's theorem, h is the inverse Fourier transform of a finite positive measure, and suppose for simplicity that this measure is of the form $\widehat{h}(\xi)d\xi$. The space $\mathcal{H}$ is the inverse Fourier transform of

$$\left\{ \widehat{f} \text{ such that } \int_{\widehat{h}(\xi)>0} |\widehat{f}(\xi)|^2 \frac{d\xi}{\widehat{h}(\xi)} < +\infty \text{ and where } \widehat{f}(\xi) = 0 \text{ if } \widehat{h}(\xi) = 0 \right\} .$$

Thus $\mathcal{H}$ is the inverse Fourier transform of a weighted L^2 space and since $\widehat{h}$ is integrable, the weight penalizes high frequencies and elements of $\mathcal{H}$ are smooth functions. Equivalently, f is equal to the convolution $\rho * h$ of a signed measure ρ with h so that

$$\int_{\mathbb{R}^n} \int_{\mathbb{R}^n} h(x - y)d\rho(y)d\rho(x) < +\infty \,.$$

Two cases are particularly interesting:

- for the Gaussian kernel $k_t(x,y) = e^{-\frac{\|x-y\|^2}{t}}$, the corresponding space is the set of all temperature distributions at time t,
- when $\widehat{h}$ is the indicator function of some bounded Borel set B, the space $\mathcal{H}$ is a set of bandlimited functions. In the case when B is a ball, the corresponding kernel will be referred to as Bessel kernel (see appendix A).

To summarize, one can say that to each positive semi-definite kernel k there corresponds a Hilbert space $\mathcal{H}$ of smooth functions defined on Ω, and that in this space

$$\langle k(x,\cdot), f\rangle_\Omega = f(x)\,.$$

3.2 Definition of the geometric harmonics

We now define the geometric harmonics. Let k be a symmetric positive semi-definite kernel on $\Omega\times\Omega$, and consider the operator $\mathbf{K} : L^2(\Gamma, d\mu) \longrightarrow \mathcal{H}$ defined by

$$\mathbf{K}f(x) = \int_\Gamma k(x,y)f(y)d\mu(y)$$

where $x\in\Omega$. Then we have the following lemma

Lemma 13. *The adjoint $\mathbf{K}^* : \mathcal{H} \longrightarrow L^2(\Gamma, d\mu)$ is the restriction operator onto the set Γ:*

$$\mathbf{K}^*g(y) = g(y) \text{ if } y\in\Gamma \text{ and } g\in\mathcal{H}\,.$$

Furthermore, if k is bounded then the operator $\mathbf{K}^*\mathbf{K} : L^2(\Gamma,d\mu)\to L^2(\Gamma,d\mu)$ is compact.

Proof. It can be checked that $\mathbf{K}^*g(y) = \langle k(y,\cdot), g\rangle_\Omega$ and the first assertion is a straightforward equivalence from the reproducing kernel identity. Now we have

$$\mathbf{K}^*\mathbf{K}f(x) = \int_\Gamma k(x,y)f(y)d\mu(y)\,,\ x\in\Gamma$$

to prove the compactness of $\mathbf{K}^*\mathbf{K}$, we prove that this operator is Hilbert-Schmidt, i.e. that

$$\int_\Gamma\int_\Gamma k(x,y)^2 d\mu(y)d\mu(x) < +\infty\,.$$

The Cauchy-Schwarz inequality in $\mathcal{H}$ implies that

$$k(x,y) = \langle k(x,\cdot), k(\cdot,y)\rangle_\Omega \le \sqrt{k(x,x)}\sqrt{k(y,y)}$$

and this entails that

$$\int_\Gamma\int_\Gamma k(x,y)^2 d\mu(y)d\mu(x) \le \left(\int_\Gamma k(x,x)d\mu(x)\right)^2 \le \left(\mu(\Gamma)\sup_{x\in\Gamma}k(x,x)\right)^2$$

which concludes the proof. □

The operator $\mathbf{K}^*\mathbf{K}$ is self-adjoint and positive, and since it is also compact, it admits a discrete set of eigenfunctions $\{\psi_j\}$ and non-negative eigenvalues $\{\lambda_j\}$. Furthermore, since the operator $\mathbf{K}^*$ is the restriction operator to Γ, the eigenfunctions and eigenvalues are obtained by diagonalizing the kernel k on Γ:

$$\int_\Gamma k(x,y)\psi_j(y)d\mu(y) = \lambda_j\psi_j(x) \text{ if } x \in \Gamma$$

So far, we have arrived at a basis $\{\psi_j\}$ of functions defined on Γ. We now describe how these functions can be extended to the whole Ω. The idea is to use a technique known as the Nyström method ([21], 18.1): if $\lambda_j > 0$ and $x \in \Omega$, define $\Psi_j(x)$ by

$$\Psi_j(x) = \frac{1}{\lambda_j}\int_\Gamma k(x,y)\psi_j(y)d\mu(y) \tag{3.1}$$

Clearly, Ψ_j and ψ_j agree on Γ, and as a consequence, Ψ_j is an extension of ψ_j. The eigenfunction is extended on Ω as an average of its values on the set Γ, and therefore the extension Ψ_j is termed *geometric harmonic*.

Numerically, the extension procedure can be very ill-conditioned as one is dividing by the singular values of a compact operator. As a consequence, for any $\delta > 0$, we introduce the following notations:

$$S_\delta = \{j \text{ such that } \lambda_j > \delta\lambda_0\}$$
$$L^2_\delta(\Gamma, d\mu) = \text{span}\{\psi_j \text{ such that } j \in S_\delta\}$$
$$\mathcal{H}_\delta = \text{span}\{\Psi_j \text{ such that } j \in S_\delta\}$$

With these notations, the extension operation has condition number $\frac{1}{\delta}$.

In the following sections, we will make use of the algebraic identities relating the extensions and restrictions:

$$\text{Restriction: } \mathbf{K}^*\Psi_j = \psi_j\,,$$
$$\text{Extension: } \mathbf{K}\psi_j = \lambda_j\Psi_j\,.$$

We conclude this section with two remarks. First, an important class of positive kernels is generated by covariance kernels, i.e. kernels of the form

$$k(x,y) = \int_{\xi\in I} e_\xi(x)e_\xi(y)p(\xi)d\xi$$

where $\{e_\xi\}_{\xi\in I}$ is a family of functions defined on Ω, and $p(\xi) \geq 0$. In this setting, each function e_ξ restricted to Γ is interpreted as a vector whose coordinates are indexed by x, and the kernel k represents the covariance information of the cloud of points generated by the mass distribution $p(\xi)d\xi$. Finding the eigenfunctions and eigenvalues associated with this kernel is equivalent to computing the axes and moments of inertia of the could of points, which is also referred to as Principal Component Analysis.

The other remark concerns a variational interpretation of the diagonalization of k. Let $\mathbf{B}$ represent the orthogonal projector on $\mathcal{H}$, defined by $\mathbf{B}f(x) = \langle k(x,\cdot), f\rangle_\Omega$ and let $\mathbf{D}$ be restriction projector defined by $\mathbf{D}f(x) = f(x)$ if $x \in \Gamma$ and $\mathbf{D}f(x) = 0$ otherwise. Then it can be checked that $\mathbf{K} = \mathbf{BD}$ and that if $x \in \Gamma$

$$\mathbf{K}^*\mathbf{K}f(x) = \mathbf{DBD}f(x)\,.$$

This decomposition of $\mathbf{K}^*\mathbf{K}$ as a product of orthogonal projections leads to a variational interpretation of the eigenfunctions that motivated Slepian *et al.* to introduce the prolate functions (see [28]).

3.3 Two properties of the geometric harmonics

The geometric harmonics feature two interesting properties: they are orthogonal both on Γ and Ω, and they have maximum concentration on Γ among all functions of $\mathcal{H}$.

Property 1 (double orthogonality). *The system $\{\Psi_j\}_{j\in S_\delta}$ forms an orthogonal basis of $\mathcal{H}_\delta$ and their restriction $\{\psi_j\}_{j\in S_\delta}$ to Γ forms an orthogonal basis of $L^2_\delta(\Gamma, d\mu)$.*

Proof. By definition, the ψ_j's on Γ are the eigenfunctions of a self-adjoint compact operator, and thus are orthogonal on Γ. In addition, if $\langle\cdot,\cdot\rangle_\Gamma$ denote the inner product on Γ:

$$\begin{aligned}\langle\Psi_i,\Psi_j\rangle_\Omega &= \frac{1}{\lambda_j}\langle\Psi_i,\mathbf{K}\mathbf{K}^*\Psi_j\rangle_\Omega \\ &= \frac{1}{\lambda_j}\langle\mathbf{K}^*\Psi_i,\mathbf{K}^*\Psi_j\rangle_\Omega \\ &= \frac{1}{\lambda_j}\langle\psi_i,\psi_j\rangle_\Gamma\end{aligned}$$

□

For a function $F\in\mathcal{H}$ with restriction f on Γ, we define the concentration of F to be the Rayleigh quotient

$$\frac{\|f\|_\Gamma}{\|F\|_\Omega}\,.$$

The geometric harmonics are also the functions of $\mathcal{H}$ that have maximum concentration on the set Γ:

Property 2 (variational optimality). *The geometric harmonic Ψ_j is a solution of the problem*

$$\max_{F\in\mathcal{H}}\frac{\|f\|_\Gamma}{\|F\|_\Omega}$$

under the constraint that $F\perp\{\Psi_0,...,\Psi_{j-1}\}$ ($f=\mathbf{K}^*F$ represents the restriction of F to Γ). In particular, Ψ_0 is the element of $\mathcal{H}$ that is the most concentrated on Γ.

Proof. By homogeneity of the ratio to be maximized, we see that we can restrict our attention to all F of norm 1. Thus we need to maximize

$$\langle f,f\rangle_\Gamma=\langle\mathbf{K}^*F,\mathbf{K}^*F\rangle_\Gamma=\langle F,\mathbf{K}\mathbf{K}^*F\rangle_\Omega$$

under the constraints

$$\langle F,F\rangle_\Omega=1 \text{ and } \langle F,\Psi_0\rangle_\Omega=...=\langle F,\Psi_{j-1}\rangle_\Omega=0$$

Using the Lagrange multipliers technique, we see that there exist numbers λ and $\alpha_0,...,\alpha_{j-1}$ such that

$$\mathbf{K}\mathbf{K}^*F=\lambda F+\alpha_0\Psi_0+...+\alpha_{j-1}\Psi_{j-1}$$

Taking the inner product with Ψ_i, for $i=0,...,j-1$ and using the constraints and the orthogonality of the geometric harmonic yields

$$\alpha_i=\langle\mathbf{K}\mathbf{K}^*F,\Psi_i\rangle_\Omega=\langle F,\mathbf{K}\mathbf{K}^*\Psi_i\rangle_\Omega=\lambda_i\langle F,\Psi_i\rangle_\Omega=0$$

As a consequence, F is a geometric harmonic associated with the eigenvalue λ, and the functional that we wish to maximize takes the form

$$\langle f,f\rangle_\Gamma=\lambda\,.$$

The maximum is therefore achieved for $F=\Psi_j$. □

3.4 Extension algorithm

Property 1 allows us to describe a simple extension procedure for any empirical function f defined on the set Γ:

- project f onto the space $L^2_\delta(\Gamma, d\mu)$ spanned by the orthogonal system $\{\psi_j\}_{\lambda_j > \delta}$ as

$$f \simeq \mathbf{P}_\delta f = \sum_{j \in S_\delta} \langle f, \psi_j \rangle_\Gamma \psi_j \,,$$

- use the extension Ψ_j of ψ_j to extend the projection $\mathbf{P}_\delta f$ as a function F defined on the set Ω:

$$F(x) = \sum_{j \in S_\delta} \langle f, \psi_j \rangle_\Gamma \Psi_j(x)$$

  with $x \in \Omega$.

