Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea in Matematica

DIFFUSION MAPS FOR DIMENSIONALITY REDUCTION

Tesi di Laurea Magistrale

Relatore:
Chiar.ma Prof.ssa
GOVANNA CITTI

Presentata da:
MARIA VIRGINIA BOLELLI

Anno Accademico 2018-2019

Introduction

In this thesis we present the diffusion maps, a framework based on diffusion processes for finding meaningful geometric descriptions of data sets. A diffusion process can be described via an iterative application of the heat kernel which has two main characteristics: it satisfies a Markov semigroup property and its level sets encode all geometric information of the space. The kernel acts by integration with respect to a measure, suitably related to the features of the space.

The Markov property ensures that the behavior of a process in the future only depends on its value at the initial time t_0, not on the values it attained in the past. First order ODEs trivially have this characteristic, since the solution of a Cauchy problem is unique

$$\begin{cases} u'(t) = Au(t), \\ u(t_0) = u_0. \end{cases}$$

The same peculiarity is shared also by the heat flow, where the linear operator $u \to Au$ is replaced by the Laplace operator $u \to \Delta u$. In chapter 1, we will describe stochastic processes, Markov semigroups and their infinitesimal generator, which are processes that have this memoryless property.

Geometrical properties of the space are classically described by Riemannian geometry. Diffusion in this setting is expressed in terms of a heat operator, associated to the Laplacian operator also called Laplace-Beltrami operator. From the general theory of self-adjoint and nonnegative operators on a Riemannian manifold, it follows the existence of a heat kernel $e^{-t\Delta}$, which is a compact operator, and it contains as much geometric information as the metric itself. A crucial result is the Hodge theorem for compact, connected, oriented Riemannian manifold, which states that there is an orthonormal basis of $L^2(\mathcal{M}, g)$ of eigenfunctions of the Laplacian. Projection on this eigenfunctions is at the basis of dimension reduction results.

The core of the thesis is the description of diffusion maps, introduced by Coifman and Lafon [14][15][16]. They are a general framework able to associate to a data set a probability measure, which describe its density, and a kernel, which contains its geometric information. Iterative application of the kernel induces a diffusion on the data. The diffu-

sion kernels a^m, obtained after m iterations of the process, will be compact operators, all with the same countable family of eigenfunctions. The projection on the first eigenvectors naturally leads to a dimensionality reduction algorithm, which maps high-dimensional vector into a low-dimensional one with minimal loss of information. These type of processes are very useful since high dimensionality is in general an obstacle to an efficient processing of the data. Therefore, if the number of variables which efficiently describe a data set is small, it is reasonable to transform the representation of the data into a more efficient low dimensional description.

This treatment will be organized in two main sections. In the first one, we focus on defining the diffusion process on a measure space (Ω, μ), where Ω is any data set and μ can be interpreted as a probability measure. In this setting, the local geometry is described by a symmetric and positive preserving function $k(x, y)$ and it allows to define an appropriate kernel

$$a(x, y) = \frac{k(x, y)}{d(x)d(y)}$$

with $d^2(x) = \int_\Omega k(x, y)d\mu(y)$. This normalization of the kernel is called *graph Laplacian*. Hence, thanks to $a(x, y)$, we can construct the diffusion operator

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

whose properties enable us to find an eigendecomposition of the kernel using an orthonormal basis for $L^2(\Omega, \mu)$, namely $a(x, y) = \sum_{i\geq 0} \lambda_i \phi_i(x)\phi_i(y)$. If for a given accuracy δ we retain only the eigenvalues $\lambda_0, \ldots, \lambda_{p-1}$ that, raised to a certain power m exceed this accuracy, it is possible, using the corresponding eigenfunctions $\phi_0, \ldots, \phi_{p-1}$, to embed the data points into $\mathbb{R}^p$:

$$\begin{aligned} \Phi_{p,m} \colon \Omega &\longrightarrow \mathbb{R}^p \\ x &\mapsto \begin{pmatrix} \lambda_0^m \phi_0(x) \\ \lambda_1^m \phi_1(x) \\ \vdots \\ \lambda_{p-1}^m \phi_{p-1}(x) \end{pmatrix}. \end{aligned}$$

The maps of the family $\Phi_{p,m}$ are called *diffusion maps*, while each component of a map is called *diffusion coordinate*. In order to understand why this embedding makes sense, we show that the weighted Euclidean distance on this space, namely $\|x - y\|^2 = \sum_{i=0}^{p-1} \lambda_i^m (\phi_i(x) - \phi_i(y))^2$, approximates the *diffusion distance* on the dataset Ω, that is

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

which is strictly related to the number of paths connecting points x and y. In particular, this approximation illustrates that nearby points in $\mathbb{R}^p$ are correlate to nearby points in

the data space Ω, where, the notion of proximity is described by a large number of paths, arisen from the diffusion process.

The second section is devoted to the case in which $\Omega = \mathcal{M}$ is a Riemannian submanifold of $\mathbb{R}^n$. Using the procedure described above, we can define a family of diffusion operators $A_{\varepsilon,\alpha}$ parametrized by a parameter $\alpha \in \mathbb{R}$ and $\varepsilon > 0$, where the kernel is defined renormalizing a rotation-invariant function $h(\frac{\|x-y\|^2}{\varepsilon})$, infinitely differentiable with an exponential decay. The main result will concern the following operator, interpreted as the infinitesimal generator of the diffusion operators

$$L_{\varepsilon,\alpha} = \frac{\mathrm{Id} - A_{\varepsilon,\alpha}}{\varepsilon}$$

as $\varepsilon \to 0$. In particular, we show that for a suitable function f

$$\lim_{\varepsilon\to 0} L_{\varepsilon,\alpha} f = \frac{\Delta(f q^{1-\alpha})}{q^{1-\alpha}} - \frac{\Delta(q^{1-\alpha})}{q^{1-\alpha}} f, \tag{1}$$

where $q(x)$ is the density of points on $\mathcal{M}$. Two values of α are investigated in details: $\alpha = 0$ and $\alpha = 1$. In the first case, i.e. $\alpha = 0$, (1) leads to

$$\lim_{\varepsilon\to 0} L_{\varepsilon,0} f(x) = \frac{\Delta f q}{q} - \frac{\Delta q}{q} f = \Delta f + 2\langle \frac{\operatorname{grad} q}{q}, \operatorname{grad} f\rangle. \tag{2}$$

Formula (2) proves that, when the density is uniform, meaning q is constant, $L_{\varepsilon,0}$ is the Laplace-Beltrami operator on $\mathcal{M}$. However, the density is not always uniform. In that case, we see that, setting $\alpha = 1$, we get:

$$\lim_{\varepsilon\to 0} L_{\varepsilon,1} f = \lim_{\varepsilon\to 0} \Delta f + R_\varepsilon = \Delta f. \tag{3}$$

Formula (3) means that, even though the density may not be uniform, we are able to recover the Laplace-Beltrami operator. For this reason, the normalization of the kernel using the parameter $\alpha = 1$ is called *Laplace-Beltrami*. Moreover, as a byproduct, in the $\alpha = 1$ case, it is possible to approximate the Neumann heat kernel $e^{-t\Delta}$ on $L^2(\mathcal{M})$, using the diffusion operator $A_{\varepsilon,1}$:

$$\lim_{\varepsilon\to 0} A_{\varepsilon,1}^{\frac{t}{\varepsilon}} = e^{-t\Delta}.$$

Finally, we illustrate the ideas previously discussed by numerical examples studying sets of data. Usually these quantities are composed by a finite numbers of points, that is, we deal with $\Omega = \{x_1, \ldots, x_N\}$. The first thing to do is to handle the implementation problems linked with the discretization of the quantities involved in our algorithm:

$$d(x_i) = \sum_{j=1}^{N} k_\varepsilon(x_i, x_j) \quad , \quad a(x_i, x_j) = \frac{k(x_i, x_j)}{d(xi)d(x_j)}$$

for all x_i , $x_j \in \Omega$. Then we generate examples of datasets and, using a Gaussian kernel,

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\varepsilon}\right)$$

we compute the eigenvectors and eigenvalues of the diffusion operator. We can achieve different interesting results:

- **dimensionality reduction**: we consider randomly arranged pictures of the word "$3D$", viewed under different angles. These pictures can be reorganized using the first non trivial eigenvectors: they recover the main parameters that describe these images, namely the angle of the rotation along the x-axis and along the y-axis.

- **representations of complex geometric structures**: we consider a set of unordered points $\Omega \in \mathbb{R}^3$ on a curve. We then compute and plot the embedding built using the first two nontrivial eigenvectors: we obtain the points reorganized on a closed curve in a coherent ways with the organization of the points following the curve.

- **global geometric information from local structures**: we generate a set that is the union of clusters. Studying the diffusion operator, we notice that it is possible to recover a block structure that gives us information about the whole set.

- **robustness to noise perturbation**: we consider a perturbed version of a set, for example a manifold, and we plot the embedding. The results obtained are not affected by the noise, being actually unalterated.

The elaborate can be divided in three main parts. The first part, covering chapters 1 and 2, describes the Markov semigroup property and the geometric properties of heat kernels in the classical Riemannian setting. Chapter 3 consists of its central core, the description of diffusion maps, and chapter 4 contains numerical examples and results.

Contents

List of Figures

Chapter 1

Introduction to Markov semigroups

The main purpose of this chapter is to introduce the notion of Markov semigroup. In order to do so, we first recall the definition of stochastic process and then, the one of Markov property.

1.1 Stochastic Process

Definition 1.1. *(Stochastic process)* Let $(\Omega, \mathcal{F}, P)$ denote a probability space and I an arbitrary nonempty index set. A stochastic process is a parametrized collection of random variables $\{X_t\}_{t\in I}$ assuming values in $\mathbb{R}^n$.

As a result, a stochastic process is a function of the variable (t, ω), where $t \in I \subseteq \mathbb{R}$ and $\omega \in \Omega$. Hence, it is sometimes usual to fix $\omega \in \Omega$ and to consider the function:

$$\begin{aligned} X_\cdot(\omega) \ : I &\longrightarrow \mathbb{R}^n \\ t &\mapsto X_t(\omega), \end{aligned}$$

for all $\omega \in \Omega$. These are called *sample paths* of the stochastic process.

One of the most important results about a stochastic process is the Kolmogorov's extension theorem. To understand it, we have to introduce the *finite-dimensional distributions*

of a stochastic process $\{X_t\}_{t\in I}$:

$$\begin{aligned} P(X_t \leq x) =& F_t(x), \\ P(X_{t_1} \leq x_1, X_{t_2} \leq x_2) =& F_{t_1,t_2}(x_1, x_2), \\ &\vdots \\ P(X_{t_1} \leq x_1, \ldots, X_{t_n} \leq x_n) =& F_{t_1,\ldots,t_n}(x_1, \ldots, x_n), \\ &\vdots \end{aligned}$$

where $t, t_i \in [t_0, T] \subseteq I$, $x, x_i \in \mathbb{R}^n, n \geq 1$ and $F_{t_1,\ldots,t_n}$ is a distribution function for all $n \geq 1$. This system of distributions satisfies the following two conditions:

1) *Symmetry condition:* if $\{i_1, \ldots, i_n\}$ is a permutation of the number $\{1, \ldots, n\}$ then, for arbitrary instants and $n \geq 1$,

$$F_{t_{i_1},\ldots,t_{i_n}}(x_{i_1}, \ldots, x_{i_n}) = F_{t_1,\ldots,t_n}(x_1, \ldots, x_n).$$

2) *Compatibility condition:* for $m < n$ and arbitrary $t_{m+1}, \ldots, t_n \in [t_0, T]$,

$$F_{t_1,\ldots,t_m,t_{m+1},\ldots,t_n}(x_1, \ldots, x_m, \infty, \ldots, \infty) = F_{t_1,\ldots,t_m}(x_1, \ldots, x_m).$$

Theorem 1.1.1 (Kolmogorov's extension theorem)**.** *For every family of distribution functions that satisfy the symmetry and compatibility conditions, there exists a probability space $(\Omega, \mathcal{F}, P)$ and a stochastic process $\{X_t\}_{t\in[t_0,T]}$ defined on it that possesses the given distributions as a finite dimensional distributions.*

This means that is equivalent to give a family of random variables or a family of distributions.

1.2 The Markov Property

We would like to give a formal mathematical definition of the Markov property, which can be roughly presented as follows (see [1]) : *"if the state of the process at a particular time s (the present) is known, additional information regarding the behavior of the process at $r < s$ (the past) has no effect on our knowledge of the probable development of the process at $t > s$ (in the future)"* .

Suppose that $I = [t_0, T]$ and let $\{X_t\}_{t\in[t_0,T]}$ be a stochastic process defined on a certain probability space $(\Omega, \mathcal{F}, P)$. We define $\mathcal{F}([t_1, t_2]) = \mathcal{F}(X_t\ ,\ t_1 \leq t \leq t_2)$ to be the smallest sub σ-algebra of $\mathcal{F}$ with respect to which all the random variables X_t for $t_1 \leq t \leq t_2$ are

measurable. In other words, $\mathcal{F}([t_1, t_2])$ contains the "history" of the process from time t_1 to time t_2 and is generated by the cylinder events

$$\{\omega : X_{s_1}(\omega) \in B_1, \ldots, X_{s_n}(\omega) \in B_n\} = (X_{s_1} \in B_1, \ldots, X_{s_n} \in B_n),$$

with $t_1 \leq s_1 \leq \ldots \leq s_n \leq t_2$, $B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R}^n)$, where $\mathcal{B}(\mathbb{R}^n)$ is the Borel set on $\mathbb{R}^n$.

Definition 1.2. *(Markov process)* A stochastic process $\{X_t\}_{t\in[t_0,T]}$ defined on the probability space $(\Omega, \mathcal{F}, P)$ is called a Markov process if the following *Markov property* is satisfied:

$$P(X_t \in B|\mathcal{F}([t_0, s])) = P(X_t \in B|X_s), \tag{1.1}$$

with $t_0 \leq s \leq t \leq T$ and $B \in \mathcal{B}(\mathbb{R}^n)$.

There are various equivalent formulations of the Markov property:

Theorem 1.2.1. *The following conditions are equivalent:*

- $\{X_t\}_{t\in[t_0,T]}$ *is a Markov's process;*
- *for* $t_0 \leq s \leq t \leq T$ *and* Y *a random variable* $\mathcal{F}([t, T])$*-measurable and integrable*

$$E(Y|\mathcal{F}([t_0, s])) = E(Y|X_s); \tag{1.2}$$

- *for* $t_0 \leq s \leq t \leq T$, $A \in \mathcal{F}([t, T])$*-measurable and integrable,*

$$P(A|\mathcal{F}([t_0, s])) = P(A|X_s);$$

- *for* $t_0 \leq t_1 \leq t \leq t_2 \leq T$, $A_1 \in \mathcal{F}([t_0, t_1])$ *and* $A_2 \in \mathcal{F}([t_2, T])$,

$$P(A_1 \cap A_2|X_t) = P(A_1|X_t)P(A_2|X_t);$$

- *for* $n \geq 1$, $t_0 \leq t_1 \leq \ldots \leq t_n \leq t \leq T$ *and* $B \in \mathcal{B}(\mathbb{R}^n)$,

$$P(X_t \in B|X_{t_1}, \ldots, X_{t_n}) = P(X_t \in B|X_{t_n}).$$

Proof of these assertions can be found, for example, in [2, p. 80-85].

1.3 Transition Probabilities

Definition 1.3. *(Transition probability)* A function $P(s, x, t, B)$ is called transition probability if satisfies the following properties:

1) $P(s, x, t, \cdot)$ is a probability on $\mathcal{B}(\mathbb{R}^n)$ for fixed $s \leq t$ and $x \in \mathbb{R}^n$.

2) $P(s,\cdot,t,B)$ is a $\mathcal{B}(\mathbb{R}^n)$-measurable function for fixed $s \leq t$ and $B \in \mathcal{B}(\mathbb{R}^n)$. Furthermore,
$$P(s,x,t,B) = \int_{\mathbb{R}^n} P(u,y,t,B)P(s,x,u,dy), \tag{1.3}$$
this identity is called *Chapman-Kolmogorov equation.*

3) For all $s \in [t_0,T]$ and $B \in \mathcal{B}(\mathbb{R}^n)$, we have
$$P(s,x,s,B) = \mathbb{1}_B(x) = \begin{cases} 1 & \text{for } x \in B \\ 0 & \text{for } x \notin B. \end{cases}$$

Remark 1.1. If $\{X_t\}_{t\in[t_0,T]}$ is a Markov process, we have that $P(X_t \in B|X_s)$ is a transition probability of the Markov process X_t for fixed $s,t \in [t_0,T]$. So, we shall also use the notation
$$P(s,x,t,B) = P(X_t \in B|X_s = x),$$
which is the probability that the observed process will be in the set B at time t if at time s, where $s \leq t$, it was in the state x.

Remark 1.2. Let us suppose that the probability $P(s,x,t,\cdot)$ has a density, i.e. for all $x \in \mathbb{R}^n$, and $B \in \mathcal{B}(\mathbb{R}^n)$
$$P(s,x,t,B) = \int_B p(s,x,t,y)dy,$$
where $p(s,x,t,y)$ is a non negative valued function that is measurable with respect to y and whose integral is equal to 1. Then, the Chapman-Kolmogorov equation reduces to
$$\begin{aligned}\int_B p(s,x,t,z)dz =& P(s,x,t,B) = \int_{\mathbb{R}^n} P(u,y,t,B)P(s,x,u,dy) \\ &= \int_{\mathbb{R}^n} P(u,y,t,B)p(s,x,u,y)dy = \int_{\mathbb{R}^n}\int_B p(u,y,t,z)p(s,x,u,y)dzdy \\ &= \int_B \int_{\mathbb{R}^n} p(u,y,t,z)p(s,x,u,y)dydz,\end{aligned}$$
from which it follows that
$$p(s,x,t,z) = \int_{\mathbb{R}^n} p(s,x,u,y)p(u,y,t,z)dy.$$
More precisely, this means that the probability of a transition from x at time s to z at time t is equal to the probability of the transition to y at an intermediate time u, multiplied by the probability of the transition from y at the time u to z at the time t, summed over all intermediate values y.

The transition probabilities for Markov processes are really importants because all finite-dimensional distributions of the process can be obtained from them and from the initial distribution at time t_0.

Theorem 1.3.1. *If $\{X_t\}_{t\in[t_0,T]}$ is a Markov process, $P(s,x,t,B)$ its transition probability and P_{t_0} the distribution of X_{t_0} (i.e. $P_{t_0}(A) = P(X_{t_0} \in A)$, for all $A \in \mathcal{B}(\mathbb{R}^n)$) then for finite dimensional distributions*

$$P(X_{t_1} \in B_1, \ldots, X_{t_n} \in B_n) \ , \ t_0 \leq t_1 < \ldots < t_n \leq T \ , \ B_i \in \mathcal{B}(\mathbb{R}^n),$$

we have

$$P(X_{t_1} \in B_1, \ldots, X_{t_n} \in B_n) =$$
$$\int_{\mathbb{R}^n}\int_{B_1}\cdots\int_{B_{n-1}} P(t_{n-1},x_{n-1},t_n,B_n)\cdot P(t_{n-2},x_{n-2},t_{n-1},dx_{n-1})\cdots P(t_0,x_0,t_1,dx_1)P_{t_0}(dx_0),$$

and hence, in particular,

$$P(X_t \in B) = \int_{\mathbb{R}^n} P(t_0,x,t,B)P_{t_0}(dx).$$

For proof see [6, pp 151].

Thanks to this theorem, we can recover a Markov process from the transition probabilities and from the initial distribution, using Theorem 1.3.1 to construct, from $P(s,x,t,B)$ and P_{t_0}, consistent finite-dimensional distributions and from them, in accordance with Kolmogorov extension theorem, the desired process. In other words, we have proved the following theorem.

Theorem 1.3.2. *Let $P(s,x,t,B)$ denote a transition probability, where $s,t \in [t_0,T]$. Then for every initial probability P_{t_0} on $\mathcal{B}(\mathbb{R}^n)$ there exists a probability space $(\Omega,\mathcal{F},P)$ and a Markov process X_t defined on it, which has transition probability $P(s,x,t,B)$ and for which X_{t_0} has the distribution P_{t_0}.*

This result is very useful because, generally, in applied problems, we have to deal with transition probabilities rather than Markov processes. Hence, it is convenient to know that is possible to construct a Markov process from a family of transition probabilities.

Definition 1.4. (*Homogeneous Markov Process*) A Markov process $\{X_t\}_{t\in[t_0,T]}$ is said to be homogeneous with respect to time if its transition probability $P(s,x,t,B)$ is stationary. In other words, the condition

$$P(s+u,x,t+u,B) = P(s,x,t,B)$$

is identically satisfied for $t_0 \leq s \leq t \leq T$ and $t_0 \leq s+u \leq t+u \leq T$.

In this case, the transition probability is a function only of $x, t-s$, and B. Hence, we can write it in the form

$$P(t-s,x,B) = P(s,x,t,B) \ , \ 0 \leq t-s \leq T-t_0.$$

Therefore, $P(t,x,B)$ is the probability of transition from x to B in time t, regardless of the actual position of the interval of length t on the time axis. For homogeneous processes, the Chapman-Kolmogorov equation becomes

$$P(t-s,x,B) = P(s,x,t,B) = \int_{\mathbb{R}^n} P(u,y,t,B)P(s,x,u,dy) =$$
$$= \int_{\mathbb{R}^n} P(t-u,y,B)P(u-s,x,dy).$$

If we denote with $h = t-u$ and $k = u-s$, we get that the equation (1.3) becomes

$$P(h+k,x,B) = \int_{\mathbb{R}^n} P(h,y,B)P(k,x,dy).$$

As a rule, homogeneous Markov processes are defined on an interval of the form $[t_0,\infty)$, so that the transition probability $P(t,x,B)$ is defined for $t \in [t_0,\infty)$.

Example 1.1 (*Wiener Process*)**.** The Wiener process is a n-dimensional homogeneous Markov process W_t defined on $[0,\infty)$ with stationary transition probability

$$P(t,x,\cdot) = \begin{cases} \mathcal{N}(x, t\,\mathrm{Id}_n), & t > 0 \\ \delta_x(\cdot), & t = 0, \end{cases}$$

where $\mathcal{N}$ is the normal distribution. That is, for $t = 0$, we have the Dirac delta centered at x, while for $t > 0$, we deal with

$$P(t,x,B) = P(W_{t+s} \in B | W_s = x) = \int_B (2\pi t)^{-n/2} e^{\frac{-|y-x|^2}{2t}} dy.$$

By virtue of the familiar formula for gaussian densities

$$\int_{\mathbb{R}^n} n(s,x,z)n(t,z,y)dz = n(s+t,x,y),$$

the Chapman-Kolmogorv equation holds for

$$p(t,x,y) = n(t,x,y) = (2\pi t)^{-n/2} e^{\frac{-|y-x|^2}{2t}}.$$

In general, we take the initial probability P_{t_0} equal to δ_0, that is $W_0 = 0$ and, since

$$n(t,x+z,y+z) = n(t,x,y) \text{ for all } z \in \mathbb{R}^n,$$

we are dealing with a space-wise as well as time-wise homogeneous process. The function W_t is frequently known as mathematical model of the Brownian motion of a free particle in absence of friction.

1.4 Markov Semigroups

Generally speaking, a semigroup $\mathbf{A} = (A_t)_{t\geq 0}$ is a family of operators acting on some suitable function space with the *semigroup property* $A_t \circ A_s = A_{t+s}$ where $t, s \geq 0$ and $A_0 = \mathrm{Id}$. These semigroups appear in the probabilistic context describing the family of laws of Markov process $(X_t)_{t\geq 0}$ living on a measurable space Ω. So, the fundamental object of investigation consist of a family $\mathbf{A} = (A_t)_{t\geq 0}$ of operators defined on some set of real-valued measurable functions on $(\Omega, \mathcal{F})$.

Definition 1.5 (*Markov operator*)**.** The operator A_t defined on some set of real-valued measurable functions on $(\Omega, \mathcal{F})$ is called Markov operator if it satisfies the following properties:

1) *mass conservation*: $A_t 1 = 1$, where 1 is the constant function ;

2) *positive preserving*: if $f \geq 0$, then $A_t f \geq 0$.

Very often property 1) may be relaxed to $A_t 1 \leq 1$.

Definition 1.6 (*Markov semigroup*)**.** A family of operators $\mathbf{A} = (A_t)_{t\geq 0}$ defined on the bounded measurable functions on a state space $(\Omega, \mathcal{F})$ is called a Markov semigroup if A_t is a Markov operator for every $t \geq 0$ and the following properties are satisfied:

1) for every $t \geq 0$, A_t is a linear operator sending bounded measurable functions on $(\Omega, \mathcal{F})$ to bounded measurable functions;

2) *initial condition*: $A_0 = \mathrm{Id}$, the identity operator;

3) *semigroup property*: for every $t, s \geq 0$ we have $A_{t+s} = A_t \circ A_s$;

4) *continuity property*: for every $f \in L^2(\Omega)$, $A_t f$ converges to f in $L^2(\Omega)$ as $t \to 0$.

It is useful to know that Markov operators $A_t,\ t \geq 0$, as given in Definition 1.5, may be represented by stochastic kernel.

Definition 1.7 (*Kernel*)**.** A (non-negative) kernel on $(\Omega, \mathcal{F})$ is a map $k := \Omega \times \mathcal{F} \longrightarrow \mathbb{R}_+$ satisfying the following two conditions:

1) for any fixed set $B \in \Omega$, the function $k(\cdot\ , B)$ is measurable;

2) for any fixed $x \in \Omega$, the set function $k(x, \cdot\ )$ is a measure on $(\Omega, \mathcal{F})$.

A kernel k is called *finite*, if

$$k(x, \Omega) < \infty \qquad \text{for all } x \in \Omega;$$

bounded, if

$$\sup_{x\in\Omega} K(x,\Omega) < \infty;$$

stochastic, if

$$k(x,\Omega) = 1 \qquad \text{for all } x \in \Omega.$$

Thanks to this notion, Markov operators can be represented by probabilistic kernels, which correspond to the transition probabilities of the associated Markov process. Namely, for every bounded measurable function $f\colon \Omega \longrightarrow \mathbb{R}$,

$$A_t f(x) = \int_\Omega f(y) p_t(x,dy), \qquad t \geq 0,\ x \in \Omega, \tag{1.4}$$

where $p_t(x,dy)$ is for every $t \geq 0$ a probability kernel.
The distribution at time t of the underlying Markov process $(X_t)_{t\geq 0}$ starting at x is thus given by the probability $p_t(x,\cdot)$.
Very often the family of kernels $p_t(x,dy)$ have densities with respect to a measure.