Using a terminology from linear algebra, our algorithm merely computes a truncated pseudo-inverse of $\mathbf{K}^*$. This algorithm is easily seen to be consistent, that is, the restriction of F to Γ is again extended as F, and this feature has some practical importance.

Two points need to be discussed here:

- It is clear that this technique does not provide an extension for f but rather for a filtered version of it, namely its orthogonal projection onto $L^2_\delta(\Gamma, d\mu)$. This set is precisely the space of functions that can (safely) be extended to Ω. Indeed, as already mentioned, the condition number of the extension operation on $L^2_\delta(\Gamma, d\mu)$ is $\frac{1}{\delta}$, that is the number of digits lost in the extension of such a function, namely $\log \frac{1}{\delta}$, is under control. Moreover, a general empirical function f on Γ can be extended if the residual $\|f - \mathbf{P}_\delta f\|$ is smaller than a prescribed error.

- Applied to $f \in L^2_\delta(\Gamma, d\mu)$, the algorithm will output a real extension that is an element of $\mathcal{H}$. As we shall see in the examples of the next section, there is in general no unique way to extend f as a function of $\mathcal{H}$, because elements of $\mathcal{H}$ might not be determined by their restrictions to Γ. However, Property 2 allows us to give an interpretation to the extension picked up by the algorithm: among all possible extensions of f as a function of $\mathcal{H}$, the algorithm will output that with maximum concentration, or equivalently, with the minimal energy on Ω. In some sense, this means that the algorithm provides the best extension given the information at our disposal.

Definition 1. A function f defined on Γ is said to be (η, δ)-*extendable* if for a given couple (η, δ) of positive numbers,

$$\sum_{j \in S_\delta} |\langle \psi_j, f \rangle_\Gamma|^2 \geq (1 - \eta) \|f\|^2_\Gamma \,.$$

As a consequence, if $j \in S_\delta$, then ψ_j is (η, δ)-extendable for all $\eta > 0$. This definition means that a function will be extendable if most of its energy is concentrated in the projection over the geometric harmonics whose extensions make numerical sense.

This extension operator is also a valuable tool for the study of the relation between the extrinsic and intrinsic geometries of the set Γ. When Γ is a submanifold of a Euclidean space $\mathbb{R}^n$, we know that an intrinsic Fourier analysis can be performed on functions via

the eigenfunctions of the Laplace-Beltrami operator. Moreover, the diffusion semigroup allows a multiscale decomposition of any function on the set. On the other hand, all these tools already exist in the ambient space $\mathbb{R}^n$, and the relation between extrinsic and intrinsic concepts such as frequency can be investigate with the help of the extension operator. But this is left to Section 3.7, and for now we move on to examples of geometric harmonics.

3.5 Examples of geometric harmonics

In these section, we consider specific instances of positive kernels and the associated extension scheme.

3.5.1 The prolate spheroidal wave functions - Bandlimited extension

In [28] and [19], Slepian *et al.* introduce the prolate spheroidal wave functions as the solution to the problem of finding functions optimally concentrated in time and frequency. The prolates are bandlimited functions of unit energy that have maximum energy within an interval in the time domain. They also generalize their results to higher dimensions (see [29]), and to do so, they define $\mathcal{H}^c$ to be the space of functions of $L^2(\mathbb{R}^n)$ whose Fourier transforms are compactly supported in the ball centered at the origin and of radius $\frac{c}{2}$. In other words, $\mathcal{H}^c$ is the space of bandlimited functions of finite energy, with bandwidth $\frac{c}{2}$. This space is a reproducing kernel Hilbert space with kernel [1] (see appendix A)

$$k_c(x,y) = \left(\frac{c}{2}\right)^{\frac{n}{2}} \frac{J_{\frac{n}{2}}\left(\pi c\|x-y\|\right)}{\|x-y\|^{\frac{n}{2}}}$$

where J_ν is the Bessel function of the first kind and of order ν. We refer to k_c as the Bessel kernel in dimension n. The bandwidth parameter c also plays the role of a scaling parameter.

The prolate spheroidal wave functions are then defined as the eigenfunctions of the operator with kernel k_c. It can be useful to think of these functions as the principal components of the set of functions generated by the complex exponentials with frequency less than $\frac{c}{2}$, all of which being equiprobable. In the prolate setting, the set Γ (in the time domain) is non-singular in $\Omega = \mathbb{R}^n$, and as a consequence, functions of $\mathcal{H}^c$ are determined by their values on Γ. On the contrary, we are mainly interested in the case when Γ is singular in $\mathbb{R}^n$; in this situation there are infinitely many ways to construct bandlimited extensions of functions defined on Γ, as two such extensions differ by a bandlimited function that vanishes on Γ. Our procedure simply finds the bandlimited extension F that maximizes the concentration

$$\frac{\|f\|_\Gamma}{\|F\|_\Omega}$$

of F on Γ, or equivalently, it finds the bandlimited extension F with minimal energy on $\mathbb{R}^n$. From now on, we will refer to $\mathbf{E}_B$ as the extension operator using the Bessel kernel with bandlimit equal to B (for a given preset accuracy δ). Therefore, $\mathbf{E}_B$ computes the bandlimited extension of band B that has the minimal energy on $\mathbb{R}^n$. In Section 3.7, we investigate the relation between the eigenfunctions and eigenvalues of $\mathbf{E}_B$ for different values of B.

[1] Note that in the case when n is odd, the expression of k_c can be further simplified as a sum of derivatives of sinc functions, as shown in appendix A.

In high dimension, all these kernel act similarly as it can be proven that they tend to the Gaussian kernel (appendix B). Finally, in [30], Slepian constructs discrete prolate functions by considering periodic functions generated by the first exponentials $\{e^{2i\pi jx}\}_{|j|\leq q}$ that they restrict to $\Gamma = [-\frac{a}{2}; \frac{a}{2}]$ where $0 < a < 1$. In this case the associated kernel is

$$k_q(x,y) = \sum_{|j|\leq q} e^{2i\pi j(x-y)} = \frac{\sin\left(\pi(2q-1)(x-y)\right)}{\sin \pi(x-y)},$$

which is often referred to us as the Dirichlet kernel. The eigenfunctions of the associated operator form a set of geometric harmonics for periodic functions of the real line.

3.5.2 Harmonic extension

Another example of importance comes from potential theory. Consider the single layer Newtonian potential in $\Omega = \mathbb{R}^n$:

$$k(x,y) = \begin{cases} -\log(\|x-y\|) & \text{if } n = 2, \\ \frac{1}{\|x-y\|^{n-2}} & \text{if } n \geq 3. \end{cases}$$

This kernel is, by definition, the Green's function for the Laplace operator on $\mathbb{R}^n$. Since the Laplace operator is positive, so is the Newtonian potential. For the sake of convenience, assume that $n = 3$ and $\Omega = \mathbb{R}^3$. Then the space $\mathcal{H}$ is the set of potentials

$$F(x) = \int_{\mathbb{R}^3} \frac{d\rho(y)}{\|x-y\|}$$

where ρ is a signed measure on $\mathbb{R}^3$ representing a distribution of charges giving rise to the electrostatic potential F. Observe that in $\mathcal{H}$, the inner product between two potentials F_1 and F_2 generated by two sets of charges ρ_1 and ρ_2 is given by the electrostatic energy of interaction between these charges:

$$\langle F_1, F_2 \rangle_\Omega = \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{d\rho_1 d\rho_2}{\|x-y\|}$$

Now if Γ is a Lipschitz surface of $\mathbb{R}^3$, then a distribution of charges with a single layer density f on Γ will induce a potential

$$F(x) = \int_\Gamma \frac{f(y)d\sigma(y)}{\|x-y\|}$$

that is a harmonic function in $\mathbb{R}^3\backslash\Gamma$. Computing the eigenfunctions of the operator associated to the single layer potential kernel amounts to minimizing the electrostatic self-energy:

$$\int_\Gamma\int_\Gamma \frac{f(x)f(y)}{\|x-y\|} d\sigma(x)d\sigma(y)$$

under the constraint that

$$\int_\Gamma |f(x)|^2 d\sigma(x) = 1$$

These eigenfunction provide a way to extend empirical functions defined on Γ as harmonic functions[2].

[2] A related type of harmonic extension is obtained by considering the double layer potential

$$\overline{k}(x,y) = \frac{\partial}{\partial \nu} k(x,y)$$

where ν is normal to Γ at x

3.5.3 Wavelet extension

Wavelet spaces allow to generate geometric harmonic extensions. Let $\{V_j\}_{j\in\mathbb{Z}}$ be a multiresolution analysis in $\mathbb{R}^n$ and set $\mathcal{H}$ to be the space of scaling functions V_j. By construction, the geometric harmonics corresponding to this space provide a way to extend empirical functions at a given scale. They also define scaling functions adapted to the set Γ. Likewise, one could work on a wavelet space W_j to construct a wavelet extension of a function, and define wavelets adapted to the set Γ.

Let's illustrate this type of construction in the context of the Haar multiresolution. Let Φ be the indicator function of the unit cube in $\mathbb{R}^2$, and let Γ be a finite length curve in the plane. Let $\mathcal{H} = V_j$ for some j, then the reproducing kernel is

$$\sum_{k\in\mathbb{Z}^2} 2^{-j}\Phi(x-2^j k)\Phi(y-2^j k)$$

One restricts this kernel to the curve Γ, to obtain

$$k(x,y) = \sum_{k\in Q_\Gamma} 2^{-j}\Phi(x-2^j k)\Phi(y-2^j k)$$

where Q_Γ is the set of indices $k \in \mathbb{Z}^2$ associated with cubes of unit area intersecting Γ. Now it is clear that the geometric harmonics are the indicator of the cubes Q intersecting Γ, and it can be checked that the eigenvalues are the ratios $2^{-j}|Q\cap\Gamma|$ where $|Q\cap\Gamma|$ is the length of the piece of curve $Q\cap\Gamma$ (see Figure 3.1 for an example).

The same procedure can be followed for more general scaling functions, as well as for each space W_j.