Definition 1.8 (*Density kernel*)**.** A Markov semigroup $(A_t)_{t\geq 0}$ on $(\Omega,\mathcal{F})$ is said to admit density kernels with respect to a reference σ-finite measure μ on $\mathcal{F}$ if there exists for every $t \geq 0$ a positive measurable function $p_t(x,y)$ defined on $\Omega\times\Omega$ (up to a set of $\mu\otimes\mu$-measure 0) such that, for every bounded or positive measurable function $f\colon \Omega \to \mathbb{R}$ and (μ-almost) every $x \in \Omega$,

$$A_t f(x) = \int_\Omega f(y) p_t(x,y) d\mu(y).$$

In this case, $\int_\Omega p_t(x,y)d\mu(y) = 1$ for μ-almost every $x \in \Omega$, reflecting the fact that $A_t 1 = 1$. In order to $A_t f$ to make sense for any $f \in L^2(\Omega,\mu)$ in this definition, it is in general required that, for all $t > 0$ and μ-almost every $x \in \Omega$,

$$\int_\Omega p_t(x,y)^2 d\mu(y) < \infty. \tag{1.5}$$

Example 1.2. For the Wiener process W_t we have

$$\begin{aligned} A_t f(x) &= (2\pi t)^{-n/2} \int_{\mathbb{R}^n} f(y) e^{-\frac{|y-x|^2}{2t}} dy \\ &= (2\pi)^{-n/2} \int_{\mathbb{R}^n} e^{-\frac{|z|^2}{2}} f(x+\sqrt{t}z) dz\ ,\ t > 0. \end{aligned}$$

Since now, we have described how to construct a Markov semigroup from a Markov process. Conversely, given a Markov semigroup, the construction of a Markov process associated to it relies on the Chapman-Kolmogorov equations which express the semigroup property from a probabilistic point of view.
Let $\mathbf{A} = (A_t)_{t\geq 0}$ be a Markov semigroup on $L^2(\Omega)$ according to Definition 1.6. The

semigroup property $A_t \circ A_s = A_{t+s}$ translates to the kernel $p_t(x, dy)$ of the representation (1.4) : for all $t, s \geq 0,\ x \in \Omega$

$$p_{t+s}(x, dy) = \int_{z\in\Omega} p_t(z, dy)p_s(x, dz)$$

and when the kernels admit densities,

$$p_{t+s}(x, y) = \int_{\Omega} p_t(z, y)p_s(x, z)d\mu(z).$$

Now, starting from any point $x \in \Omega$, we can construct a Markov process $(X_t)_{t\geq 0}$ on Ω by specifying the distribution of $(X_{t_1}, \ldots, X_{t_k})$ with $0 \leq t_1 \leq \ldots \leq t_k$, see [4, pp 17]. In particular, if we are in $\mathbb{R}^n$ we can recover the finite-dimensional distributions and consequently the Markov process thanks to Theorems 1.3.1 and 1.3.2.

1.5 Infinitesimal generator

A Markov semigroup $\mathbf{A} = (A_t)_{t\geq 0}$, as defined in Definition 1.6, is driven by an operator called the infinitesimal generator of the Markov semigroup. We suppose that our semigroup is defined on the Hilbert space $L^2(\Omega, \mu)$, and we call $\mathcal{D}$ the domain of the semigroup $(A_t)_{t\geq 0}$ on which the derivative at $t = 0$ of A_t exists in $L^2(\Omega, \mu)$.

Definition 1.9 (*Infinitesimal generator*)**.** Let $\mathbf{A} = (A_t)_{t\geq 0}$ be a Markov semigroup with state space $(\Omega, \mathcal{F})$ and a measure μ. The operator that maps $f \in \mathcal{D}$ to the derivative Lf at $t = 0$ of A_t is a linear operator, called the infinitesimal generator L of $\mathbf{A}$ in $L^2(\Omega, \mu)$.

In view of the connection with the Markov process $(X_t)_{t\geq 0}$ associated with the semigroup $\mathbf{A} = (A_t)_{t\geq 0}$, the generator L will also be called the Markov generator of $(X_t)_{t\geq 0}$. The linearity of the operators A_t ,for every $t \geq 0$, together with the semigroup property, shows that L is the derivative of A_t at any time $t > 0$. Namely, for $t, s > 0$,

$$\frac{1}{s}[A_{t+s} - A_t] = A_t\frac{1}{s}[A_s - I_d] = \left(\frac{1}{s}[A_s - I_d]\right) A_t.$$

Letting $s \to 0$ then yields

$$\partial_t A_t = A_t L = LA_t. \tag{1.6}$$

The semigroup $(A_t)_{t\geq 0}$ will often be called the *heat semigroup or heat flow with respect to the generator* L and thus solving the heat equation (1.6).

Example 1.3. For the n-dimensional Wiener process W_t, we must calculate

$$Lf(x) = (2\pi)^{-n/2} \lim_{t\searrow 0} \frac{\int_{\mathbb{R}^n} e^{-|z|^2/2}(f(x + \sqrt{t}z) - f(x))dz}{t}.$$

For this we use Taylor's theorem, which for every twice continuous partially differentiable function f yields

$$f(x+\sqrt{t}z) - f(x) = \sqrt{t}\sum_{i=1}^{d} z_i f_i(x) + \frac{t}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} z_i z_j f_{i,j}(x) + \frac{t}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} z_i z_j (f_{i,j}(\bar{x}) - f_{i,j}(x)),$$

where $\bar{x}$ is a point between x and $x+\sqrt{t}z$. When we substitute this into the expression given above for $Lf(x)$, we get

$$Lf = \frac{1}{2}\sum_{i=1}^{d} \frac{\partial^2 f}{\partial x_i^2} = \frac{1}{2}\Delta f,$$

where Δ is the Laplace operator.

Chapter 2

Elements of Riemannian geometry

We recall here a few elements of Riemannian geometry, and in particular the notion of Laplace and Heat operator in this setting, so that we will generalize it to diffusion maps in the next chapter. The presentation is taken from [10] [11] [12] [13].

2.1 Riemannian Manifolds

Definition 2.1 (*Riemannian manifold*)**.** A Riemannian manifold (M, g) is a smooth manifold M with a family of smoothly varying positive definite inner products $g = g_p$ on T_pM for each $p \in M$. The family g is called *Riemannian metric*.

Example 2.1. (1) On $\mathbb{R}^n$, the standard Riemannian metric is given by the standard inner product $g_p(v, w) = v \cdot w$ for all $v, w \in T_p\mathbb{R}^n$ for all $p \in \mathbb{R}^n$. We call $\mathbb{R}^n$ with this metric *Euclidean space*.

(2) If M is a submanifold of Euclidean space, then M has a natural Riemann metric given by $g_p(v, w) = v \cdot w_{|T_pM}$. This so called *induced metric* is the metric used in the classical theory of curves and surfaces in Euclidean three space.

Let us indicate by $\chi(M)$ the set of all vector fields of class C^∞ on M and by $\mathcal{D}(M)$ the ring of real valued functions of class C^∞ defined on M.

Definition 2.2 (*Connection*)**.** Let M be a differentiable manifold and let ∇ be a mapping

$$\begin{aligned} \nabla \colon \chi(M) \times \chi(M) &\to \chi(M) \\ (X, Y) &\mapsto \nabla_X Y. \end{aligned}$$

We say that ∇ is a connection on M if it satisfies the following properties:

1. $\nabla_{fX+gY}Z = f\nabla_X Z + g\nabla_Y Z$;

2. $\nabla_X(Y+Z) = \nabla_X Y + \nabla_X Z$;

3. $\nabla_X(fY) = X(f)Y + f\nabla_X Y$;

with $X, Y, Z \in \chi(M)$ and $f, g \in C^\infty(M)$.

Proposition 2.1.1 (Covariant derivative)**.** *Let M be a differentiable manifold with a connection ∇. There exists a unique correspondence which associates to a vector field V along the differentiable curve $c : I \longrightarrow M$ another vector field $\frac{DV}{dt}$ along c called the covariant derivative of V along c such that:*

(a) $\frac{D(V+W)}{dt} = \frac{DV}{dt} + \frac{DW}{dt}$ *where W is a vector field along c;*

(b) $\frac{D(fV)}{dt} = \frac{df}{dt}V + f\frac{DV}{dt}$ *where f is a differentiable function on I;*

(c) if V is induced by a vector field $Y \in \chi(M)$, i.e. $V(t) = Y(c(t))$, then $\frac{DV}{dt} = \nabla_{dc/dt}Y$.

Proof. First, suppose that exists a correspondence satisfying the conditions $(a), (b), (c)$. Let $\varphi : U \subset \mathbb{R}^n \longrightarrow M$ be a system of coordinates with $c(I) \cap \varphi(U) \neq \emptyset$, let $(c^1(t), \ldots, c^n(t))$ be the local expression of $c(t)$, $t \in I$ and $X_i = \frac{\partial}{\partial x^i}$. Then we can express the field V locally as $V = \sum_{j=1}^n v^j X_j$, where n is the dimension of the manifold, $v^j = v^j(t)$ and $X_j = X_j(c(t))$. By (a) and (b) we have

$$\frac{DV}{dt} = \sum_j \frac{dv^j}{dt} X_j + \sum_j v^j \frac{DX_j}{dt}.$$

By (c) and Definition 2.2, since $X_j = X_j(c(t))$ we have:

$$\begin{aligned}\frac{DX_j}{dt} = \nabla_{dc/dt}X_j =& \nabla_{\sum \frac{dc^i}{dt}X_i} X_j \\ =& \sum_i \frac{dc^i}{dt}\nabla_{X_i}X_j, \qquad i,j = 1,\ldots n.\end{aligned}$$

Therefore,

$$\frac{DV}{dt} = \sum_j \frac{dv^j}{dt} X_j + \sum_{i,j} \frac{dc^i}{dt} v^j \nabla_{X_i} X_j. \tag{2.1}$$

The expression (2.1) shows that if there is a correspondence satisfying the conditions above, then such a correspondence is unique.

To show existence, define $\frac{dV}{dt}$ in $\varphi(U)$ as in (2.1). It is easy to verify that this quantity verifies the desired properties:

(a): Suppose V and W vector fields along c, then we have

$$\begin{aligned}\frac{D(V+W)}{dt} &= \sum_j \frac{d(v^j+w^j)}{dt}X_j + \sum_{i,j}\frac{dc^i}{dt}(v^j+w^j)\nabla_{X_i}X_j \\ &= \sum_j \frac{dv^j}{dt}X_j + \sum_j \frac{dw^j}{dt}X_j + \sum_{i,j}\frac{dc^i}{dt}v^j\nabla_{X_i}X_j + \sum_{i,j}\frac{dc^i}{dt}w^j\nabla_{X_i}X_j \\ &= \sum_j \frac{dv^j}{dt}X_j + \sum_{i,j}\frac{dc^i}{dt}v^j\nabla_{X_i}X_j + \sum_j \frac{dw^j}{dt}X_j + \sum_{i,j}\frac{dc^i}{dt}w^j\nabla_{X_i}X_j \\ &= \frac{DV}{dt} + \frac{DW}{dt}.\end{aligned}$$

(b): suppose$f$ a differentiable function on I

$$\begin{aligned}\frac{D(fV)}{dt} &= \sum_j \frac{d(fv^j)}{dt}X_j + \sum_{i,j}\frac{dc^i}{dt}fv^j\nabla_{X_i}X_j \\ &= \sum_j \frac{df}{dt}v^jX_j + \sum_j f\frac{dv^j}{dt}X_j + \sum_{i,j}f\frac{dc^i}{dt}v^j\nabla_{X_i}X_j \\ &= \frac{df}{dt}V + f\frac{DV}{dt}.\end{aligned}$$

(c): suppose V is induced by a vector field Z along c, then

$$\begin{aligned}\nabla_{dc/dt}Z &= \sum_i \frac{dc^i(t)}{dt}\nabla_{X_i}Z = \sum_{i,h}\frac{dc^i(t)}{dt}\nabla_{X_i}(z^h(c(t))X_h) \\ &= \sum_{i,h}\frac{dc^i(t)}{dt}\big(X_i(z^h(c(t)))X_h + z^h(c(t))\nabla_{X_i}X_h\big) \\ &= \sum_h \frac{dc}{dt}(z^h(c(t)))X_h + \sum_{i,h}\frac{dc^i(t)}{dt}z^h(c(t))\nabla_{X_i}X_h \\ &= \frac{D(Z(c(t)))}{dt} = \frac{DV}{dt}\end{aligned}$$

If $\psi(W)$ is another coordinate neighborhood , with $\psi(W)\cap\varphi(U)\neq\emptyset$ and we define by $\frac{dV}{dt}$ in $\psi(W)$ by (2.1), the definition agree in $\psi(W)\cap\varphi(U)$, by the uniqueness of $\frac{dV}{dt}$ in $\varphi(U)$. It follows that the definition can be extended over all of M, and this concludes the proof. □

Remark 2.1. Choosing a system of coordinates $(x^1,\dots,x^n)$, where $n=\dim M$, about p and writing

$$Y=\sum_j y^jX_j \qquad Z=\sum_j z^jX_j$$

where $X_j=\frac{\partial}{\partial x^j}$, we have

$$\nabla_Y Z = \nabla_{\sum_j y^jX_j}\sum_i z^iX_i = \sum_{j,i}y^j\nabla_{X_j}z^iX_i = \sum_{j,i}y^j(X_j(z^i)+z^i\nabla_{X_j}X_i).$$

Setting $\nabla_{X_j} X_i = \sum_k \Gamma^k_{j,i} X_k$, we conclude that the $\Gamma^k_{j,i}$ are differentiable functions and that

$$\nabla_Y Z = \sum_k (Y(z^k) + \sum_{i,j} y^i z^j \Gamma^k_{i,j}) X_k.$$

From now on denote $g(v,w)$ with $\langle v, w\rangle$ and $g_{ij} = \langle X_i, X_j\rangle$.

Definition 2.3 (*Connection compatible with the metric*)**.** A connection ∇ on a Riemannian manifold M is said to be compatible with the metric if and only if

$$X\langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle, \qquad X, Y, Z \in \chi(M).$$

Definition 2.4 (*Symmetric connection*)**.** A connection ∇ on a smooth manifold M is said to be symmetric if

$$\nabla_X Y - \nabla_Y X = [X, Y] \quad \text{for all} \quad X, Y \in \chi(M).$$

Theorem 2.1.2 (Levi-Civita)**.** *Given a Riemannian manifold M, there exists a unique connection ∇ on M satisfying the conditions:*

a) ∇ is symmetric;

b) ∇ is compatible with the Riemannian metric.

The connection given by the theorem will be referred as the *Levi-Civita connection* or *Riemannian connection* on M.