3.6 Bi-Lipschitz parametrization of sets

In the previous chapter, we explained that the eigenfunctions of diffusion operators provide us with a system of coordinates and that in the corresponding embedding space, the Euclidean distance is equal to the diffusion metric on the set Γ. We also noted that it could be useful to obtain a parametrization Ψ of the data that is bi-Lipschitz with a small distortion. Recall that the distortion of a map Ψ between two metric spaces $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y})$ is defined as

$$\mathrm{dist}(\Psi) = \left(\sup_{u\neq v\in\mathcal{X}} \frac{d_\mathcal{Y}(\Psi(u),\Psi(v))}{d_\mathcal{X}(u,v)}\right)\left(\sup_{u\neq v\in\mathcal{X}} \frac{d_\mathcal{X}(u,v)}{d_\mathcal{Y}(\Psi(u),\Psi(v))}\right)$$

In particular, we are interested in finding mappings Ψ that achieve substantial reduction of the dimension while keeping their distortion as close to 1 as possible:

$$1 \leq \mathrm{dist}(\Psi) \leq 1+\delta \tag{3.2}$$

However, in many applications, one merely needs that the above identity hold locally, i.e. for the local distortion:

$$\mathrm{dist}_R(\Psi) = \left(\sup_{d_\mathcal{X}(u,v)<R} \frac{d_\mathcal{Y}(\Psi(u),\Psi(v))}{d_\mathcal{X}(u,v)}\right)\left(\sup_{d_\mathcal{X}(u,v)<R} \frac{d_\mathcal{X}(u,v)}{d_\mathcal{Y}(\Psi(u),\Psi(v))}\right)$$

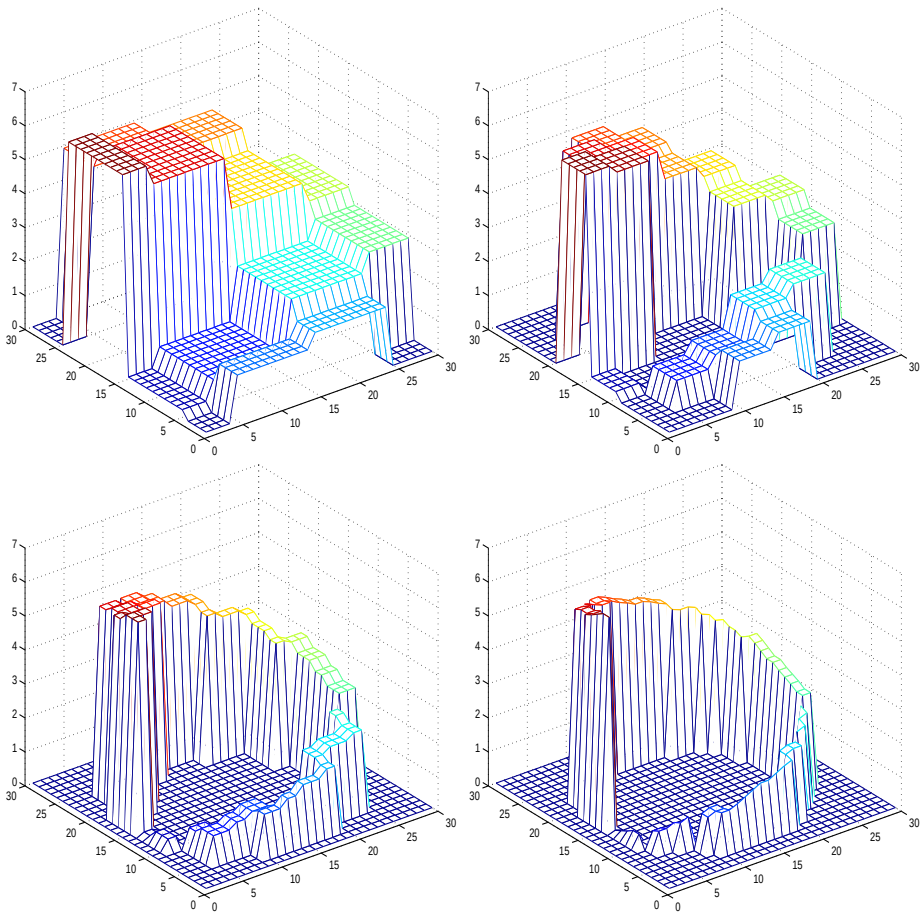

Figure 3.1: Extension of the function $f(\theta) = \theta$ on the circle using Haar scaling functions at different scales.

where $R > 0$. We refer to this problem as the relaxed distortion problem. The benefit of the relaxation is a significant reduction of dimension. Experimentally, we observed that the geometric harmonics provide a simple solution to the relaxed problem, in the sense that it is always possible to find a selection of these coordinates that embed large subsets of the data into a lower dimensional space, such that the local distortion remains close to 1. In this new space, the processing of the data points is usually easier, and a major remaining question concerns the way to extrapolate the inverse of the mapping outside the data set.

3.6.1 Constructing the parametrization

Consider a rotation invariant kernel

$$k_\varepsilon(x,y) = h\left(\frac{\|x-y\|^2}{\varepsilon}\right)$$

where h is normalized so that $h(0) = 1$. In addition we suppose that for $r > 0$,

$$h(r) = 1 - \alpha r + r^2 g(r)$$

where g is bounded on $\mathbb{R}^+$. This is the case for instance with the Gaussian and Bessel kernels. In addition, note that by positivity of k_ε, the quantity $k_\varepsilon(x,y)$ is maximum when $x = y$. Consequently, $\alpha \geq 0$. We will now assume that $\alpha \neq 0$. Let $k_\varepsilon^{(m)}$ be the kernel of the m^{th} power of the operator with kernel k_ε. In what follows, we assume that Γ is a compact submanifold of dimension d.

We suppose that Γ is a C^∞ submanifold of dimension d. Let

$$\Psi_\infty(x) = \begin{pmatrix} \lambda_0^{\frac{1}{2}}\psi_0(x) \\ \lambda_1^{\frac{1}{2}}\psi_1(x) \\ \vdots \end{pmatrix}$$

be the mapping that consists in taking all geometric harmonics as coordinates.

By definition,

$$\|\Psi_\infty(x) - \Psi_\infty(y)\|^2 = 2h(0) - 2h\left(\frac{\|x-y\|^2}{\varepsilon}\right) = 2\alpha\frac{\|x-y\|^2}{\varepsilon} - 2\frac{\|x-y\|^4}{\varepsilon^2}g\left(\frac{\|x-y\|^2}{\varepsilon}\right)$$

Equivalently,

$$\frac{\|\Psi_\infty(x) - \Psi_\infty(y)\|^2}{\|x-y\|^2} = \frac{2\alpha}{\varepsilon}\left(1 - \frac{\|x-y\|^2}{\alpha\varepsilon}g\left(\frac{\|x-y\|^2}{\varepsilon}\right)\right)$$

Thus, clearly, if $\|x-y\|^2 \leq \delta\frac{\alpha}{\|g\|_\infty}\varepsilon$ then we have Lipschitz bounds with ratio $\sqrt{\frac{1+\delta}{1-\delta}}$. This proves that

$$\mathrm{dist}_{R_\varepsilon}(\Psi_\infty) \leq \sqrt{\frac{1+\delta}{1-\delta}}$$

with

$$R_\varepsilon = \sqrt{\delta\frac{\alpha}{\|g\|_\infty}\varepsilon}$$

The size R_ε of the balls for which the bi-Lipschitz identity holds is optimal since if $\|x-y\| > CR_\varepsilon$, then $\|\Psi_\infty(x) - \Psi_\infty(y)\| \simeq 2$. In fact this says that $\|\Psi_\infty(x) - \Psi_\infty(y)\|$ is equivalent to $\frac{\|x-y\|}{\varepsilon+\|x-y\|}$.

But of course Ψ_∞ does not permit to reduce the dimensionality as it employs all coordinate maps. In order to reduce the number of geometric harmonics, it suffices to consider powers $k_\varepsilon^{(m)}$of the operator with kernel k_ε. The eigenvalues are then raised to same power and decay faster. In fact, the decay of the eigenvalues of $k_\varepsilon^{(m)}$ is directly related to the size of this kernel. More precisely, in the asymptotic $\delta \to 0$, the number of eigenvalues λ_j^m that are above the threshold δ is proportional to

$$\left(\frac{1}{m}\log\left(\frac{1}{\delta}\right)\right)^d$$

This estimate, the analog of Weyl's asymptotic law, corresponds to the minimum number of bumps of the form $k_\varepsilon(x,.)$ necessary to approximately cover Γ. If in addition Γ verifies a chord-arc condition, then Euclidean balls can be used to cover the set.

As an illustration, we consider a homemade subset of points in $\mathbb{R}^2$ called "MOUSE". We compute a collection of bi-Lipschitz maps (an atlas) acting on this data set by the following elementary algorithm:

Atlas Computation

1. At each point x of the data set, define a neighborhood N_x by considering the k-nearest neighbors or Euclidean balls.
2. Find a subset of eigenfunctions for which the bi-Lipschitz identity holds on N_x, i.e. such that the local distortion is bounded by some reasonable constant (5 in our example). One way to do so is to follow the algorithm **Eigenfunction selection** described below.
3. Agglomerate all neighborhoods corresponding to the same choice of geometric harmonics.

To find a small set of eigenfunctions that form a bi-Lipschitz parametrization of the neighborhood N_x, we used the following elementary greedy approach:

Eigenfunction selection

1. Given two fixed integers n_c and d_m, and a bound $B > 1$ on the bi-Lipschitz distortion.
2. For all i between 1 and d_m, {
   for all tuples $(j_1, ..., j_i)$ of integers between 1 and n_c, {
   if the bi-Lipschitz distortion of $(\psi_{j_1}, ..., \psi_{j_i})$ is less than B, then return $(j_1, ..., j_i)$.
   }
   }
3. return -1

The integer n_c is the maximal value of the indices of the eigenfunctions one considers in the search, and d_m is the maximal dimensionality of the set. In the case when the function returns -1, it means that the algorithm was unable to find a bi-Lipschitz parametrization with distortion less than B within all tuples of eigenfunctions in the range considered. In this case, one has to allows bigger values of B, n_c or d_m. For the MOUSE set, we chose $B = 5$, $n_c = 20$ and $d_m = 4$.

The result is shown on Figure 3.2. The body and the ear of the mouse are divided into 6 regions, each of which is parameterized by 2 coordinates. The tail of the mouse appears divided into 6 domains, each of them being parameterized by 1 geometric harmonic. Thus the correct dimension is detected. To illustrate the robustness to noise perturbation, we also plotted the output of the algorithm on versions of the MOUSE perturbed by additive random noise.

3.6.2 Inverting the parametrization

The reduction of dimension provided by any embedding $\Psi : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ described above makes it possible to apply some algorithms that are untractable in high dimension. The idea is to reduce the dimension of the data by employing the embedding Ψ, and to apply the treatment of our choice to the data in the embedding space $\mathbb{R}^m$. After this step, it is often necessary to get back to the original space, that is to invert the embedding. In addition, the data at our disposal is generally finite, and we need to invert the embedding at points that are outside the data set. In the following, we assume that the dimension m of the embedding space is fairly small, typically $m \leq 3$.

Let's give a formal statement of the problem. Let $\{x_1, ..., x_p\}$ be our data points in $\mathbb{R}^n$. These points can be thought of as being distributed on a submanifold Γ. We compute the embedding Ψ that maps the data into $\mathbb{R}^m$, where $m < n$. In reality, this means that we form the $p \times p$ matrix whose (i, j)-th entry is given by $a_\varepsilon(x_i, x_j)$, and that we compute its first m eigenfunctions. Thus, we merely have the knowledge of the points $\Psi(x_1), ..., \Psi(x_p)$ in $\mathbb{R}^m$. If the density of points is sufficiently high, then these points are approximately those that we would obtain by computing the actual embedding on the continuous submanifold Γ. As a consequence, we can assume that $\Psi(x_1), ..., \Psi(x_p)$ are approximately the samples of the actual heat embedding of Γ at the points $x_1, ..., x_p$. Equivalently, we have the samples of Ψ^{-1} at the points $\Psi(x_1), ..., \Psi(x_p)$ in the embedding space.

The mapping Ψ^{-1} is a parametrization of Γ defined on $\Psi(\Gamma)$, and to be able to extend it is of crucial importance in several applications. The main virtue of the mapping Ψ is that it provides a dimension reduction and an organization of the points.

3.7 Relation between the intrinsic and extrinsic geometries

Suppose that Γ is a smooth submanifold of $\mathbb{R}^n$. Two different Fourier (or Littlewood-Paley) analyses can be performed on a function f defined on Γ. The first one is purely intrinsic and can be obtained using the Fourier basis on Γ, that is to say the eigenfunctions of the Laplace-Beltrami operator. The other one is the classical Fourier analysis of the ambient space $\mathbb{R}^n$, that one can apply to extensions of f to the whole space.

In this section, we investigate the relation between the two analyzes via the action of the operators of restriction and extension. This approach is equivalent to the question of whether the intrinsic diffusion of heat is equivalent to an extrinsic diffusion.

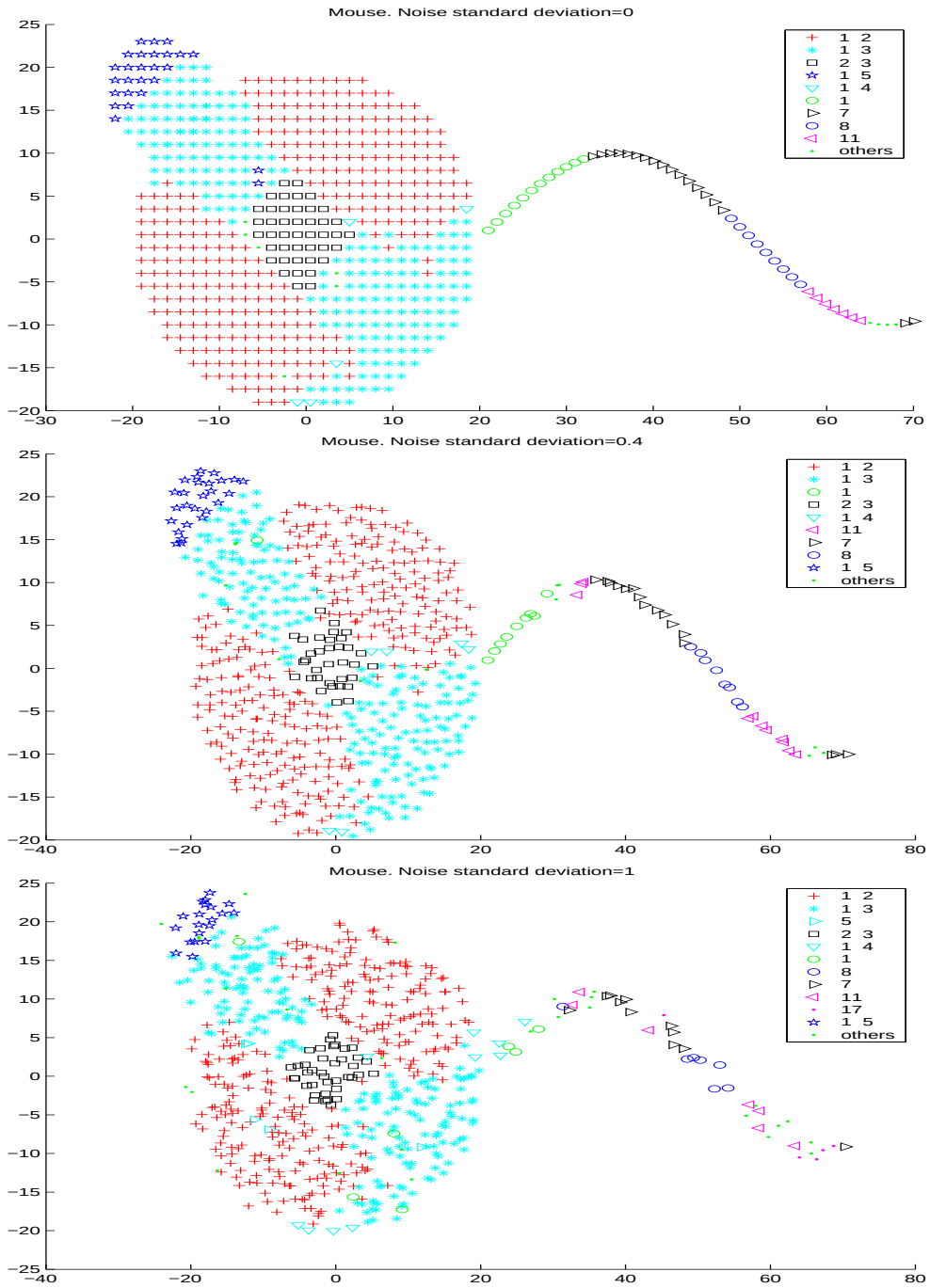

Figure 3.2: Top: the MOUSE set and its decomposition in local patches, each of which having bi-Lipschitz coordinates. The numbers in the legend are the indices of the selected geometric harmonics. To illustrate the robustness to noise, we also present the output of the algorithm on a version of the MOUSE set perturbed by additive Gaussian noise.

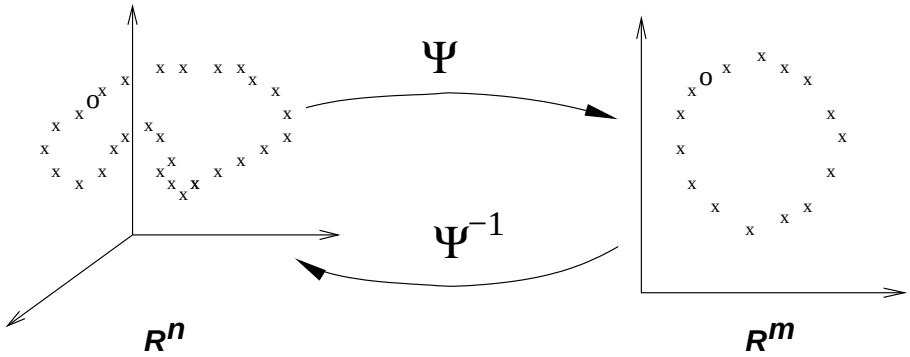

Figure 3.3: The values of Ψ are known at the samples points in $\mathbb{R}^n$. Equivalently, the values of Ψ^{-1} are known at the sample points in $\mathbb{R}^m$. These values are used to interpolate Ψ^{-1} at the new point (circle), and to interpolate the submanifold.

3.7.1 Restriction operator

Since the set Γ is a smooth submanifold, the restriction of a function F to Γ is a well-behaved operation, in the sense that if F is smooth, then so is its restriction f.

We can give a precise meaning of this statement using a space characterization of smoothness: suppose that F is differentiable with a bounded derivative, then f is obviously differentiable as a map from Γ into $\mathbb{R}$, and its (intrinsic) gradient at a point $x \in \Gamma$ is nothing else but the orthogonal projection of gradient of F onto the tangent plane at that same point x.

The same idea can be characterized by a frequency argument. Consider plane waves in $\mathbb{R}^n$:

$$F_\xi(x) = e^{2i\pi\langle\xi,x\rangle}$$

Let Δ be the Laplace-Beltrami operator on Γ that we can assume to be a curve for the sake of simplicity. To relate the extrinsic frequency ξ to the intrinsic frequency ν_j corresponding to the eigenfunction ϕ_j of Δ:

$$\Delta\phi_j = \nu_j^2\phi_j\,,$$

we have the following result:

Proposition 14. *Suppose that the curvature of Γ is bounded by a number $M > 0$. Then if $\|\xi\| > \frac{M}{\pi}$,*

$$|\langle\phi_j, f_\xi\rangle| \leq \sqrt{\mu(\Gamma)}\left(\frac{4\pi^2\|\xi\|^2}{\nu_j^2}\right)^m$$

for all $m \geq 0$. As a conclusion, a function with only low extrinsic frequencies is also a function with low intrinsic frequencies when restricted to Γ.

Proof. Locally, on the curve around $x \in \Gamma$, the function F_ξ has the form

$$f_\xi(x) = F_\xi(x) = \exp(2i\pi(\langle\xi, x\rangle + u\xi_T + a(x)u^2\xi_N)$$

where u is the local coordinate in the tangent plane of the point y, $a(x)$ is the scalar curvature at x, and ξ_T and ξ_N are the tangent and normal projections of ξ in the osculatory plane at x. A Taylor expansion yields:

$$f_\xi(x) = e^{2i\pi\langle\xi,x\rangle}\left(1 + 2i\pi\xi_T u + 2i\pi(a(x)\xi_N + i\pi\xi_T^2)u^2 + ...\right)\,.$$

Now by using Lemma 6 of Section 2.3.2, we identify

$$\Delta f_\xi(x) = 4i\pi(a(x)\xi_N + i\pi\xi_T^2) f_\xi(x)\,.$$

Therefore, if the curvature is bounded by $M > 0$, we have the trivial estimate for $\|\xi\| > \frac{M}{\pi}$,

$$\|\Delta f_\xi\|_\Gamma \leq 4\pi^2 \|\xi\|^2 \|f_\xi\|_\Gamma\,.$$

In fact it is easily seen that for all $m \geq 0$,

$$\|\Delta^m f_\xi\|_\Gamma \leq \left(4\pi^2\|\xi\|^2\right)^m \|f_\xi\|_\Gamma\,.$$

Since

$$\Delta^m f_\xi(x) = \sum_{j\geq 0} \nu_j^{2m} \langle \phi_j, f\rangle \phi_j(x)$$

and

$$\|f_\xi\|_\Gamma^2 \leq \mu(\Gamma)\,,$$

we must have, by Parseval,

$$\sum_{j\geq 0} \nu_j^{4m} |\langle \phi_j, f_\xi\rangle|^2 \leq \mu(\Gamma) \left(4\pi^2\|\xi\|^2\right)^{2m}\,.$$

In particular,

$$|\langle \phi_j, f_\xi\rangle| \leq \sqrt{\mu(\Gamma)} \left(\frac{4\pi^2\|\xi\|^2}{\nu_j^2}\right)^m$$

for all $m \geq 0$. Therefore the coefficients of expansions in the eigenfunctions of the Laplace-Beltrami operator are negligible, except for finitely many, namely those for which the eigenvalue ν_j^2 is less than $4\pi^2\|\xi\|^2$. □

3.7.2 Extension operator

The extension algorithm described in Section 3.4 is a two-step procedure

- the function f on the data is first pre-filtered by a projection on the geometric harmonics that numerically admit an extension,
- then the extension is computed.

Of course, most functions f on the data set Γ cannot be extended, but checking whether this is the case of not is relatively simple. Indeed, one just needs to verify that not too much energy of f is lost by the projection operation, in which case one can conclude that, up to a small residual, f belongs to the space $L^2_\delta(\Gamma, d\mu)$ spanned by the geometric harmonics that can be extended with a prescribed condition number $\frac{1}{\delta}$. In other words, to use the terminology introduced earlier, the relevant concept is that of (η, δ)-extendable functions.

For the study of the restriction operator, we looked at the restriction of the "Fourier basis" in $\mathbb{R}^n$. For the extension problem, it is natural to extend the Fourier basis on Γ, i.e. the set of eigenfunctions of the Laplace-Beltrami operator on Γ. In order to relate the intrinsic frequencies that these functions represent to the extrinsic spectrum (provided by the Fourier analysis in $\mathbb{R}^n$), it can be instructive to compute how many of these functions admit a bandlimited extension, with a given bandwidth, or a Gaussian extension, with a

given scale. More precisely, one can compute the maximal value of the index j such that it is possible to extend ϕ_j using a Gaussian of fixed variance ε. Similarly, for an eigenfunction ϕ_j of the Laplace-Beltrami operator on Γ, one can compute the largest scale that allows to extend this function, that is to say how far ϕ_j can be extended away from the set. Last, since this procedure is global on the set Γ, it is desirable to multiply ϕ_j by a window to localize the analysis. Therefore we consider the following intrinsic wave packets:

$$w_x^{j,t}(y) = w_{x,t}(y)\phi_j(y)$$

where $w_{x,t}$ is a window function centered at x and of width $\sqrt{t}$. For instance,

$$w_{x,t}(y) = \exp\left(-\widetilde{D}_t^2(x,y)\right)$$

is a Gaussian window in the diffusion space. The distance $\widetilde{D}_t(x,y) = C_t D_t(x,y)$ is a multiple of the diffusion distance, normalized so that its maximum value is 1 (therefore C_t is an increasing exponential).

In some sense, $w_x^{j,t}$ defines an intrinsic local cosine waveform: x represents the time location parameter, j is the intrinsic frequency location, and $\sqrt{t}$ is the time-width.

For instance consider the set shown on Figure 3.4. This curve has low frequencies on the left part, but has wild variations on the right part. We have plotted the domains of the plane where the 10^{th} eigenfunction of the Laplace-Beltrami operator can be locally extended. In other words, we have used the local cosine wave packet describe above with $j = 10$. These domains reflect the relation between the intrinsic and extrinsic frequencies.