Proof. Suppose initially the existence of such a ∇. Then

$$X\langle Y, Z\rangle = \langle \nabla_X Y, Z\rangle + \langle Y, \nabla_X Z\rangle, \tag{2.2}$$

$$Y\langle Z, X\rangle = \langle \nabla_Y Z, X\rangle + \langle Z, \nabla_Y X\rangle, \tag{2.3}$$

$$Z\langle X, Y\rangle = \langle \nabla_Z X, Y\rangle + \langle X, \nabla_Z Y\rangle. \tag{2.4}$$

Adding (2.2) a (2.3) and subtracting (2.4), using the symmetry of ∇ we have that

$$X\langle Y, Z\rangle + Y\langle Z, X\rangle - Z\langle X, Y\rangle = \langle [X, Z], Y\rangle + \langle [Y, Z], X\rangle + \langle [X, Y], Z\rangle + 2\langle Z, \nabla_Y X\rangle.$$

Therefore, we find the *Koszul formula*

$$\langle Z, \nabla_Y X\rangle = \frac{1}{2}\{X\langle Y, Z\rangle + Y\langle Z, X\rangle - Z\langle X, Y\rangle - \langle [X, Z], Y\rangle - \langle [Y, Z], X\rangle - \langle [X, Y], Z\rangle\}. \tag{2.5}$$

The expression (2.5) shows that ∇ is uniquely determined from the metric $\langle\ ,\ \rangle$ since Z is an arbitrary vector field. Hence, if it exists, it will be unique.

To prove the existence, define ∇ by (2.5). It is easy to verify that ∇ is well defined and satisfies the desired conditions:

a):

$$\begin{aligned} g(\nabla_Y X - \nabla_X Y, Z) &= g(\nabla_Y X, Z) - g(\nabla_X Y, Z) \\ &= -\frac{1}{2} g([X,Y],Z) + \frac{1}{2} g([Y,X],Z) \\ &= -g([X,Y],Z) = g(-[X,Y],Z). \end{aligned}$$

b):

$$\begin{aligned} g(\nabla_X Y, Z) + g(Y, \nabla_X Z) =& X g(Y,Z) + \frac{1}{2}\{g([X,Y],Z) + g([Z,Y],X) + g([Z,X],Y) \\ & - g([Z,Y],X) - g([Z,Y,X]) - g([X,Y],Z)\} \\ =& X g(Y,Z). \end{aligned}$$

□

Remark 2.2. Suppose that we have a coordinate system (U, φ) on M. It is customary to call the functions $\Gamma^k_{i,j}$ defined on U by $\nabla_{X_i} X_j = \sum_k \Gamma^k_{ij} X_k$ the *Christoffel symbols* of the connection. It is possible to recover these symbols from the metric g. In fact, if we consider $X_i = \frac{\partial}{\partial x^i}, X_j = \frac{\partial}{\partial x^j}, X_k = \frac{\partial}{\partial x^k}$, it follows from the proof of the Theorem 2.1.2 that

$$\begin{aligned} \frac{\partial}{\partial x^i} g_{jk} + \frac{\partial}{\partial x^j} g_{ik} - \frac{\partial}{\partial x^k} g_{ij} &= X_i g(X_j, X_k) + X_j g(X_k, X_i) - X_k g(X_i, X_j) = \\ &= g([X_i, X_k], X_j) + g([X_j, X_k], X_i) + g([X_i, X_j], X_k) + 2g(X_k, \nabla_{X_j} X_i) \\ &= 2g(X_k, \nabla_{X_j} X_i), \end{aligned}$$

where we have used the fact that $[X_i, X_j] = 0$ for all $i \neq j$. So we have,

$$\sum_l \Gamma^l_{ij} g_{lk} = \frac{1}{2}\Big\{ \frac{\partial}{\partial x^i} g_{jk} + \frac{\partial}{\partial x^j} g_{ik} - \frac{\partial}{\partial x^k} g_{ij} \Big\}.$$

Since the matrix (g_{km}) admits an inverse (g^{km}), recalling that $\sum_i g_{ik} g^{ji} = \delta_{jk}$, we obtain that

$$\sum_l \Gamma^l_{ij} \underbrace{\sum_k g_{lk} g^{mk}}_{=\delta_{l,m}} = \frac{1}{2} \sum_k \Big\{ \frac{\partial}{\partial x^i} g_{jk} + \frac{\partial}{\partial x^j} g_{ik} - \frac{\partial}{\partial x^k} g_{ij} \Big\} g^{mk},$$

so

$$\Gamma^m_{ij} = \frac{1}{2} \sum_k \Big\{ \frac{\partial}{\partial x^i} g_{jk} + \frac{\partial}{\partial x^j} g_{ik} - \frac{\partial}{\partial x^k} g_{ij} \Big\} g^{mk}. \tag{2.6}$$

The equation (2.6) is a classical expression for the Christoffel symbols of the Riemannian connection in terms of g_{ij} , i.e. given by the metric.

Example 2.2. For the Euclidean space $\mathbb{R}^n$ we have $\Gamma_{ij}^k = 0$. In terms of the Christoffel symbols, the covariant derivative has the classical expression

$$\frac{DV}{dt} = \sum_k \Big\{ \frac{dv^k}{dt} + \sum_{i,j} \Gamma_{ij}^k v^j \frac{dX_i}{dt} \Big\} X_k.$$

Observe that $\frac{DV}{dt}$ differs from the usual derivative in the Euclidean space by terms which involve the Christoffel symbols. Therefore, in Euclidean spaces the covariant derivative coincides with the usual derivative.

2.2 Connection of a Riemannian submanifold

Now, we would like to investigate the case of the Levi Civita connection for Riemannian submanifold.

Definition 2.5 (*Riemannian submanifold*)**.** Let $(\tilde{M}, \tilde{g})$ be a Riemannian manifold. (M, g) is a Riemannian submanifold of $(\tilde{M}, \tilde{g})$ if:

1) M is a submanifold of $\tilde{M}$;

2) for any $p \in M$, h_p is the restriction of g_p to T_pM.

So, let (M, g) be a Riemannian submanifold of $(\tilde{M}, \tilde{g})$ and assume that M is n-dimensional, and is of codimension d in $\tilde{M}$.

Remark 2.3. Recall that if $f\colon M \longrightarrow \tilde{M}$ is an immersion, then we can split $T_p\tilde{M}$ into the direct sum

$$T_p\tilde{M} = T_pM \oplus (T_pM)^\perp,$$

where $(T_pM)^\perp$ is the orthogonal complement of T_pM in $T_p\tilde{M}$. Clearly, this decomposition is valid because T_pM is naturally identified with a subspace of $T_{f(p)}\tilde{M}$ via the map $df\colon T_pM \longrightarrow T_{f(p)}\tilde{M}$. Here df is injective on T_pM because f is required to be an immersion. In this set, if v is an element of $T_p\tilde{M}$ with $p \in M$, we can write

$$v = v^T + v^N \text{ with } v^T \in T_pM \text{ and } v^N \in (T_pM)^\perp.$$

We call v^T the *tangential component* of v and v^N the *normal component* of v.

Denote with ∇ and $\tilde{\nabla}$ the Riemannian connection of M and $\tilde{M}$ respectively and consider $\tilde{X}$ and $\tilde{Y}$ local extension of the field X and Y on M. We would like to show that

$$\nabla_X Y = (\tilde{\nabla}_{\tilde{X}} \tilde{Y})^T$$

and this is the Riemannian connection relative to the metric induced on M by $\tilde{M}$.

In particular, let X be a vector field on M, and $p \in M$. Then there exists an open subset U of $\tilde{M}$ containing p, and a vector field $\tilde{X}$ defined on U, which restriction to $M \cap U$ is X. Let then (U, ϕ) be a chart for $\tilde{M}$ around p such that $\phi(U \cap M) = \mathbb{R}^n \times \{0\} \subset \mathbb{R}^{n+d}$. In this chart, let $X = \sum_i a^i \frac{\partial}{\partial x^i}$, where a^i is a smooth function on $U \cap M$. Just set

$$\tilde{X}\big(\phi^{-1}(x,y)\big) = \sum_i a^i(\phi^{-1}(x,0))\frac{\partial}{\partial x^i}.$$

Proposition 2.2.1. *In the previously notation we have*

$$(\nabla_X Y)_p = (\tilde{\nabla}_{\tilde{X}}\tilde{Y})^T. \tag{2.7}$$

Proof. Let $\tilde{X}, \tilde{Y}$ and $\tilde{Z}$ be local extensions for the vector field X, Y and Z on M. Then $\tilde{X} = \sum_{i=1}^{n+d} \tilde{X}^i \frac{\partial}{\partial d_i}$, with $\tilde{X}^i_{|U\cap M} = X^i$ for $i \leq n, \tilde{X}^i_{|U\cap M} = 0$ for $i > n$, and likewise for Y and Z. This shows that $\big([\tilde{X}, \tilde{Y}]\big)_p = \big([X, Y]\big)_p$, and $\tilde{g}(\tilde{X}, \tilde{Y}) = g(X, Y)$ for all $p \in U \cap M$. Just apply to both connections ∇ and $\tilde{\nabla}$ to find that

$$\tilde{g}(\tilde{\nabla}_{\tilde{X}}\tilde{Y}, \tilde{Z}) = g(\nabla_X Y, Z)$$

on $U \cap M$, and so we can conclude that

$$(\nabla_X Y)_p = (\tilde{\nabla}_{\tilde{X}}\tilde{Y})^T$$

because we are dealing with elements of T_pM since we are on $U \cap M$. □

Example 2.3. Let us analyse the case of Riemannian submanifolds of Euclidean space. If M is a Riemannian submanifold of $\mathbb{R}^n$, then formula (2.7) implies that a smooth curve γ in M is a geodesic of M if and only if its second derivative γ'' in $\mathbb{R}^n$ is everywhere normal to M:

$$\gamma''(t) = \underbrace{\gamma''(t)^T}_{\in T_{\gamma(t)}M} + \underbrace{\gamma''(t)^N}_{\in (T_{\gamma(t)}M)^\perp}$$

but,

$$0 = \nabla_{\dot{\gamma}(t)}\dot{\gamma}(t) = \gamma''(t)^T$$

and so we can conclude that the geodesics of M are the curves with normal acceleration.

2.3 Geodesics and Exponential map

In what follows, M will be a Riemannian manifold, together with its Levi-Civita connection and $I \subset \mathbb{R}$.

Definition 2.6 (*Geodesic*)**.** A parametrized curve $\gamma : I \longrightarrow M$ is a geodesic at $t_0 \in I$ if $\frac{D}{dt}(\frac{d\gamma}{dt}) = 0$ at the point t_0. If γ is a geodesic at t for all $t \in I$, we say that γ is a geodesic.

At times, by abuse of language, we refer to the image $\gamma(I)$, of a geodesic γ, as a geodesic.

We can determine the local equations satisfied by a geodesic γ in a system of coordinates (U, x) about $\gamma(t_0)$. In U a curve γ

$$\gamma(t) = (x^1(t), \ldots, x^n(t))$$

will be a geodesic if and only if

$$0 = \frac{D}{dt}(\frac{d\gamma}{dt}) = \sum_k \Big(\frac{d^2x^k}{dt^2} + \sum_{i,j} \Gamma^k_{ij} \frac{dx^i}{dt}\frac{dx^j}{dt} \Big) \frac{\partial}{\partial x^k}.$$

Hence the second order system

$$\frac{d^2x^k}{dt^2} + \sum_{i,j} \Gamma^k_{ij} \frac{dx^i}{dt}\frac{dx^j}{dt} = 0 \qquad k = 1, \ldots, n, \tag{2.8}$$

yields the desired equations. In order to study the system (2.8) it is convenient to consider the tangent bundle.

Definition 2.7 (*Vector bundle*)**.** A smooth n-dimensional vector bundle is defined by a triple (E, M, π) where E and M are a pair of smooth manifolds, E the *total space*, M the *base*, and π is a surjective map, $\pi : E \longrightarrow M$ the *projection*, satisfying the following conditions:

a) Each set $E_p = \pi^{-1}(p)$ called the *fiber* of E over p is endowed with the structure of vector space.

b) For each $p \in M$ there exists a neighborhood U of p and a diffeomorphism $\phi \colon \pi^{-1}(U) \longrightarrow U \times \mathbb{R}^n$ called a *local trivialization* of E such that the following diagram commutes:

$$\begin{array}{ccc} \pi^{-1}(U) & \xrightarrow{\ \phi\ } & U \times \mathbb{R}^n \\ {\scriptstyle \pi}\big\downarrow & & \big\downarrow{\scriptstyle \pi_1} \\ U & = & U \end{array}$$

(where π_1 is the projection onto the first factor.)

c) The restriction of ϕ to each fiber $\phi : E_p \longrightarrow \{p\} \times \mathbb{R}^n$ is a linear isomorphism.

Example 2.4 (Tangent bundle)**.** Let us define the tangent bundle defining the base as M, the total space as the set

$$TM \coloneqq \bigsqcup_{p \in M} T_pM = \cup_{p \in M} \{p\} \times T_pM = \{(p, v);\ p \in M\ ,\ v \in T_pM\}$$

and the projection $\pi : TM \longrightarrow M$ such that $\pi(v_p) = p$. The triple (TM, M, π) is a vector bundle, the *tangent bundle*. In fact, the first property follows from the definition of the

tangent space. Then, if we consider a chart (U, x) of M around $p \in M$, for all $p \in U$ we define

$$\begin{aligned} \phi_U \ &: \pi^{-1}(U) \longrightarrow U \times \mathbb{R}^n \\ &v_p \mapsto (x^1, \dots, x^n, v^1, \dots, v^n), \end{aligned}$$

where $v^1, \dots, v^n$ are the coordinates of v_p with respect to the basis $\{\frac{\partial}{\partial x^1}, \dots, \frac{\partial}{\partial x^n}\}$. This map clearly satisfies the property b) and c).

So, using the tangent bundle we can say that if γ is a geodesic then the curve

$$t \mapsto (x^1(t), \dots, x^n(t), \frac{dx^1(t)}{dt}, \dots, \frac{dx^n(t)}{dt})$$

satisfies the system

$$\begin{cases} \frac{dx^k}{dt} = y^k \\ \frac{dy^k}{dt} = -\sum_{i,j} \Gamma^k_{ij} y^i y^j \quad k = 1, \dots, n \end{cases} \tag{2.9}$$

in terms of coordinates $(x^1, \dots, x^n, y^1, \dots, y^n)$ on TU. Therefore, the second order system (2.8) on U is equivalent to the first order system (2.9) on TU.
Let us recall the following result from differential equation:

Theorem 2.3.1. *If X is a C^∞ field on the open set V in the manifold M and $p \in V$ then there exists an open set $V_0 \subset V$, a number $\delta > 0$ and a C^∞ mapping $\varphi : (-\delta, \delta) \times V_0 \longrightarrow V$ such that the curve $t \mapsto \varphi(t, p), t \in (-\delta, \delta)$ is the unique trajectory of X which at the instant $t = 0$ passes through the point p for every $p \in V_0$.*

Regarding the notation, the mapping $\varphi_t : V_0 \longrightarrow V$ given by $\varphi_t(p) = \varphi(t, p)$ is called the *flow* of X on V; the smooth map $p \mapsto X(p) \in T_pM$ is called *vector field* in TM.

Lemma 2.3.2. *There exists a unique vector field G in TM whose trajectories are of the form $t \mapsto (\gamma(t), \gamma'(t))$ where γ is a geodesic on M.*

Proof. We shall first prove the uniqueness of G, supposing its existence. Consider a system of coordinates (U, ψ) on M. From the hypothesis, the trajectories of G on TU are given by $t \mapsto (\gamma(t), \gamma'(t))$ where γ is a geodesic. It follows that $t \mapsto (\gamma(t), \gamma'(t))$ is a solution of the system of differential equation (2.9). From the uniqueness of the trajectories of such a system, we conclude that if G exists, then it is unique.

To prove the existence of G, define it locally by the system (2.9). Using the uniqueness, we conclude that G is well defined on TM. □

Definition 2.8 (*Geodesic field and flow*)**.** The vector field G defined above is called the geodesic field on TM and its flow is called geodesic flow on TM.

Applying Theorem 2.3.1 at the point $(p, 0) \in TM$ we obtain the following fact.

Proposition 2.3.3. *For each $p \in M$ there exist an open set $\mathcal{U}$ in TU where (U, x) is a system of coordinates at p and $(p, 0) \in \mathcal{U}$, a number $\delta > 0$ and a C^∞ mapping $\varphi : (-\delta, \delta) \times \mathcal{U} \longrightarrow TU$ such that $t \mapsto \varphi(t, p, v)$ is the unique trajectory of G which satisfies the initial condition $\varphi(0, p, v) = (p, v)$ for each $(p, v) \in \mathcal{U}$.*

It is possible to choose $\mathcal{U}$ in the form

$$\mathcal{U} = \{(p, v) \in TU;\ p \in V \text{ and } v \in T_pM \text{ with } |v| < \varepsilon_1\},$$

where $V \subset U$ is a neighborhood of $p \in M$. We are saying that defining the geodesic as the composition of π and φ, namely $\gamma = \pi \circ \varphi$, while $|v| < \varepsilon_1$, the geodesic $\gamma(t, p, v)$ exists in an interval $(-\delta, \delta)$ and is unique. Actually it is possible to increase the velocity of a geodesic by decreasing its interval of definition or vice-versa. This follows from the following Lemma of homogeneity.

Lemma 2.3.4 (Homogeneity of a geodesic)**.** *If the geodesic $\gamma(t, p, v)$ is defined on the interval $(-\delta, \delta)$, then the geodesic $\gamma(t, p, av), a \in \mathbb{R}, a > 0$, is defined on the interval $(-\frac{\delta}{a}, \frac{\delta}{a})$ and*

$$\gamma(t, p, av) = \gamma(at, p, v).$$

Proof. Let $h\colon (-\frac{\delta}{a}, \frac{\delta}{a}) \longrightarrow M$ be a curve given by $h(t) = \gamma(at, p, v)$. Then $h(0) = p$ and since $h'(t) = a\gamma'(at, p, v)$, we have $\frac{dh(0)}{dt} = av$. In addition, considering the covariant derivative

$$\frac{D}{dt}\Big(\frac{dh}{dt}\Big) = \nabla_{h'(t)}h'(t) = a^2\nabla_{\gamma'(at,p,v)}\gamma'(at, p, v) = 0,$$

where for the first inequality, we extend $h'(t)$ to a neighborhood of $h(t)$ in M. Therefore, h is a geodesic passing through p with velocity av at the instant $t = 0$. By uniqueness

$$\gamma(at, p, v) = h(t) = \gamma(t, p, av).$$

□

Proposition 2.3.5. *Given $p \in M$ there exist a neighborhood V of $p \in M$, a number $\varepsilon > 0$ and a C^∞ mapping $\gamma : (-2, 2) \times \mathcal{U} \longrightarrow M$, $\mathcal{U} = \{(p, w) \in TM; p \in V, w \in T_pM, |w| < \varepsilon\}$ such that $t \mapsto \gamma(t, p, w), t \in (-2, 2)$ is the unique geodesic of M which, at the instant $t = 0$ passes through p with velocity w, for every p in V and for every $w \in T_pM$, with $|w| < \varepsilon$.*

Proof. The geodesic $\gamma(t, p, av)$ of Proposition 2.3.3 is defined for $|t| < \delta$ and for $|v| < \varepsilon_1$. From the Lemma of homogeneity, $\gamma(t, q, \frac{\delta v}{2})$ is defined for $|t| < 2$. Taking $\varepsilon < \frac{\delta\varepsilon_1}{2}$, we obtain that the geodesic $\gamma(t, q, w)$ is defined for $|t| < 2$ and $|w| < \varepsilon$. □

We saw that any initial point $p \in M$ and any initial velocity $v \in T_pM$ determine a unique geodesic, this implicitly defines a map from the tangent bundle to the set of geodesic in M. More importantly, it allow us to define a map from the tangent bundle to M itself, by sending the vector v to the point obtained following the geodesic for time $t = 1$.

Definition 2.9 (*Exponential map*)**.** Let $p \in M$ and let $\mathcal{U} \subset TM$ be the open set previously defined. Then, the map $\exp : \mathcal{U} \longrightarrow M$ given by

$$\exp(p, v) = \gamma(1, p, v) = \gamma(|v|, p, \frac{v}{|v|}) \qquad (p, v) \in \mathcal{U},$$

is called the exponential map on $\mathcal{U}$.

It is clear that exp is differentiable, in fact it is the evaluation of the smooth map $\gamma(t, p, v)$ at $t = 1$, and hence it is smooth. Frequently, we shall utilize the restriction of exp to an open subset of the tangent space T_pM, that is, we define

$$\exp_p : B_\varepsilon(0) \subset T_pM \longrightarrow M$$

by $\exp_p(v) = \exp(p, v)$.

Proposition 2.3.6. *Given $p \in M$, there exists an $\varepsilon > 0$ such that $\exp_p : B_\varepsilon(0) \subset T_pM \longrightarrow M$ is a diffeomorphism of $B_\varepsilon(0)$onto an open subset of M.*

Proof. Let us calculate $d(\exp_p)_0$:

$$\begin{aligned} d(\exp_p)_0(v) =& \frac{d}{dt}(\exp_p(tv))\Big|_{t=0} = \frac{d}{dt}(\gamma(1, p, tv))\Big|_{t=0} \\ =& \frac{d}{dt}(\gamma(t, p, v))\Big|_{t=0} = v. \end{aligned}$$

Hence, $d(\exp_p)_0$ is the identity of T_pM, and it follows from the inverse function theorem that $\exp_p$ is a local diffeomorphism on a neighborhood of 0. □

Definition 2.10 (*Normal neighborhood*)**.** Any open neighborhood U of $p \in M$ that is the diffeomorphic image under $\exp_p$ of the ball $B_\varepsilon(0) \subset T_pM$ is called a normal neighborhood of p.

Remark 2.4. A choice of orthonormal basis $\{e_i\}$ for T_pM is equivalent to an isomorphism $E : \mathbb{R}^n \longrightarrow T_pM$ by $E(x^1, \ldots, x^n) = \sum_i x^i e_i$. If U is a normal neighborhood of p, we can combine the isomorphism with the exponential map to get a coordinate chart

$$\varphi = E^{-1} \circ \exp_p^{-1} : U \longrightarrow \mathbb{R}^n.$$

Such coordinates are called *normal coordinates* centered at p.

2.4 Laplace operator on a Riemannian manifold

At this point we can develop analysis on manifold with the metric and we would like to define the Laplace operator on function on M. We begin to define the Hilbert space of real valued functions on M.

2.4.1 L^2 space of functions

From this point on, we assume that the given manifold M is oriented and connected. We are looking for an n-form $\alpha(x)$ such that $\langle f, g\rangle = \langle f, g\rangle_M = \int_M f(x)g(x)\alpha(x)$ defines a positive definite inner product; such an α is callled *volume form.*

First of all, we would like to understand what the volume of a Riemannian manifold should be. For the sake of simplicity, we will just compute the volume of a coordinate chart and then, using a partition of unity we can recover the full volume of M. So, consider a positively oriented coordinate neghborhood U around p with coordinates $(x^1, \dots, x^n)$. Let $v_1, \dots, v_n$ be a positively oriented basis of $T_{p_j}M$. Then $\frac{\partial}{\partial_{x^i}} = \sum_k a_i^k v_k$.
With all this notation we define the volume of U to be

$$\mathrm{vol}(U) = \int_U \det(a_i^k) dx^1 \wedge \ldots \wedge dx^n.$$

Since $g_{ij} = \langle \frac{\partial}{\partial_{x^i}}, \frac{\partial}{\partial_{x^j}} \rangle$ we have

$$\begin{aligned} g_{ij} =& \langle \frac{\partial}{\partial_{x^i}}, \frac{\partial}{\partial_{x^j}} \rangle = \langle \sum_k a_i^k v_k, \sum_l a_j^l v_l \rangle = \sum_{k,l} a_i^k a_j^l \delta_{kl} \\ =& \sum_k a_i^k a_j^k = (AA^t)_{ij} \end{aligned}$$

where $A = (a_i^j)$. Thus

$$\det g = \det(AA^t) = (\det A)^2,$$

and so $\mathrm{vol}(U)$ should be $\int_U \sqrt{\det g}\ dx^1 \wedge \ldots \wedge dx^n$.

Definition 2.11 (*Volume form*)**.** We define the volume form of a Riemannian metric to be the top dimensional form, i.e. a differential form of top degree, which in local coordinates is given by

$$\mathrm{dvol} = \sqrt{\det g} dx^1 \wedge \ldots \wedge dx^n,$$

whenever $(\frac{\partial}{\partial_{x^1}}, \dots, \frac{\partial}{\partial_{x^n}})$ is a positively oriented basis of the tangent space.

Remark 2.5. The volume form dvol is well defined on M. In fact, for the Riemannian manifold (M, g) consider an atlas $\{(U_\alpha, \phi_\alpha)\}_{\alpha \in I}$ and suppose that exist $\alpha, \beta \in I$ such that $p \in U_\alpha \cap U_\beta$ with $p \in M$. If we suppose the coordinates in U_α are $(x^1, \dots, x^n)$ and in U_β are $(y^1, \dots, y^n)$, we would like to show that

$$\mathrm{dvol}\,|_{U_\alpha}(p) = \sqrt{\det g} dx^1 \wedge \ldots \wedge dx^n = \sqrt{\det g} dy^1 \wedge \ldots \wedge dy^n = \mathrm{dvol}\,|_{U_\beta}(p).$$

We know that

$$dx^1 \wedge \ldots \wedge dx^n = \det J \left(\frac{\partial x^i}{\partial y^j} \right) dy^1 \wedge \ldots \wedge dy^n.$$

Denoting with $g_{ij}(x) = g(\frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j})$ and $g_{ij}(y) = g(\frac{\partial}{\partial y^i}, \frac{\partial}{\partial y^j})$ we have

$$g(x) = J^{-T} g(y) J^{-1}$$

and so,

$$\det g(x) = \det J^{-2} \det g(y)$$

using the Binet formula and the properties of the determinant. We can then conclude that

$$\sqrt{\det g(x)} dx^1 \wedge \ldots \wedge dx^n = \sqrt{\det g(y) \det J^{-2}} \det J dy^1 \wedge \ldots dy^n = \sqrt{\det g(y)} dy^1 \wedge \ldots \wedge dy^n.$$

We have already said that our goal is to study the Laplace operator associated to a Riemannian manifold. The natural setting to do this is the Hilbert space of quadratic integral functions.

Definition 2.12 ($L^2(M, g)$)**.** Let (M, g) be an oriented Riemannian manifold and consider the space $C_c^\infty(M)$ of smooth functions with compact support with the global inner product defined by

$$\langle f, g \rangle_M = \int_M f(x) g(x) \mathrm{dvol}(x). \tag{2.10}$$

We define the Hilbert space $L^2(M, g)$ to be the completion of $C_c^\infty(M)$ with respect to the inner product (2.10).