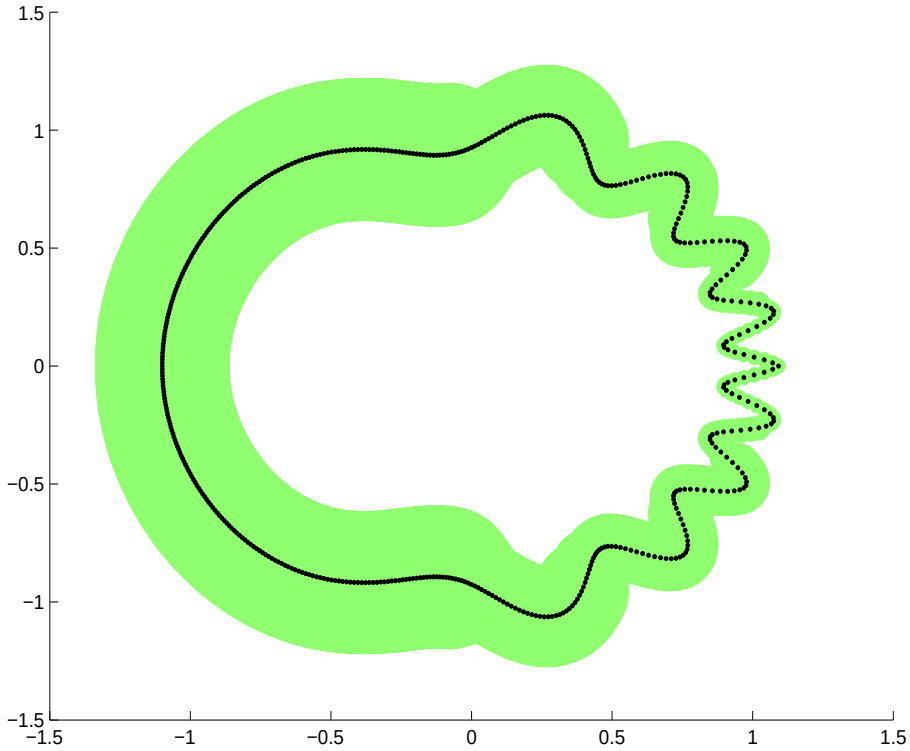

Figure 3.4: Original set and domains of extension for ϕ_{10}.

This simple example shows that extending a function off the set Γ can result into an ill-conditioned operation if the set is complicated. In the following proposition, we show that a function f with intrinsic bandlimit B on Γ admits an approximate extension that is a bandlimited function with band CB where C is some universal constant that depends on the geometry of Γ. In order to adopt a broader point of view, instead of considering the eigenfunctions of the Laplace-Beltrami operator Δ, we will study the extension of eigenfunctions of the following elliptic operator:

$$\mathbf{\Delta} = \Delta + E\,,$$

where E is a bounded potential function. We have seen in the previous chapter, Sections 2.3.2, 2.3.3 and 2.3.4, that this type of differential operator arises naturally as the limit of several families of operators. We keep the same notations for the eigenfunctions and eigenvalues, namely,

$$\mathbf{\Delta}\phi_j = \nu_j^2 \phi_j\,.$$

Proposition 15. *Let $\delta > 0$ be a preset accuracy. There exists a constant $C > 0$ such that for all $j \geq 0$, one can construct a function F_j defined on $\mathbb{R}^n$ satisfying:*

- *F_j is an extension of ϕ_j, i.e.,*

$$F_j(x) = \phi_j(x) \text{ for all } x \in \Gamma\,,$$

- *F_j can be approximated to relative precision δ by a bandlimited function B_j of band $C\nu_j$:*

$$\frac{\left(\int_{\mathbb{R}^n} |F_j(x) - B_j(x)|^2\, dx\right)^{\frac{1}{2}}}{\left(\int_{\mathbb{R}^n} |F_j(x)|^2\, dx\right)^{\frac{1}{2}}} \leq \delta\,.$$

The constant C depends on the precision δ and on the geometry of Γ. More precisely, if the diffusion metric on Γ is comparable to the Euclidean metric of $\mathbb{R}^n$, then C can be controlled. For instance, in the example of figure 3.4, C can be quite large because of the oscillations of the curve.

Proof. For any $x \in \mathbb{R}^n$, let $x' \in \Gamma$ be such that

$$\|x - x'\| = \inf_{y \in \Gamma} \|x - y\|\,.$$

Define F_j by

$$F_j(x) = e^{-\nu_j^2 \|x - x'\|^2} \phi_j(x')\,.$$

The function F_j is an extension of ϕ_j to $\mathbb{R}^n$. To estimate the decay of its spectrum, we need to bound its gradient $\nabla^m F_j(x)$ (tensor of order m). The Leibnitz formula yields:

$$\|\nabla^m F_j(x)\| \leq \sum_{i=0}^{m} \binom{m}{i} \|\nabla^i (e^{-\nu_j^2 \|x - x'\|^2})\| . \|\nabla^{m-i}(\phi_j(x'))\|\,.$$

The triangular inequality in $L^2(\mathbb{R}^n)$ gives

$$\left(\int_{\mathbb{R}^n} \|\nabla^m F_j(x)\|^2 dx\right)^{\frac{1}{2}} \leq \sum_{i=0}^{m} \binom{m}{i} \left(\int_{\mathbb{R}^n} \|\nabla^i (e^{-\nu_j^2 \|x - x'\|^2})\|^2 . \|\nabla^{m-i}(\phi_j(x'))\|^2 dx\right)^{\frac{1}{2}}\,.$$

To evaluate each term of the right-hand side, we make use of the following lemma:

Lemma 16. *Let f_ν be a function on $\mathbb{R}^n$ of the form*

$$f_\nu(x) = g(\nu\|x - x'\|)h_\nu(x')\,,$$

where g has an exponential decay. Again, let $M > 0$ be a bound on the curvature of Γ. Then, if $\nu > 4M$,

$$\int_{\mathbb{R}^n} |f(x)|dx \asymp \nu^{-(n-d)} \int_0^{+\infty} |g(r)|r^{n-d}dr \int_\Gamma |h_\nu(u)|du\,.$$

Because of the decay of g, up to exponentially small terms, this integral can be computed on the set Ω_ν of all points at distance less than or equal to $\frac{1}{\nu}$. We can associate to any $x \in \mathbb{R}^n$ a pair (u,t) where $x' = u$ is the closest point to $x \in \Omega_\nu$ and $t = x - u$. Conversely, to any $u \in \Gamma$ and t normal to Γ at u, we can associate the point $x = u + t$. Let $J(u,t)$ denote the Jacobian of the change of variable $(u,t) \mapsto x$. The lemma follows from the fact that J is bounded from below and above for all $x \in \Omega_\nu$. Indeed, first, a variation dt of t entails the same variation of x. Second, a variation du in the tangent plane at u entails a variation of x of order $(1+2\alpha(u)|\|t\|)du$ in the tangent direction, where $|\alpha(u)| \leq M$, and of order $\|t\|du^2$ in the normal direction. To conclude, since $\|t\| < \frac{1}{\nu} < \frac{1}{4M}$, we obtain that $1-2\alpha(u)\|t\| > 1-\frac{1}{2}$ and $1 + 2\alpha(u)\|t\| < 1 + \frac{1}{2}$. Finally,

$$\frac{1}{2} < J(u,t) < \frac{3}{2}\,.$$

Therefore, with the same constants,

$$\int_{\mathbb{R}^n} |f_\nu(x)|dx \asymp \int_\Gamma |h_\nu(u)|du \int_{\mathbb{R}^{n-d}} g(\nu\|t\|)dt\,,$$

which ends the proof of the lemma.

Going back to the proof of the proposition, the lemma implies that

$$\left(\int_{\mathbb{R}^n} \|\nabla^m F_j(x)\|^2 dx\right)^{\frac{1}{2}} \leq \sum_{i=0}^{m} \binom{m}{i} K_i \nu_j^{i-\frac{n-d}{2}} \left(\int_\Gamma \|\nabla^{m-i}\phi_j(u)\|du\right)^{\frac{1}{2}}\,, \tag{3.3}$$

where K_i is a constant of the order of magnitude of the L^2 norm of the i^{th} derivative of the univariate Gaussian at scale 1. What remains to be done is to bound the $L^2(\Gamma)$-norm of the derivatives of ϕ_j. To do so, we need the following result:

Lemma 17. *For all $s \geq 0$, there exists C'_s such that*

$$\|\phi_j\|_s = \left(\sum_{|\alpha|\leq s} \int_\Gamma |\partial^\alpha \phi_j(u)|^2 du\right)^{\frac{1}{2}} \leq C'_s \nu_j^i\,.$$

This lemma follows from the classical theory of elliptic operators, which says that since we can bound the norm of $\boldsymbol{\Delta}^k \phi_j$, we have a bound on all derivatives of order less than or equal to $2k$.

Let

$$C_E = \sup_{u\in\Gamma} |E(u)|\,.$$

For $i = 0$, the lemma is trivial, and for $i = 1$, it results from an integration by parts as

$$\begin{aligned} \|\nabla\phi_j\|_\Gamma^2 &= \langle\Delta\phi_j, \phi_j\rangle_\Gamma \quad \text{by the Stokes formula,} \\ &= \langle\boldsymbol{\Delta}\phi_j, \phi_j\rangle_\Gamma - \langle E\phi_j, \phi_j\rangle_\Gamma\,, \\ &\leq \nu_j^2 + C_E\,, \end{aligned}$$

and therefore

$$\|\phi_j\|_1^2 = \|\phi_j\|_\Gamma^2 + \|\nabla\phi_j\|_\Gamma^2 = 1 + \nu_j^2 + C_E \leq C_1'^2\nu_j^2\,.$$

In [12], p 262, it is shown that if L is an elliptic operator of order k, then for all $i \geq k$ and all f defined on Γ, we have:

$$\|f\|_i \leq C'(\|Lf\|_{i-k} + \|f\|_{i-1})\,. \tag{3.4}$$

We can now proceed by induction:

- for $i = 2s$ and $L = \boldsymbol{\Delta}^s$, identity (3.4) yields

$$\begin{aligned} \|\phi_j\|_{2s} &\leq C'(\|\boldsymbol{\Delta}^s\phi_j\|_\Gamma + \|\phi_j\|_{2s-1}) \\ &\leq C'(\nu_j^{2s} + C'_{2s-1}\nu_j^{2s-1}) \\ &\leq C'_{2s}\nu_j^{2s}\,, \end{aligned}$$

- for $i = 2s+1$ and $L = \boldsymbol{\Delta}^s$, identity (3.4) becomes

$$\|\phi_j\|_{2s+1} \leq C'(\|\boldsymbol{\Delta}^s\phi_j\|_1 + \|\phi_j\|_{2s})\,.$$

We have

$$\begin{aligned} \|\boldsymbol{\Delta}^s\phi_j\|_1^2 &= \|\boldsymbol{\Delta}^s\phi_j\|_\Gamma^2 + \|\nabla\boldsymbol{\Delta}^s\phi_j\|_\Gamma^2 \quad \text{by definition,} \\ &= \nu_j^{4s} + \langle\Delta\boldsymbol{\Delta}^s\phi_j, \boldsymbol{\Delta}^s\phi_j\rangle_\Gamma \quad \text{by the Stokes formula,} \\ &= \nu_j^{4s} + \langle\boldsymbol{\Delta}^{s+1}\phi_j, \boldsymbol{\Delta}^s\phi_j\rangle_\Gamma - \langle E\boldsymbol{\Delta}^s\phi_j, \boldsymbol{\Delta}^s\phi_j\rangle_\Gamma\,, \\ &\leq \nu_j^{4s} + \nu_j^{2(2s+1)} + C_E\nu_j^{4s}\,, \end{aligned}$$

and finally,

$$\|\phi_j\|_{2s+1} \leq C'\left(\sqrt{(1+C_E)\nu_j^{4s} + \nu_j^{2(2s+1)}} + C'_{2s}\nu_j^{2s}\right) \leq C'_{2s+1}\nu_j^{2s+1}\,.$$

The lemma is now proven, and it allows us to finish the proof of the proposition as from equation (3.3), we can conclude that

$$\left(\int_{\mathbb{R}^n} \|\nabla^m F_j(x)\|^2 dx\right)^{\frac{1}{2}} \leq C_m\nu_j^{m-\frac{n-d}{2}}\,.$$

Now for a fixed value of m, define

$$\widehat{B}_j(\xi) = \begin{cases} \widehat{F}_j(\xi) & \text{if } \|\xi\| < C\nu_j, \\ 0 & \text{otherwise,} \end{cases}$$

then, by the Parseval identity, we have

$$\begin{aligned}\int_{\mathbb{R}^n} |F_j(x) - B_j(x)|^2\, dx &= \int_{\|\xi\|>C\nu_j} |\widehat{F}_j(\xi)|^2 d\xi\,,\\ &\leq \int_{\|\xi\|>C\nu_j} |\widehat{F}_j(\xi)|^2 \frac{\|\xi\|^{2m}}{(C\nu_j)^{2m}} d\xi\,,\\ &\leq \frac{1}{(C\nu_j)^{2m}} \int_{\mathbb{R}^n} \|\nabla^m F_j(x)\|^2 dx\,,\\ &\leq \frac{C_m^2}{C^{2m}} \nu_j^{-(n-d)}\,.\end{aligned}$$

Form lemma 16, we have

$$\int_{\mathbb{R}^n} |F_j(x)|^2 dx \geq K^2 \nu_j^{-(n-d)}$$

for some $K > 0$, and we merely have to pick C so that

$$\frac{C_m^{\frac{1}{m}}}{KC} < \delta\,.$$

□

Recall that $\mathbf{E}_B$ is the bandlimited extension operator corresponding to band B. From now on, we set $B = C\nu_j$. The consequence of this proposition is that, because of its optimal property, the extension provided by $\mathbf{E}_B$ must have an energy on $\mathbb{R}^n$ that is less than or equal to that of the extension that we constructed in the proof above. This means that the numerical support of $\mathbf{E}_B\phi_j$ will be included in a tube of radius $\frac{1}{\nu_j}$ around Γ. Theoretically, it could be much thinner, but because of the Heisenberg principle, then the support cannot really be smaller:

Lemma 18. *The standard deviation of the extension $\mathbf{E}_B\phi_j$ along any normal direction to Γ is at least equal to $\frac{C'}{\nu_j}$ for some $C' > 0$ independent of ν_j.*

Proof. Let f be the restriction of $\mathbf{E}_B\phi_j$ on a line that is normal to Γ. Then f is a univariate bandlimited function of band $C\nu_j$. Let

$$\mathrm{Var}(f) = \frac{\int_{\mathbb{R}} (x-\overline{x})^2 |f(x)|^2 dx}{\int_{\mathbb{R}} |f(x)|^2 dx}$$

and

$$\mathrm{Var}(\widehat{f}) = \frac{\int_{\mathbb{R}} (\xi-\overline{\xi})^2 |\widehat{f}(\xi)|^2 d\xi}{\int_{\mathbb{R}} |\widehat{f}(\xi)|^2 d\xi}$$

be the variances of f in the space and frequency domains, $\overline{x}$ and $\overline{\xi}$ being the corresponding means. Then since f is bandlimited,

$$\int_{\mathbb{R}} \xi^2 |\widehat{f}(\xi)|^2 d\xi \leq (C\nu_j)^2 \int_{\mathbb{R}} |\widehat{f}(\xi)|^2 d\xi\,,$$

and consequently, $\mathrm{Var}(\widehat{f}) \leq (C\nu_j)^2$ (the variance is always smaller than the second moment). The Heisenberg uncertainty principle implies that $\mathrm{Var}(f) \geq C'^2 \nu_j^{-2}$. □

As a conclusion, the extension operation satisfies a certain version of the Heisenberg principle relating the spectrum of the operator $\boldsymbol{\Delta}$ to the space and frequency localizations of the extensions of its eigenfunctions ϕ_j. This principle says that if $\boldsymbol{\Delta}\phi_j = \nu_j^2\phi_j$, then the operator $\mathbf{E}_B$ extends ϕ_j to a bandlimited function of band $\mathcal{O}(\nu_j)$ and localized in a tube of radius $\mathcal{O}(\frac{1}{\nu_j})$ around Γ.

It is worthy to mention that similar results can be obtained for, say, Gaussian kernels. Let's finish by an illustration, in the Gaussian case, as the calculations are easy. Let Γ be the unit circle in the plane, and let's look at the Gaussian extensions. In this case, the Fourier basis $\{\psi_j(x) = \frac{1}{\sqrt{2\pi}}e^{ij\theta}\}$ constitutes the set of eigenfunctions of any rotation invariant operator (such as $\Delta + E$). For $x \in \mathbb{R}^2$ with polar coordinates (r, α), and $y \in \Gamma$ of polar coordinates $(1, \beta)$, we have $\|x - y\|^2 = r^2 + 1 - 2r\cos(\alpha - \beta)$ and thus

$$\begin{aligned}
\mathbf{K}_B\psi_j(x) &= \frac{1}{\sqrt{2\pi}}\int_0^{2\pi} e^{-B^2(r^2+1-2r\cos(\alpha-\beta))}e^{ij\beta}d\beta\,, \\
&= \frac{1}{\sqrt{2\pi}}e^{ij\alpha}e^{-B^2(r^2+1)}\int_0^{2\pi} e^{2B^2r\cos\beta}\cos(j\beta)d\beta\,, \\
&= \psi_j(\frac{x}{\|x\|})e^{-B^2(r^2+1)}2\pi i^j J_j(2iB^2r^2)\,,
\end{aligned}$$

where the last equality comes from 9.1.21 in [1]. Therefore taking $r = 1$ allows to identify $\lambda_j = 2\pi i^j e^{-2B^2}J_j(2iB^2)$, which means that the eigenvalues decay roughly like $B^j j^{-j}$ by 9.3.1 in [1]. We deduce that the extension of ψ_j has the form

$$\Psi_j(x) = \mathbf{E}_B\psi_j(x) = \frac{e^{ij\alpha}}{\sqrt{2\pi}}\frac{J_j(2iB^2r^2)}{J_j(2iB^2)}e^{-B^2(r^2-1)}\,,$$

which implies that each extension Ψ_j decays approximately like a Gaussian at scale $\frac{1}{B}$.

3.7.3 Multiscale extension

For a given $B' > 0$, let $\mathbf{K}_{B'}$ be the integral operator with Bessel kernel of band B' acting on functions defined on Γ. In other words, to obtain the geometric harmonics of band B', we need to diagonalize $\mathbf{K}_{B'}$ on the set Γ. As we have explained, from these geometric harmonics, we can define an extension operator $\mathbf{E}_{B'}$ that will extend functions from the set Γ as bandlimited functions of band B' on $\mathbb{R}^n$

In the previous chapter, Section 2.3.2, Proposition 7 shows that

$$\mathbf{K}_{B'}\phi_j(x) = C_1 B'^{-d}\left[\left(1 - C_2^2\frac{\nu_j^2}{B'^2}\right)\phi_j(x) + \mathcal{O}\left(\frac{1}{B'^{\frac{3}{2}}}\right)\right] \quad \text{for } x \in \Gamma\,, \tag{3.5}$$

where C_1 and C_2 are constants that can be computed from the moments of the Bessel kernel. This identity establishes the relation between the eigenvalues and eigenfunctions of the different operators $\mathbf{K}_{B'}$ for $B' > 0$. In particular, it asserts that if $B' \gg \nu_j$, then the eigenfunction ϕ_j of $\boldsymbol{\Delta}$ is an approximate eigenfunction of $\mathbf{K}_{B'}$ with eigenvalue $B'^{-d}(1 - C_2^2\frac{\nu_j^2}{B'^2})$:

$$\lambda_j(B') \approx \lambda_0(B')\left(1 - C_2^2\frac{\nu_j^2}{B'^2}\right)\,.$$

Consequently, as soon as

$$B' > \frac{C_2 \nu_j}{\sqrt{1-\delta}},$$

ϕ_j belongs to the set $L^2_\delta(\Gamma, d\mu)$ of functions that can be extended with condition number δ. More generally, we know that $K^{tB'^2}_{B'}\phi_j(x) \approx C_3 e^{-C_2^2 t \nu_j^2}$ and therefore $C_2^2\nu_j^2 \approx B'^2 \log(\frac{\lambda_0(B')}{\lambda_i(B')})$. Equation (3.5) also says that the first few eigenfunctions of $\mathbf{K}_{B'}$ converge rapidly to their limit values as $B' \to +\infty$, and as a consequence, the first eigenfunctions of $\mathbf{K}_{B_1}$ and $\mathbf{K}_{B_2}$ are approximately the same if B_1 and B_2 are sufficiently large.

On another hand, from the result of the previous section, we also know that by setting $B = C\nu_j$, for some constant C, the operator $\mathbf{E}_B$ extends ϕ_j to a distance of the order of $\frac{1}{\nu_j}$ to the set Γ. All these simple observations give rise to a natural *multiscale extension scheme*:

Multiscale extension scheme

1. Fix a condition number δ and the finest scale 2^{-I}. Let $C > 0$ be larger than the constant in Proposition 15 and satisfying

   $$\left(1 - \frac{1}{C^2}\right) > \delta\,.$$

   Define $B_I = C2^I$ and compute the geometric harmonics at the finest scale

   $$\mathbf{K}_{B_I}\psi_j(x) = \lambda_j(B_I)\psi_j(x)\,.$$

   Also, define

   $$\overline{\nu}_j = \frac{B_I}{C_2}\sqrt{\log\left(\frac{\lambda_0(B_I)}{\lambda_j(B_I)}\right)}\,.$$

2. Group the $\overline{\nu}_j$'s in dyadic packets: for $i \leq I$, define

   $$S_i = \{j \text{ such that } 2^{i-1} \leq \overline{\nu}_j < 2^i\}\,,$$

   and define $B_i = C2^i$. If $j \in S_i$, then ψ_j can be extended at scale 2^{-i} as a bandlimited function of band B_i by:

   $$\Psi_j = \mathbf{E}_{B_i}\psi_j\,.$$

3. For any function f defined on Γ, compute its expansion in $\{\psi_j\}$:

   $$f \approx \sum_{\lambda_J \geq \delta\lambda_0} c_j\psi_j\,.$$

4. Extend f as

   $$F = \sum_{i\leq I}\sum_{j\in S_i} c_j\Psi_j = \sum_{i\leq I}\sum_{j\in S_i} c_j\mathbf{E}_{B_i}\psi_j\,.$$

Let's justify the steps of the algorithm. First, if the smallest scale is sufficiently fine, *i.e*,

if I is large enough, then the condition

$$\left(1 - \frac{1}{C^2}\right) > \delta$$

ensures that $\lambda_j(B_i) > \delta\lambda_0(B_i)$ and all extensions are well-defined. Second, the expression of $\overline{\nu}_j$ entails that $\overline{\nu}_j \approx \nu_j$ for small values of j (first eigenfunctions). Last, the scale B_i is picked up so that we extend ψ_j, $j \in S_i$, to an optimal distance ($\approx 2^{-i}$) given the frequency band 2^i of this function (here, optimality refers to the Heisenberg principle explained in the previous section).