2.4.2 The Laplacian on functions

The Laplace operator on functions, sometimes called Laplace-Beltrami operator, represents the first important tool to study a Riemannian manifold from an analytic point of view. The definition one can give of Laplace operator is analogue to (minus) the Laplacian $-\left(\frac{\partial^2}{(\partial x^1)^2} + \ldots + \frac{\partial^2}{(\partial x^n)^2} \right)$ in $\mathbb{R}^n$. Recall that in $\mathbb{R}^n$, for a smooth function $f \colon \mathbb{R}^n \to \mathbb{R}$, we have

$$-\sum_{i=1}^n \frac{\partial^2 f}{(\partial x^i)^2} = -\mathrm{div}(\mathrm{grad}\, f).$$

In particular we see that the second definition is somehow intrinsic, so it can be given in more general setting.
First of all we define the gradient of a smooth function i.e.

$$\mathrm{grad} \colon C^\infty(M) \to TM,$$

where (M, g) is a Riemannian manifold. In Euclidean setting $\mathrm{grad}\, f = \sum_i \frac{\partial f}{\partial x^i} \frac{\partial}{\partial x^i}$ and so it is a vector field on $\mathbb{R}^n$, while $df = \sum_i \frac{\partial f}{\partial x^i} dx^i$ is a 1-form on $\mathbb{R}^n$ such that

$$df(X) = \langle \mathrm{grad}\, f, X \rangle \,, \ \forall X \in \chi(\mathbb{R}^n)$$

where $\langle \cdot, \cdot \rangle$ here denotes the Euclidean metric. Now this kind of relation between vector fields and 1-forms can be recovered in a Riemannian manifold since we have a metric. For a Riemannian manifold (M, g), define, $\forall x \in M$

$$\begin{aligned} \alpha_{g,p} \colon T_pM &\to T_p^*M \\ v &\mapsto v^*;\ v^*(w) = \langle v, w \rangle_p . \end{aligned}$$

So $\alpha_{g,p}$ is just the canonical isomorphism between a finite dimensional vector space with a metric and its dual. We denote this isomorphism again with α. Moreover the Riemannian metric g induces an inner product on each cotangent space T_p^*M under the isomorphism α, namely for $v^*, w^* \in T_p^*M$

$$\langle v^*, w^* \rangle_p = \langle v, w \rangle_p .$$

We want to compute the expression in local coordinates of the metric on the cotangent. First suppose that $v = \frac{\partial}{\partial x^i}$ then

$$v^*(\frac{\partial}{\partial x^j}) = \left\langle \frac{\partial}{\partial x^i}, \frac{\partial}{\partial x^j} \right\rangle = g_{ij}.$$

So $v^* = (\frac{\partial}{\partial x^i})^* = \sum_j g_{ij} dx^j$, hence $(\sum_i g^{ij} \frac{\partial}{\partial x^i})^* = dx^j$. Therefore

$$\begin{aligned} g(dx^i, dx^j) &= \sum_{k,l} g(g^{ik} \frac{\partial}{\partial x^k}, g^{jl} \frac{\partial}{\partial x^l}) = \sum_{k,l} g^{ik} g^{jl} g(\frac{\partial}{\partial x^k}, \frac{\partial}{\partial x^l}) \\ &= \sum_{k,l} g^{ik} g^{jl} g_{kl} = \sum_k g^{ik} \delta_{jk} \\ &= g^{ij}. \end{aligned}$$

Therefore we can say that in local coordinates, the metric of the cotangent is represented by the inverse matrix of g.
Now we are ready to define the gradient of a smooth function on an arbitrary Riemannian manifold.

Definition 2.13 (*Gradient*)**.** Let (M, g) be a Riemannian manifold. We set the gradient to be the composition

$$\operatorname{grad} \colon C^\infty(M) \xrightarrow{d} \Lambda^1 T^*M \xrightarrow[\cong]{\alpha^{-1}} \chi(M)$$

It is easy to check that this produces the ordinary gradient in Euclidean space, in fact in local coordinates, for $f \in C^\infty(M)$ we have

$$\begin{aligned} \operatorname{grad} f = \alpha^{-1}(df) &= \alpha^{-1}(\sum_i \frac{\partial f}{\partial x^i} dx^i) = \sum_i \frac{\partial f}{\partial x^i} \alpha^{-1}(dx^i) \\ &= \sum_{i,j} \frac{\partial f}{\partial x^i} g^{ij} \frac{\partial}{\partial x^j}. \end{aligned}$$

Note that the gradient of f is a vector field on M, so the divergence shall be a functional defined on $\chi(M)$. In order to give this definition, we need to introduce the concept of contraction of a form along a vector field.

Definition 2.14 (*Contraction of ω along X*)**.** Let (M, g) be a Riemannian manifold and X a vector field on M, ω a k-form. The contraction of ω along X is defined as the $k-1$ form $i_X\omega$ that satisfies the following request: suppose $V_1, \dots, V_{k-1}$ elements of the tangent space, then

$$i_X\omega = \omega(X, V_1, \dots, V_{k-1}).$$

Example 2.5. Let us calculate the contraction of $\omega = \text{dvol}$ along $X = \frac{\partial}{\partial x^i}$. We have

$$i_X\omega = (-1)^{i-1}\sqrt{\det g}dx^1 \wedge \ldots \wedge d\hat{x}^i \wedge \ldots dx^n,$$

where with $d\hat{x}^i$ we indicate that dx^i does not appear in the wedge product.

It is easy to check that the contraction is linear with respect to X and ω, in fact:

1) suppose ω, η to be k-form, f, g differental functions on M

$$i_X(f\omega + g\eta) = (f\omega + g\eta)(X, \ldots) = f\omega(X, \ldots) + g\eta(X, \ldots)$$

2) suppose X, Y vector field on M, a, b differential function and ω a k-form

$$i_{aX+bY}\omega = \omega(aX + bY, \ldots) = a\omega(X, \ldots) + b\omega(Y, \ldots)$$

thanks to the multilinearity of the form.

From now on, suppose $\omega = \text{dvol} = \sqrt{\det g}dx^1 \wedge \ldots \wedge dx^n$.

Definition 2.15 (*Divergence*)**.** Let (M, g) be a Riemannian manifold and X a vector field on M. The divergence of X is defined as the quantity that satisfies the following identity

$$\text{div}X \cdot \omega = d(i_X\omega).$$

Example 2.6. Let us calculate the divergence of $X = \frac{\partial}{\partial x^i}$:

$$\begin{aligned}
\text{div}(X) \cdot \omega = d(i_X\omega) &= (-1)^{i-1}\frac{\partial\sqrt{\det g}}{\partial x^i}dx^i \wedge dx^1 \wedge \ldots \wedge d\hat{x}^i \wedge \ldots \wedge dx^n \\
&= \frac{\partial\sqrt{\det g}}{\partial x^i}dx^1 \wedge \ldots \wedge dx^n \\
&= -\frac{1}{\sqrt{\det g}}\frac{\partial \det g}{\partial x^i}dx^1 \wedge \ldots \wedge dx^n \\
&= -\frac{1}{\sqrt{\det g}}\det g\,\text{tr}(g^{-1}\frac{\partial}{\partial x^i}g)dx^1 \wedge \ldots \wedge dx^n \\
&= -\,\text{tr}(g^{-1}\frac{\partial}{\partial x^i}g)\omega,
\end{aligned}$$

using Jacobi's formula for the derivative of the determinant. We can conclude that

$$\operatorname{div}(\frac{\partial}{\partial x^i}) = -\operatorname{tr}(g^{-1}\frac{\partial}{\partial x^i}g).$$

Clearly, the divergence is linear thanks to the linearity of the differential and the contraction.

Proposition 2.4.1. *Let (M, g) be a Riemannian manifold, let a be a function of $C^\infty(M)$ and X a vector field on M, then we have*

$$\operatorname{div}(aX) = a\operatorname{div}(X) + da(X). \tag{2.11}$$

Proof. Let us calculate

$$\operatorname{div}(aX)\cdot\omega = d(i_{aX}\omega) = d(ai_X\omega) = da\wedge i_X\omega + ad(i_X\omega),$$

where

$$\begin{aligned} da\wedge i_X\omega &= \sum_{i,j=1}^{n}\sqrt{\det g}(-1)^{i-1}\frac{\partial a}{\partial x^j}dx^j dx^i(X)dx^1\wedge\ldots\wedge d\hat{x}^i\wedge\ldots dx^n \\ &= \sum_{i=1}^{n}\frac{\partial a}{\partial x^i}dx^i(X)\sqrt{\det g}dx^1\wedge\ldots\wedge dx^n \\ &= da(X)\cdot\omega. \end{aligned}$$

So we conclude that

$$\operatorname{div}(aX)\cdot\omega = da(X)\cdot\omega + a\operatorname{div}(X)\cdot\omega = (da(X)+\operatorname{div}(X))\omega.$$

□

Example 2.7. If we suppose $X = \sum_i b_i\frac{\partial}{\partial x^i}$, we have

$$\begin{aligned} \operatorname{div}(X) &= \sum_i \operatorname{div}\left(b_i\frac{\partial}{\partial x^i}\right) \\ &= \sum_i b_i\operatorname{div}\left(\frac{\partial}{\partial x^i}\right) + \sum_i db_i(\frac{\partial}{\partial x^i}) \\ &= \sum_i b_i(-tr(g^{-1}\frac{\partial}{\partial x^i}g) + \sum_i\sum_k\frac{\partial b_i}{\partial x^k}dx^k(\frac{\partial}{\partial x^k}) \\ &= \sum_i b_i(-tr(g^{-1}\frac{\partial}{\partial x^i}g) + \sum_i\frac{\partial b_i}{\partial x^i}. \end{aligned}$$

Now we are ready to give the definition of the Laplacian.

Definition 2.16 (*Laplacian*)**.** Let (M, g) be a Riemannian manifold and let $f \in C^\infty(M)$. The Laplacian of f, denoted by Δf, is defined as minus the divergence of grad f, i.e.

$$\Delta f = -\mathrm{div}(\mathrm{grad}\, f).$$

In local coordinates, if f is in $C^\infty(M)$, we know that $\mathrm{grad}(f) = \sum_i \left(\sum_j \frac{\partial f}{\partial x^j} g^{ij} \right) \frac{\partial}{\partial x^i}$, so

$$\begin{aligned}
\mathrm{div}(\mathrm{grad}\, f) &= \sum_i \left(\sum_j \frac{\partial f}{\partial x^j} g^{ij} \right) (-tr(g^{-1} \frac{\partial}{\partial x^i} g) + \sum_i \frac{\partial}{\partial x^i} \left(\sum_j \frac{\partial f}{\partial x^j} g^{ij} \right) \\
&= -\sum_i \sum_j \frac{\partial f}{\partial x^j} g^{ij} tr(g^{-1} \frac{\partial}{\partial x^i} g + \sum_{i,j} \frac{\partial^2 f}{\partial x^i \partial x^j} g^{ij} + \sum_{ij} \frac{\partial f}{\partial x^j} \frac{\partial g^{ij}}{\partial x^i} \\
&= -\frac{1}{2} \frac{1}{\det g} \sum_{i,j} \frac{\partial f}{\partial x^j} g^{ij} \frac{\partial \det g}{\partial x^i} + \sum_{i,j} \frac{\partial^2 f}{\partial x^i \partial x^j} g^{ij} + \sum_{ij} \frac{\partial f}{\partial x^j} \frac{\partial g^{ij}}{\partial x^i}.
\end{aligned}$$

Since

$$\begin{aligned}
&\frac{1}{\sqrt{\det g}} \sum_j \frac{\partial}{\partial x^j} \left(\sqrt{\det g} \sum_k g^{jk} \frac{\partial f}{\partial x^k} \right) = \\
&\frac{1}{\sqrt{\det g}} \sum_j \frac{\partial \sqrt{\det g}}{\partial x^j} \sum_k g^{jk} \frac{\partial f}{\partial x^k} + \frac{1}{\sqrt{\det g}} \sqrt{\det g} \sum_{j,k} \frac{\partial g^{jk}}{\partial x^j} \frac{\partial f}{\partial x^k} + \frac{1}{\sqrt{\det g}} \sqrt{\det g} \sum_{j,k} g^{jk} \frac{\partial^2 f}{\partial x^k \partial x^j} = \\
&\frac{1}{\sqrt{\det g}} \sum_j \frac{-1}{2} \frac{1}{\sqrt{\det g}} \frac{\partial \det g}{\partial x^j} \sum_k g^{jk} \frac{\partial f}{\partial x^k} + \sum_{j,k} \frac{\partial g^{jk}}{\partial x^j} \frac{\partial f}{\partial x^k} + \sum_{j,k} g^{jk} \frac{\partial^2 f}{\partial x^k \partial x^j} = \\
&-\frac{1}{2} \frac{1}{\det g} \sum_{i,j} \frac{\partial f}{\partial x^j} g^{ij} \frac{\partial \det g}{\partial x^i} + \sum_{i,j} \frac{\partial^2 f}{\partial x^i \partial x^j} g^{ij} + \sum_{ij} \frac{\partial f}{\partial x^j} \frac{\partial g^{ij}}{\partial x^i} = \mathrm{div}(\mathrm{grad}\, f).
\end{aligned}$$

As a consequence, in local coordinates, for $f \in C^\infty(M)$

$$\begin{aligned}
\Delta f &= -\frac{1}{\sqrt{\det g}} \sum_j \frac{\partial}{\partial x^j} \left(\sqrt{\det g} \sum_k g^{jk} \frac{\partial f}{\partial x^k} \right) \\
&= -\frac{1}{\sqrt{\det g}} \frac{\partial}{\partial x^j} \left(\sqrt{\det g}\; g^{ij} \frac{\partial f}{\partial x^i} \right)
\end{aligned}$$

where in the last equality we used the Einstein convention for the sum.

Remark 2.6. We can give another definition of the Laplacian on functions, studying the first variation of the functional

$$A(f) = \int |\mathrm{grad}\, f|^2 \,\mathrm{dvol}\,.$$

Suppose $h \in C_0^\infty$, let us compute the derivative of the functional in the direction h:

$$\begin{aligned}
\frac{d}{dt}A(f+th)_{|t=0} =& 2\int \operatorname{grad} f \cdot \operatorname{grad} h \,\mathrm{dvol} \\
=& 2\int \langle g^{-1}\nabla f, g^{-1}\nabla h\rangle_g \,\mathrm{dvol} \\
=& 2\int g^{-1}gg^{-1}\langle \nabla f, \nabla h\rangle \,\mathrm{dvol} \\
=& 2\int \sqrt{\det g}g^{-1}\langle \nabla f, \nabla h\rangle dx \\
=& 2\int \sqrt{\det g}g^{ij}\frac{\partial}{\partial x^j}f\frac{\partial}{\partial x^i}h dx \\
=& 2\int \frac{1}{\sqrt{\det g}}\frac{\partial}{\partial x^i}(\sqrt{\det g}g^{ij}\frac{\partial}{\partial x^j}f)h \,\mathrm{dvol},
\end{aligned}$$

where we used the integration by parts formula and $dx = dx^1 \wedge \ldots \wedge dx^n$. So, we can conclude that the first variation of the gradient is

$$\int -\Delta f h \,\mathrm{dvol};$$

in this way we have given an analytical definition of the Laplacian on functions, a valid alternative to the geometric one using forms.