Note that this algorithm can be adapted to the eigenfunctions of the Laplace-Beltrami operator: f is developed onto the eigenfunctions of Δ, each of which being extended to the corresponding distance. Also, there is no reason to work on the global level, and it is possible to localize functions using windows.

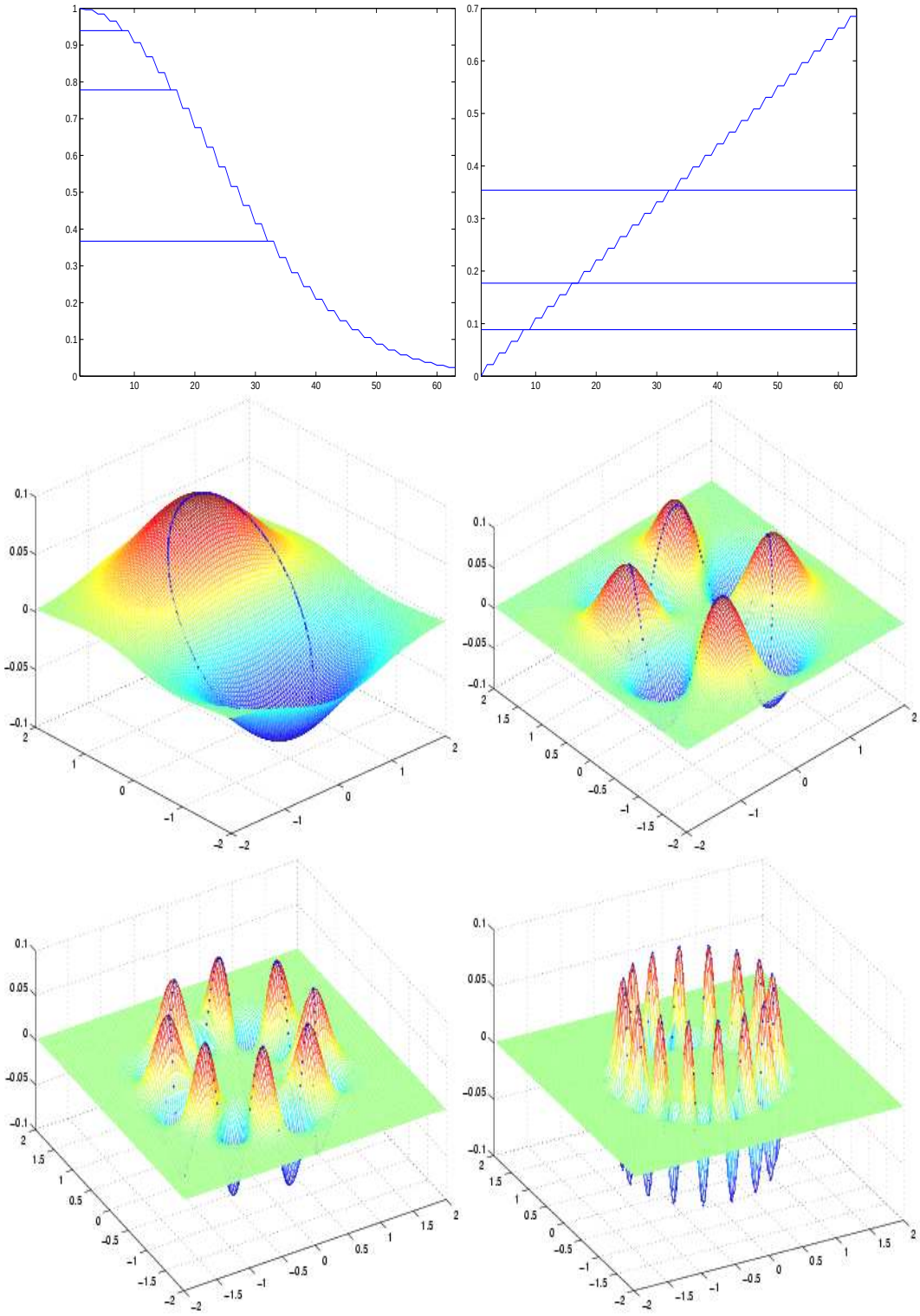

Figure 3.5: Top: the distribution of $\frac{\lambda_j}{\lambda_0}$ (left) and of ν_j (right). The groups of frequencies are indicated by the horizontal lines. Middle and bottom, clockwise: extensions of $\phi_1(s) = \cos(2\pi s)$, $\phi_7(s) = \cos(8\pi s)$, $\phi_{15}(s) = \cos(16\pi s)$ and $\phi_{31}(s) = \cos(32\pi s)$ from the unit circle to the plane.

Conclusion and future work

Diffusion processes provide a unified framework for addressing the problem of finding relevant geometric descriptions of data sets. It is possible to design a specific diffusion matrix that will locally preserve some specific metric or exhibit a particular behavior, and the diffusion maps that it defines will produce a global representation of the data that aggregates the local information. However, a certain number of questions still need to be answered. For instance it would be interesting to know how to extend the different normalizations presented in this thesis to fractal sets. The point here is that on those sets, it might be difficult to directly define a Laplacian, whereas constructing a relevant diffusion kernel might be possible as this object is very regular. How would the points be embedded? Another range of applications concerns differential equations and dynamical systems. As we have shown, eigenfunctions of diffusion matrices allow to integrate, compare and organize trajectories of a differential system. This is really the first step towards a generalization of differential calculus in terms of diffusion processes.

By their capacity to perform out-of-sample extension of functions defined on a data set, the geometric harmonics are a valuable tool for statistical learning. The fact they also provide a low distortion embedding of the data underlines the potential interest in visualization applications. The properties of the associated restriction and extension operators provide a simple way to investigate the relation between the intrinsic and extrinsic geometries. In particular, the geometric harmonics allow to define a multiscale extension scheme, and their ability to transpose signal processing concepts to manifolds opens the door to the construction of a sampling theory for sets.

Last, practical concerns as well as the need for a better understanding of the geometry of data sets in high dimension motivate the development of fast and efficient methods for the computation of the eigendecomposition of all these kernels.

Appendix A

Expression of the Bessel kernels

In what follows, we derive the form of the kernel corresponding to functions whose Fourier transform is the indicator function of the ball of radius $\frac{c}{2}$ centered at the origin, namely:

$$k_c(x,y) = \int_{\|\xi\|<\frac{c}{2}} e^{2i\pi\langle\xi,x-y\rangle} d\xi = \left(\frac{c}{2}\right)^{\frac{n}{2}} \frac{J_{\frac{n}{2}}(\pi c\|x-y\|)}{\|x-y\|^{\frac{n}{2}}},$$

where J_ν is the Bessel function of the first kind of order ν. This kernel will be termed "Bessel kernel".

Since the kernel is really a function of $\|x-y\|$, we are looking for the form of the Fourier transform of the indicator of the unit ball in dimension n. To do so, we make use of a result known under the name of the Bochner-Coifman-Howe periodicity relations:

Lemma 19. *Let f be a radial function, and let $\mathcal{F}_n f(\xi) = h_n(\|\xi\|^2)$ be its Fourier transform in dimension n.*

Then the Fourier transforms of f in dimension n and $n+2$ are related in the following manner:

$$h_{n+2}(u) = -\frac{1}{\pi} h_n'(u)\,.$$

In other words, to compute the Fourier transform $h_{n+2}(\|\xi\|^2)$ of f in $\mathbb{R}^{n+2}$, one can start from the Fourier transform $h_n(\|\xi\|^2)$ in dimension n, view this function as a function of $\|\xi\|^2$ and compute its derivative in this variable.

Proof. Since any radial function or tempered distribution can be approximated as a sum of Gaussians, one merely needs to verify the relation for $f(x) = e^{-\alpha r^2}$. In this case,

$$\mathcal{F}_n f(\xi) = \left(\frac{\pi}{\alpha}\right)^{\frac{n}{2}} e^{-\frac{\pi^2\xi^2}{\alpha}}\,.$$

Thus

$$h_n(u) = \left(\frac{\pi}{\alpha}\right)^{\frac{n}{2}} e^{-\frac{\pi^2 u}{\alpha}},$$

and the identity is satisfied. □

Using this lemma we can now conclude:

Proposition 20. *In dimension n, the Bessel kernel has the following form:*

$$k_c(x,y) = \left(\frac{c}{2}\right)^{\frac{n}{2}} \frac{J_{\frac{n}{2}}(\pi c\|x-y\|)}{\|x-y\|^{\frac{n}{2}}}\,.$$

Moreover, if n is odd, then the simpler formula can be used:

$$k_c(x,y) = M_c \left(\frac{1}{r}\frac{d}{dr}\right)^{\frac{n-1}{2}} sinc(cr)\,,$$

where $r = \|x-y\|$ and

$$M_c = \left(\frac{c}{2}\right)^{\frac{n}{2}} \sqrt{2c}(-1)^{\frac{n-1}{2}}\,.$$

Proof. By a trivial scaling argument, we may assume that $c = 2$.

Then if $n = 1$, then

$$k(x,y) = \int_{-1}^{1} e^{2i\xi\pi(x-y)}d\xi = 2\mathrm{sinc}(2\|x-y\|) = \frac{J_{\frac{1}{2}}(2\pi\|x-y\|)}{\|x-y\|^{\frac{1}{2}}}\,,$$

where the third equality is obtained using 10.1.1 and 10.1.11 in [1].

If $n = 2$, then in polar coordinates (ρ,θ):

$$\begin{aligned}
\int_{\|\xi\|<1} e^{2i\pi\langle\xi,x-y\rangle}d\xi &= \int_0^1\int_0^{2\pi} e^{2i\pi r\rho\cos\theta}d\theta\rho d\rho\,,\\
&= 2\pi\int_0^1 J_0(2\pi r\rho)\rho d\rho \text{ by 9.1.21 in [1]},\\
&= \frac{1}{2\pi r^2}\int_0^{2\pi r} uJ_0(u)du\,,\\
&= -\frac{1}{r}J_0'(2\pi r) \text{ since by 9.1 in [1] } (uJ_0(u))' = -uJ_0(u)\,,\\
&= \frac{J_1(2\pi r)}{r} \text{ by 9.1.28 in [1].}
\end{aligned}$$

For higher orders we proceed by induction on n, noting that if

$$h(u) = \frac{J_{\frac{n-2}{2}}(2\pi\sqrt{u})}{u^{\frac{n-2}{4}}}\,,$$

then

$$\begin{aligned}
h'(u) &= \frac{J'_{\frac{n-2}{2}}(2\pi\sqrt{(u)})\frac{\pi}{\sqrt{(u)}}u^{\frac{n-2}{4}} - J_{\frac{n-2}{2}}(2\pi\sqrt{u})\frac{n-2}{4}u^{\frac{n-2}{4}-1}}{u^{\frac{n-2}{2}}}\,,\\
&= \frac{2\pi\sqrt{u}J'_{\frac{n-2}{2}}(2\pi\sqrt{u}) - \frac{n-2}{2}J_{\frac{n-2}{2}}(2\pi\sqrt{u})}{2u^{\frac{n-2}{4}+1}}\,,\\
&= -\frac{\pi\sqrt{u}J_{\frac{n}{2}}(2\pi\sqrt{u})}{u^{\frac{n-2}{4}+1}} \text{ according 9.1.27 in [1]}\,,\\
&= -\frac{\pi J_{\frac{n}{2}}(2\pi\sqrt{u})}{u^{\frac{n}{2}}}\,.
\end{aligned}$$

Now invoking lemma 19 yields the result. Finally, to obtain a formula in terms of the variable r instead of r^2, notice that $d(r^2) = rdr$, and this implies that for odd values of n

$$k(x,y) = 2(-1)^{\frac{n-1}{2}} \left(\frac{1}{r}\frac{d}{dr}\right)^{\frac{n-1}{2}} \operatorname{sinc}(2r)\,.$$

□

Appendix B

Bessel kernels in high dimension

In high dimension n, the Bessel kernels rescaled by a factor $\sqrt{n}$ converge to a Gaussian function. This fact was pointed out by von Neumann, and a detailed proof was given by Schoenberg [26]. For the sake of completeness, we reproduce his proof here. For a positive real number z, let $z!$ denote the number

$$\Gamma(z+1) = \int_0^{+\infty} e^{-t} t^z dt$$

Proposition 21. *For $n \in \mathbb{N}\backslash\{0\}$ and $r \geq 0$, let*

$$K_n(r) = \frac{J_{\frac{n}{2}}(2\pi r)}{r^{\frac{n}{2}}}\,.$$

Then

$$\lim_{n\to+\infty} \frac{K_n(\sqrt{2n}r)}{K_n(0)} = e^{-2\pi r^2}$$

uniformly for all $r \geq 0$. Moreover,

$$K_n(0) = \frac{\pi^{\frac{n}{2}}}{(\frac{n}{2})!}$$

is equal to the volume V_n of the unit ball in $\mathbb{R}^n$.