2.5 Hodge theory for functions

In this section we would like to investigate the Hodge theorem for compact, connected, oriented Riemannian manifold, which states that there is an orthonormal basis of $L^2(M,g)$ of eigenfunctions of the Laplacian; hence if we restrict our attention to the space spanned by the first N eigenvectors, we can write

$$\Delta \sim \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_N \end{pmatrix}$$

where λ_i's are the eigenvalues.
A complete presentation of the proof of the theorem can be found in [13]. Here we will point out some remarkable definitions and properties that allow to demonstrate the result.

2.5.1 Heat operator

The proof of the theorem exploit the heat operator, so we begin with its definition.

Definition 2.17 (Heat kernel)**.** Let (M, g) be an oriented, compact, connected Riemannian manifold. Suppose there exists a function $e(t, x, y) \in C^\infty(\mathbb{R}^+ \times M \times M)$ such that

$$\begin{cases} (\partial_t + \Delta_y)e(t, x, y) = 0 \\ \lim_{t\to 0^+} \int_M e(t, x, y)f(y)\,\mathrm{dvol}(y) = f(x),\ \forall f \in L^2(M, g). \end{cases}$$

Here Δ_y denotes the Laplacian acting in the y variable. We say that $e(t, x, y)$ is the heat kernel on M.

Even if it's not clear from the definition, if the heat kernel exists, it is symmetric and unique. Hence we can rewrite the first condition as

$$(\partial_t + \Delta_x)e(t, x, y) = 0.$$

So we can interpret the heat kernel as the kernel of the resolvent operator for the heat equation associated with the Laplacian: $\partial_t + \Delta_x = 0$. Indeed the resolvent operator, also called heat operator, is given by:

$$e^{-t\Delta} f(x) = \int_M e(t, x, y)f(y)\,\mathrm{dvol}(y),\ \forall x \in M.$$

This means that, given a temperature distribution on our manifold, say $f \in L^2(M, g)$, then setting $f(t, x) \coloneqq e^{-t\Delta} f(x)$, we have

$$\begin{cases} (\partial_t + \Delta_x)f(t, x) = 0 \\ f(0, x) = f(x), \quad \forall x \in M \end{cases}$$

where $f(0, x) = \lim_{t\to 0} f(t, x)$.

The definition we gave corresponds to a general setting, now we specialize it in the case of the Laplacian and the operator defined below is actually a heat operator.

Definition 2.18 (Heat Operator)**.** Let (M, g) be a compact, oriented, connected Riemannian manifold and let $e(t, x, y)$ be its heat kernel. We define the heat operator as

$$\begin{aligned} e^{-t\Delta} \colon L^2(M, g) &\to L^2(M, g) \\ f &\mapsto \int_M e(t, \cdot, y)f(y)\,\mathrm{dvol}(y). \end{aligned}$$

We may regard the heat kernel as a compact operator, thanks to compactness theorem in Sobolev spaces

$$L^2(M, g) \xrightarrow{e^{-t\Delta}} L^2(M, g).$$

Moreover, the heat operator is self-adjoint:

$$\begin{aligned}\left\langle e^{-t\Delta}f, g\right\rangle &= \int_M \left(\int_M e(t,x,y)f(y)dy\right) g(x)dx \\ &= \int_M \left(\int_M e(t,y,x)g(x)dx\right) f(y)dy \\ &= \left\langle f, e^{-t\Delta}g\right\rangle,\end{aligned}$$

by the symmetry of the heat kernel $e(t,x,y)$. By the spectral theorem for self-adjoint compact operators on Hilbert spaces, there exists an orthonormal basis of $L^2(M,g)$ consisting of eigenfunctions for the heat operator $e^{-t\Delta}$ with egenvalues $\gamma_i(t) \xrightarrow{i\to\infty} 0$. We can prove that the $\gamma_i(t)$'s are strictly positive, i.e. 0 is not an eigenvalue of $e^{-t\Delta}$ thanks to the semigroup property of the heat operator

Lemma 2.5.1. $e^{-t\Delta}e^{-s\Delta} = e^{-(t+s)\Delta}$.

It is possible even to show that

Lemma 2.5.2. *There exist $\lambda_i \in \mathbb{R}$ such that, for all t*

$$\gamma_i(t) = e^{-\lambda_i t}.$$

2.5.2 Hodge theorem

Theorem 2.5.3 (Hodge Theorem for Functions)**.** *Let (M,g) be a compact, connected, oriented Riemannian manifold. There exists an orthonormal basis of $L^2(M,g)$ consisting of eigenfunctions of the Laplacian. All the eigenvalues are positive with a finite multiplicity, except that zero is an eigenvalue with multiplicity one. Moreover the eigenvalues accumulate only at infinity.*

A straightforward consequence of the Hodge theorem for functions is that the formal sum

$$\sum_i e^{-\lambda_i t}\phi_i(x)\phi_i(y),$$

converges point-wise to the heat kernel $e(t,x,y)$.

Proposition 2.5.4. *Let (M,g) be a compact, connected, oriented Riemannian manifold and let $e(t,x,y) \in C^\infty(\mathbb{R}^+ \times M \times M)$ be the heat kernel for functions. Then we have the point-wise convergence*

$$e(t,x,y) = \sum_i e^{-\lambda_i t}\phi_i(x)\phi_i(y).$$

Corollary 2.5.5. *The series $tr(e^{-t\Delta}) \coloneqq \sum_i e^{-\lambda_i t}$ converges for each $t > 0$ and sum is $\int_M e(t,x,x)dx$.*

Chapter 3

Diffusion maps

In this chapter we extend the notion of heat kernel, initially given in the Riemannian setting, to a general diffusion process for geometric description of data sets, called diffusion maps.

3.1 Definition of the diffusion process

As we saw in the previous chapter, a diffusion process in a Riemannian setting can be realized via a geometric heat kernel, which contains the properties of the space. In analogy, starting with a geometric kernel, we will show how to define an operator satisfying the Markov property, whose eigendecomposition produces an embedding of the data in $\mathbb{R}^n$ via *diffusion maps*. In this space, the Euclidean distance defines a diffusion metric that measure the proximity of points in the given set of data.

3.1.1 Construction of a diffusion operator

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, where Ω is a set whose points are abstract objects and μ can be thought as a counting measure or a probability one, with the request that $\mu(\Omega) < \infty$.
Consider then a function $k : \Omega \times \Omega \longrightarrow \mathbb{R}$ which assigns a real number to a given object pair that satisfies the following *admissibility conditions*:

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

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

The number $k(x, y)$ is usually interpreted as a measure of similarity between the points x and y and it represents our a priori information on Ω.

Remark 3.1. The admissibility conditions allow the function k to define a notion of neighborhood, i.e. the neighborhood of x corresponds to all points y such that $k(x, y)$ is numerically significant. In other words, if we suppose $\varepsilon > 0$ we can define:

$$\mathcal{U}_{x,\varepsilon} = \{y \in \Omega \ ; \ 0 < k(x, y) < \varepsilon\}.$$

Furthermore, the positive preservation property will allow to renormalize k in order to define a diffusion process on the data and the third condition is necessary for imposing the positivity of the diffusion metric.

We would like to construct a diffusion process on the data. This result can be achieved by interpreting the function k as a kernel; we need to renormalized it, in order to interpret it as a probabilistic density. Denote

$$d^2(x) = \int_\Omega k(x, y) d\mu(y),$$

we notice that this quantity is well defined because $k(x, y) \geq 0$. Then $\tilde{a}(x, y) = \frac{k(x,y)}{d^2(x)}$ satisfies

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

moreover $\tilde{a}(x, y) \geq 0$, for all x and $y \in \Omega$.

Definition 3.1. *(Kernel operator)* The diffusion operator $\tilde{A}$ corresponding to the kernel $\tilde{a}(x, y)$ is defined naturally as:

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

From an analysis perspective, this operator can be viewed as an averaging operator as it fixes constant functions and it is also positive-preserving: if $f \geq 0$ then $Af \geq 0$. We can easily check that the function $\tilde{a}$ satisfies the positive-preserving property and it is positive semi-definite, but it does not posses the symmetric feature.
So we can renormalized $\tilde{a}$:

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

in this way we have $a(x, y) = a(y, x)$. Then, we can defined naturally the kernel operator associated:

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

Remark 3.2. The normalization $a(x, y) = \frac{k(x,y)}{d(x)d(y)}$ is sometimes called *graph Laplacian normalization* because of a connection with the spectral theory of a graph [17]. More details about this association can be found in Chapter 4.

The new kernel $a(x,y)$ is therefore conjugate to the stochastic one $\tilde{a}(x,y)$, and share the same spectrum and its eigenfunctions are obtained by conjugation by d. In what follows we will use A rather than $\tilde{A}$ as a diffusion operator.

Theorem 3.1.1. *The diffusion operator A with kernel a*

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

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

$$\|A\| = 1$$

and it is achieved by the eigenfunction $d(x)$:

$$Ad(x) = d(x).$$

Moreover, A is symmetric and positive semi-definite.

Proof. Let us begin the proof by showing the boundness of the operator, hence we need to determine $M > 0$ such that $\|Af\|_{L^2(\Omega,\mu)} \leq M\|f\|_{L^2(\Omega,\mu)}$. First, we notice that:

$$\begin{aligned}
\int_\Omega a(x,y) f(y) d\mu(y) &= \int_\Omega \frac{k(x,y)}{d(x)d(y)} f(y) d\mu(y) \\
&= \frac{1}{d(x)} \int_\Omega k(x,y) \frac{f(y)}{d(y)} d\mu(y) \\
&\leq \frac{1}{d(x)} \left(\int_\Omega k(x,y) d\mu(y) \right)^{1/2} \left(\int_\Omega k(x,y) \frac{|f(y)|^2}{d(y)^2} d\mu(y) \right)^{1/2} \\
&= \left(\int_\Omega k(x,y) \frac{|f(y)|^2}{d(y)^2} d\mu(y) \right)^{1/2},
\end{aligned}$$

using the Schwartz inequality with $f(x,y) = k(x,y)^{1/2}, g(x,y) = k(x,y)^{1/2} \frac{f(y)}{d(y)}$. It follows that

$$\begin{aligned}
\|Af\|_{L^2}^2 = \langle Af, Af \rangle_{L^2} &= \int_\Omega |Af(x)|^2 d\mu(x) = \\
&= \int_\Omega |\int_\Omega a(x,y) f(y) d\mu(y)|^2 d\mu(x) \leq \int_\Omega \int_\Omega k(x,y) \frac{|f(y)|^2}{d(y)^2} d\mu(y) d\mu(x) \\
&= \int_\Omega \frac{|f(y)|^2}{d(y)^2} \int_\Omega k(x,y) d\mu(x) d\mu(y) = \int_\Omega |f(y)|^2 d\mu(y) = \|f\|_{L^2}^2.
\end{aligned}$$

We can affirm that $M = 1$ and A is bounded from L^2 into itself.
Then we have, by definition,

$$\|A\|_{op} = \sup_{\|f\|_{L^2}=1} \|Af\|_{L^2}.$$

Thanks to the previous point we can say that

$$\|A\|_{op} \leq 1$$

and, since $d(x)$ is an eigenfunction of A corresponding to eigenvalue 1:

$$\begin{aligned} Ad(x) &= \int_\Omega a(x,y)d(y)d\mu(y) \\ &= \int_\Omega \frac{k(x,y)}{d(x)}d\mu(y) = d(x), \end{aligned}$$

we can then conclude that $1 = \|A\|$.
The operator A is symmetric, in fact for all $f, g \in L^2(\Omega, \mu)$ we have

$$\langle Af, g\rangle = \int_{\Omega\times\Omega} a(x,y)f(y)d\mu(y)g(x)d\mu(x) = \int_{\Omega\times\Omega} a(x,y)f(y)g(x)d\mu(x)d\mu(y),$$

$$\langle f, Ag\rangle = \int_{\Omega\times\Omega} f(x)d\mu(x)a(x,y)g(y)d\mu(y) = \int_{\Omega\times\Omega} a(x,y)f(x)g(y)d\mu(x)d\mu(y)$$

Hence, $\langle Af, g\rangle = \langle f, Ag\rangle$ from which it follows that the operator is not only symmetric, but also self-adjoint.
Lastly, we show that A is positive semi-definite: for all $f \in L^2(\Omega, \mu)$ we have

$$\langle Af, f\rangle = \int_\Omega\int_\Omega a(x,y)f(x)f(y)d\mu(x)d\mu(y) = \int_\Omega\int_\Omega k(x,y)\frac{f(x)}{d(x)}\frac{f(y)}{d(y)}d\mu(x)d\mu(y) \geq 0$$

thanks to the positivity of k. □

In addition, if we assume that

$$\int_{\Omega\times\Omega} a(x,y)^2 d\mu(x)d\mu(y) < \infty, \tag{3.1}$$

we can show that A is a compact operator. In fact, we can apply the following proposition, whose a complete proof can be found in [7, pp 43].

Proposition 3.1.2. *If $(\Omega, \mathcal{F}, \mu)$ is a measure space and $a \in L^2(\Omega\times\Omega, \mathcal{F}\times\mathcal{F}, \mu\times\mu)$, then*

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

is a compact operator and $\|A\| \leq \|a\|_{L^2}$.

Remark 3.3. The request on the kernel of equation (3.1) is compatible from a probabilistic point of view with the equation (1.5).