Proof. The Bessel functions of the first type are analytic on the real line, and their power series expansion is [1]:

$$J_\nu(2\pi r) = \sum_{l\geq 0} \frac{(-1)^l}{2^{2l+\nu} l!(\nu+l)!} (2\pi r)^{2l+\nu}\,.$$

Consequently,

$$\begin{aligned} K_n(\sqrt{2n}r) &= \sum_{l\geq 0} \frac{(-1)^l (2\pi)^{2l+\frac{n}{2}}}{2^{2l+\frac{n}{2}} l!\,(\frac{n}{2}+l)!} 2^l n^l r^{2l}\,, \\ &= \frac{\pi^{\frac{n}{2}}}{(\frac{n}{2})!} \sum_{l\geq 0} \frac{(\frac{n}{2})^l}{(\frac{n}{2}+1)\ldots(\frac{n}{2}+l)} \frac{(-4\pi^2)^l}{l!} r^{2l}\,. \end{aligned} \tag{B.1}$$

Since for each value of $l \geq 0$,

$$\lim_{n \to +\infty} \frac{\left(\frac{n}{2}\right)^l}{\left(\frac{n}{2}+1\right) ... \left(\frac{n}{2}+l\right)} = 1\,,$$

and for all $l \geq 0$ and $n > 0$

$$\frac{\left(\frac{n}{2}\right)^l}{\left(\frac{n}{2}+1\right) ... \left(\frac{n}{2}+l\right)} \leq 1\,,$$

it can easily be checked that

$$\lim_{n \to +\infty} \frac{1}{\pi^{\frac{n}{2}}} \left(\frac{n}{2}\right)! K_n(\sqrt{2n}r) = e^{-2\pi r^2}$$

uniformly for r in a bounded interval.

To prove the uniform convergence for all $r \geq 0$, it suffices to bound $2^{\frac{n}{2}} \left(\frac{n}{2}\right)! K_n(\sqrt{2n}r)$ for all n by the same function tending to zero at $+\infty$.

By definition,

$$K_n(r) = \int_{\|\xi\|<1} e^{2i\pi r\langle u,\xi\rangle} d\xi\,,$$

where u is any unit vector. Notice that from (B.1),

$$K_n(0) = \frac{\pi^{\frac{n}{2}}}{\left(\frac{n}{2}\right)!}\,,$$

and in particular the volume V_n of the unit ball in dimension n is equal to this number. By decomposing the unit ball in $\mathbb{R}^n$ into $(n-1)-$dimensional balls, we obtain

$$K_n(r) = \int_{-1}^{1} e^{2i\pi r\rho} \left(\sqrt{1-\rho^2}\right)^{n-1} V_{n-1} d\rho = 2V_{n-1} \int_0^1 \cos(2\pi r\rho) \left(\sqrt{1-\rho^2}\right)^{n-1} d\rho\,,$$

It follows that

$$K_n(0) = 2V_{n-1} \int_0^1 \left(\sqrt{1-\rho^2}\right)^{n-1} d\rho\,,$$

thus

$$\frac{K_n(\sqrt{2n}r)}{K_n(0)} = \int_0^1 \cos(2\pi\sqrt{2n}r\rho) \frac{\left(1-\rho^2\right)^{\frac{n-1}{2}}}{\int_0^1 \left(1-s^2\right)^{\frac{n-1}{2}} ds} d\rho\,.$$

For $\rho \in [0,1]$, define the variable $t \in [0,1]$ by

$$t = \int_0^\rho \frac{\left(1-t^2\right)^{\frac{n-1}{2}}}{\int_0^1 \left(1-s^2\right)^{\frac{n-1}{2}} ds} dt\,,$$

It can be verified that since $\log(1-s^2) \geq -s^2 - \frac{s^4}{2}$,

$$\begin{aligned}
\rho'(t) &\geq \int_0^1 (1-s^2)^{\frac{n-1}{2}} ds\,, \\
&\geq \int_0^1 e^{-\frac{n-1}{2}(s^2+\frac{s^4}{2})} ds\,, \\
&\geq \int_0^1 e^{-ns^2} ds\,, \\
&\geq \frac{1}{\sqrt{n}} \int_0^{\sqrt{n}} e^{-s^2} ds\,, \\
&\geq \frac{C}{\sqrt{n}}\,.
\end{aligned}$$

Therefore

$$\begin{aligned}
\frac{K_n(\sqrt{2n}r)}{K_n(0)} &= \int_0^1 \cos(2\pi\sqrt{2n}r\rho(t))dt\,, \\
&= \int_0^1 \frac{1}{2\pi\sqrt{2n}r\rho'(t)} \cos(2\pi\sqrt{2n}r\rho(t))2\pi\sqrt{2n}r\rho'(t)dt\,, \\
&\leq \frac{C'}{r} \int_0^1 \cos(2\pi\sqrt{2n}r\rho(t))2\pi\sqrt{2n}r\rho'(t)dt\,, \\
&\leq \frac{C'}{r}\,.
\end{aligned}$$

since $\rho(0) = 0$ and $\rho(1) = 1$.

Finally, it has been proven that for all $n \geq 1$,

$$\frac{K_n(\sqrt{2n}r)}{K_n(0)} \leq \frac{C'}{r}\,,$$

and the uniform convergence follows. □

Bibliography

[1] M. Abramowitz and I. Stegun, "Handbook of mathematical functions", Dover, 1965.

[2] N. Aronszajn, "Theory of reproducing kernels", *Transactions of the American Mathematical society*, Vol. 68, No. 3, May 1950.

[3] M. Belkin, "Problems of Learning on Manifolds", *Ph.D. Dissertation*, August 2003

[4] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation", *Neural Computation*, Vol 13, pp 1373-1396, 2003.

[5] R. Bellman, "Adaptive control process: a guided tour", *Princeton University Press*, 1961.

[6] S. Bochner, "Hilbert distances and positive definite functions", *the Annals of Mathematics*, Vol. 42, No. 3, July 1941.

[7] F. R. K. Chung, "Spectral graph theory", *CBMS regional conference series in mathematics*, AMS 1997

[8] R. R. Coifman, "Challenges in Analysis", *Geometric and Functional Analysis*, Special volume "GAFA2000 - visions in Mathematics, towards 2000", pp 471-480, 2000.

[9] D. L. Donoho, "High-dimensional data analysis: the curses and blessings of dimensionality", *Aide-mémoire, MAS conference: mathematical challenges of the 21st century*, August 2000.

[10] D. L. Donoho, "The Kolmogorov sampler", *technical report, Stanford University*, 2002.

[11] D. L. Donoho and C. Grimes, "Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data", *technical report, Stanford Statistics Department*, 2003.

[12] G. B. Folland, "Introduction to partial differential equations", *Princeton University Press*, 1976.

[13] A. Fireze, R. Kannan, S. Vempala, "Fast monte-carlo algorithms for finding low-rank approximations", *Proc. IEEE Foundations of Computer Science*, 1998.

[14] D. B. Graham and N. M. Allinson, "Characterizing virtual eigensignatures for general purpose face recognition", in *in Face Recognition: From Theory to Applications NATO ASI Series F, Computer and Systems Sciences*, Vol. 163, pp 446-456, 1998.

[15] J. Ham, D. D. Lee, S. Mika, and B. Schölkopf, "A kernel view of the dimensionality reduction of manifolds", *technical report TR-110*, Max Planck Institute for Biological Cybernectics, July 2003.

[16] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning", *Springer series in statistics*, 2001.

[17] P. S. Huggins and S. W. Zucker, "Representing Edge Models via Local Principal Component Analysis", *Computer Vision - ECCV 2002 : 7th European Conference on Computer Vision, Copenhagen, Denmark* Proceedings, Part I, May 2002.

[18] W. Johnson and J. Lindenstrauss, "Extensions of Lipschitz maps into a Hilbert space", *Contemporary Mathematics*, Vol. 26, pp 189-206, 1984.

[19] H. J. Landau and H. O. Pollak, "Prolate spheroidal wave functions, Fourier analysis and uncertainty II", *The Bell System technical journal*, Vol. 40, pp 65-84, January 1961.

[20] N. Linial, E. London and Y. Rabinovitch, "The geometry of graphs and some of its algorithmic applications", *Combinatorica*, Vol. 15, pp 215-245, 1995.

[21] W. H. Press, S. A. Teukolsky, W. Vetterling and B. P. Flannery, "Numerical recipes in C: the art of scientific computing", *Cambridge University Press*, 1992.

[22] G. Reinert, S. Schbath and M. S. Waterma, "Probabilistic and statistical properties of words: an overview", *Journal of Computational Biology*, Vol. 7, pp 1-46, 2000.

[23] S. Rosenberg, "The Laplacian on a Riemannian manifold", *Cambridge University Press*, 1997.

[24] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by local linear embedding", *Science*, Vol. 290, Issue 550, pp 2323-2326, December 2000.

[25] I. J. Schoenberg, "On certain metric spaces arising from Euclidean spaces by a change of metric and their imbedding in Hilbert space", *The Annals of Mathematics*, 2nd ser., Vol. 38, No. 4, pp 787-793, October 1937.

[26] I. J. Schoenberg, "Metric spaces and completely monotone functions", *The Annals of Mathematics*, 2nd ser., Vol. 39, No. 4, pp 811-841, October 1938.

[27] J. Shi and J. Malik, "Normalized cut and image segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, Issue 8, August 2000.

[28] D. Slepian and H. O. Pollak, "Prolate spheroidal wave functions, Fourier analysis and uncertainty I", *The Bell System technical journal*, Vol. 40, pp 43-64, January 1961.

[29] D. Slepian, "Prolate spheroidal wave functions, Fourier analysis and uncertainty IV: extensions to many dimensions; generalized prolate spheroidal wave functions", *The Bell System technical journal*, Vol. 43, pp 3009-3058, November 1964.

[30] D. Slepian, "Prolate spheroidal wave functions, Fourier analysis and uncertainty V: the discrete case", *The Bell System technical journal*, Vol. 57, pp 1371-1430, May-June 1978.

[31] O. G. Smolyanov, H. V. Weizsäcker and O. Wittich, "Brownian motion on a manifold as limit of stepwise conditioned standard Brownian motions", *Volume in Honour of S. Albeverio's 60th Birthday*, 2000

[32] G. W. Stewart, "Perturbation theory for the singular value decomposition", *technical report CS-TR 2539, university of Maryland*, September 1990.

[33] J. B. Tenenbaum, V. de Silva and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction", *Science*, Vol. 290, Issue 5500, December 2000.

[34] Z. Zhang and H. Zha, "Principal manifolds and nonlinear dimension reduction via Local Tangent Space Alignment", *technical report, CSE-02-019, Department of Computer Science & Engineering, Pennsylvania State University*, 2002.