3.1.2 Spectral decomposition of the diffusion kernel

Previously, we proved that the diffusion operator is compact and self-adjoint. We can then apply the spectral decomposition theorem for an operator on an Hilbert space and find an orthonormal basis of eigenfunctions for $L^2(\Omega, \mu)$ with respect to the operator A. We denote with $\{\lambda_i\}_{i\in\mathbb{N}}$ the eigenvalues of the operator A and $\{\phi_i\}_{i\in\mathbb{N}}$ the corresponding eigenfunctions.

Lemma 3.1.3. *Let $\{\lambda_n\}_{n\in\mathbb{N}}$ be the eigenvalues sequence for a compact self-adjoint operator A. If the sequence is infinite, then $\lambda_n \to 0$ for $n \to \infty$.*

Proof. Suppose that the sequence $\{\lambda_n\}_{n\in\mathbb{N}}$ does not converge to 0, for $n \to \infty$. So, there exists $\varepsilon > 0$ such that $|\lambda_n| \geq \varepsilon$ for all $n \geq \bar{n}$. If $m \neq n$ *with* $n, m \geq \bar{n}$, denoting with $\{e_n\}_{n\in\mathbb{N}}$ the set of the corresponding eigenfunctions, we have

$$\|Ae_n - Ae_m\|^2 = \|\lambda_n e_n - \lambda_m e_m\|^2 = \lambda_n^2 + \lambda_m^2 > \varepsilon^2,$$

and this shows that $(Ae_n)_{n\in\mathbb{N}}$ has no convergent subsequences, a contradiction to the compactness of A. □

Lemma 3.1.4. *If A is a compact self-adjoint operator, then either $\pm\|A\|$ is an eigenvalue of A.*

Proof. Denote with $r(A) = \max\{|\lambda_n| \ ; \ \lambda_n \text{ eigenvalue of } A\}$ and recall the spectral radius formula: $r(A) = \lim_{n\to\infty} \|A^n\|^{1/n}$. Then, also

$$r(A) = \lim_{n\to\infty} \|A^n\|^{1/n} = \lim_{n\to\infty} \|A^{2^n}\|^{1/2^n} = \lim_{n\to\infty} \|A\|^{\frac{2^n}{2^n}} = \|A\|.$$

The only thing we need to prove to conclude is that $\|A\|^{2^n} = \|A^{2^n}\|$. First, suppose $n = 1$:

$$\|Ax\|^2 = \langle Ax, Ax\rangle = \langle A^*Ax, x\rangle = \langle A^2x, x\rangle \leq \|A^2x\|\|x\|,$$

taking the supremum over all x of norm 1 we obtain

$$\|A\|^2 \leq \|A^2\|.$$

On the other hand,

$$\|A^2x\| \leq \|A\|\|Ax\| \leq \|A\|\|A\|\|x\| = \|A\|^2\|x\|,$$

taking the supremum

$$\|A^2\| \leq \|A\|^2.$$

Let be $n > 1$ and prove the statement by induction: first,

$$\|A^{2^n}x\| \leq \|A^{2^{n-1}}\|\|A^2x\| \leq \|A\|^{2^{n-1}}\|A\|\|x\| = \|A\|^{2^n}\|x\|,$$

hece,

$$\|A^{2^n}\| \leq \|A\|^{2^n}.$$

Finally, notice that

$$\|A\|^{2^n} = \sup_{\|x\|=1} \|Ax\|^{2^n} = \sup_{\|x\|=1} (\|Ax\|^2)^{2^{n-1}};$$

if we consider the quantity

$$(\|Ax\|^2)^{2^{n-1}} = (\langle Ax, Ax\rangle)^{2^{n-1}} \leq (\|A^2x\|\|x\|)^{2^{n-1}},$$

taking the supremum we have

$$\|A\|^{2^n} \leq (\|A^2\|)^{2^{n-1}} = \|A^{2\cdot 2^{n-1}}\| = \|A^{2^n}\|.$$

□

Remark 3.4. It follows from these Lemmas that the eigenvalues of the diffusion operator are distributed between 0 and 1. In other words we have

$$1 = \lambda_0 \geq \lambda_1 \geq \dots$$

with $\lambda_i \geq 0$, for all $i \in \mathbb{N}$.

As we previously stated, we would like to find a spectral decomposition of the kernel. In order to do so we need a special basis for $L^2(\Omega \times \Omega)$.

Proposition 3.1.5. *Let be $\{\phi_n\}_{n\in\mathbb{N}}$ an orthonormal basis for $L^2(\Omega)$, then $\{\phi_n\phi_m\}_{n,m\in\mathbb{N}}$ is an orthonormal basis for $L^2(\Omega \times \Omega)$.*

Proof. First thing to do is to check that $\{\phi_n\phi_m\}_{n,m\in\mathbb{N}}$ is an orthonormal set for $L^2(\Omega\times\Omega)$:

$$\begin{aligned}\langle \phi_n(x)\phi_m(y), \phi_i(x)\phi_j(y)\rangle &= \int_{\Omega\times\Omega} \phi_n(x)\phi_m(y)\phi_i(x)\phi_j(y)d\mu(x)d\mu(y) \\ &= \delta_{n,i}\delta_{m,j} = \begin{cases} 1 & \text{if } n = i \ , \ m = j \\ 0 & \text{otherwise.}\end{cases}\end{aligned}$$

Then, we can see if the orthogonal of the space generated by $\{\phi_n\phi_m\}$ is empty. In other words, suppose $h \in L^2(\Omega \times \Omega)$ such that $\langle h, \phi_n\phi_m\rangle = 0$, we would like to show that h is the null function. Consider

$$0 = \int_{\Omega\times\Omega} h(x,y)\phi_n(x)\phi_m(y)d\mu(x)d\mu(y) = \int_\Omega \left(\int_\Omega h(x,y)\phi_n(y)d\mu(y)\right)\phi_m(x)d\mu(x).$$

and note that, since $\{\phi_m\}_{m\in\mathbb{N}}$ is a basis for $L^2(\Omega)$ the function $x \to \int_\Omega h(x,y)\phi_n(y)d\mu(y)$ is zero almost everywhere for each n. Denoting with

$$E_n = \left\{x \in \Omega \ ; \ \int_\Omega h(x,y)\phi_n(y)d\mu(y) \neq 0\right\}$$

we have that $E = \cup_{n\geq 0} E_n$ is a null set. We can say that $h(x,y) = 0$ for all $x \in \Omega \setminus E$ almost everywhere.
Finally, since $h \in L^2(\Omega \times \Omega)$ we have

$$\int_{\Omega\times\Omega} |h(x,y)|^2 d\mu(x)d\mu(y) = \int_\Omega \int_\Omega |h(x,y)|^2 d\mu(x)d\mu(y) = \int_{\Omega\setminus E} \int_\Omega |h(x,y)|^2 d\mu(x)d\mu(y) = 0,$$

and this concludes the proof. □

Let us consider then the kernel $a(x,y)$, seeing this as a function of y we can apply the following decomposition:

$$a(x,y) = \sum_{i\geq 0} c_i(x)\phi_i(y). \tag{3.2}$$

We can easily check that the coefficients $c_i(x)$ are in $L^2(\Omega)$, in fact

$$\begin{aligned}
\infty > \int_{\Omega\times\Omega} |a(x,y)|^2 d\mu(x)d\mu(y) &= \int_{\Omega\times\Omega} |\sum_{i\geq 0} c_i(x)\phi_i(y)|^2 d\mu(x)d\mu(y) \\
&\geq \int_\Omega \int_\Omega \sum_{i\geq 0} |c_i(x)|^2 |\phi_i(y)|^2 d\mu(x)d\mu(y) = \sum_{i\geq 0} \underbrace{\int_\Omega |\phi_i(y)|^2 d\mu(y)}_{=1} \int_\Omega |c_i(x)|^2 d\mu(x) \\
&= \sum_{i\geq 0} \int_\Omega |c_i(x)|^2 d\mu(x) \geq \int_\Omega |c_i(x)|^2 d\mu(x),
\end{aligned}$$

and then we can write

$$c_i(x) = \sum_{k\geq 0} a_{i,k}\phi_k(x). \tag{3.3}$$

Putting (3.3) in (3.2) we obtain:

$$a(x,y) = \sum_{i\geq 0}\sum_{k\geq 0} a_{i,k}\phi_k(x)\phi_i(y). \tag{3.4}$$

Observing that :

$$\begin{aligned}
\lambda_h\phi_h(x) = A\phi_h(x) = \int_\Omega a(x,y)\phi_h(y)d\mu(y) &= \int_\Omega \sum_{i\geq 0}\sum_{k\geq 0} a_{i,k}\phi_k(x)\phi_i(y)\phi_h(y)d\mu(y) \\
= \sum_{i,k\geq 0} a_{i,k}\phi_k(x)\langle\phi_i(y),\phi_h(y)\rangle &= \sum_{i,k\geq 0} a_{i,k}\phi_k(x)\delta_{i,h} = \sum_{k\geq 0} a_{h,k}\phi_k(x),
\end{aligned}$$

it follows that $a_{h,h} = \lambda_h$ while $a_h, k = 0$ for all $k \neq h$. We can conclude:

$$a(x,y) = \sum_{j\geq 0} \lambda_j\phi_j(x)\phi_j(y). \tag{3.5}$$

An analogous decomposition holds for the kernel related to the diffusion operator A^m, with $m \geq 1$.

Proposition 3.1.6. *Let $a^{(m)}(x,y)$ denote the kernel of A^m. Then we have*

$$a^{(m)}(x,y) = \sum_{j\geq 0} \lambda_j^m \phi_j(x)\phi_j(y). \tag{3.6}$$

Proof. Let us begin with writing explicitly the operator

$$A^m f(x) = \int_{\underbrace{\Omega \times \cdots \times \Omega}_{m}} a(x,x_1)a(x_1,x_2)\cdots a(x_{m-1},x_m)f(x_m)d\mu(x_1)\cdots d\mu(x_m). \tag{3.7}$$

We have already shown that for the kernel holds the following decomposition:

$$a(x,x_1)a(x_1,x_2)\cdots a(x_{m-1},x_m) = \sum_{i\geq 0} \lambda_i^m \phi_i(x)\phi_i^2(x_1)\cdots\phi_i^2(x_{m-1})\phi_i(x_m),$$

reorganizing the equation (3.7) we obtain

$$A^m f(x) = \sum_{i\geq 0} \Bigg(\int_\Omega \phi_i^2(x_1)d\mu(x_1) \int_\Omega \phi_i^2(x_2)d\mu(x_2)\cdot \ldots \cdot \int_\Omega \phi_i^2(x_{m-1})d\mu(x_{m-1})\cdot \\ \cdot \int_\Omega \lambda_i^m \phi_i(x)\phi_i(x_m)f(x_m)d\mu(x_m)\Bigg).$$

Since $\{\phi_n\}_{n\in\mathbb{N}}$ is an orthonormal basis for $L^2(\Omega)$, we have $\int_\Omega \phi_i^2(x_k)d\mu(x_k) = 1$, for all $i \geq 0$ and $k = 1,\ldots m-1$. It follows that

$$A^m f(x) = \int_\Omega \sum_{i\geq 0} \lambda_i^m \phi_i(x)\phi_i(x_m)f(x_m)d\mu(x_m), \tag{3.8}$$

and defining $a^m(x,y) = \sum_{i\geq 0} \lambda_i^m \phi_i(x)\phi_i(y)$ we can conclude. □

Remark 3.5. Using the definition and the properties of diffusion operators $A = (A^m)_{m\geq 0}$, defining $A^0 = \mathrm{Id}$, we may notice that there are some analogies with the Markov semigroup, even if, here, there is a discretization of the continuous parameter of the definition. In fact, each A^m with $m \geq 0$ satisfies the properties of a Markov operator as in Definition 1.5 thanks to Theorem 3.1.1. Another straightforward consequence of Theorem 3.1.1 is that the first property of Definition 1.6 is verified. Furthermore, the semigroup property holds: on one hand we have

$$A^{m+n} f(x) = \int_\Omega a^{m+n}(x,y)f(y)d\mu(y) = \int_\Omega \sum_i \lambda_i^{m+n}\phi_i(x)\phi_i(y)f(y)d\mu(y),$$

on the other hand

$$\begin{aligned}
A^m(A^n f(x)) =& A^m\left(\int_\Omega a^n(x,y)f(y)d\mu(y)\right) = A^m\Bigg(\underbrace{\int_\Omega \sum_i \lambda_i^n \phi_i(x)\phi_i(y)f(y)d\mu(y)}_{F(x)}\Bigg)\\
&= \int_\Omega \sum_i \lambda_i^m \phi_i(x)\phi_i(z)F(z)d\mu(z)\\
&= \int_\Omega \sum_i \lambda_i^m \phi_i(x)\phi_i(z) \int_\Omega \sum_i \lambda_i^n \phi_i(z)\phi_i(y)f(y)d\mu(y)d\mu(z)\\
&= \int_\Omega \sum_i \lambda_i^{m+n} \phi_i(x)\phi_i(z)^2\phi_i(y)f(y)d\mu(y)d\mu(z)\\
&= \underbrace{\int_\Omega \phi_i(z)^2 d\mu(z)}_{=1} \int_\Omega \sum_i \lambda_i^{m+n} \phi_i(x)\phi_i(y)f(y)d\mu(y)\\
&= \int_\Omega \sum_i \lambda_i^{m+n} \phi_i(x)\phi_i(y)f(y)d\mu(y).
\end{aligned}$$

Finally, it is verified also the continuity property: for every $f \in L^2(\Omega)$

$$\|A^m f - f\|_2^2 = \int_\Omega \left(A^m f(x) - f(x)\right)^2 d\mu(x) = \int_\Omega \left(\int_\Omega a^m(x,y)f(y)d\mu(y) - f(x)\right)^2 d\mu(x),$$

sending m to 0 we have

$$\int_\Omega \Bigg(\underbrace{\int_\Omega f(y)d\mu(y)}_{=C} - f(x)\Bigg)^2 d\mu(x) = \int_\Omega C^2 + f(x)^2 - 2Cf(x)d\mu(x) \leq C^2 + C^2 - 2C^2 = 0,$$

if we suppose that μ is a probability measure.

3.1.3 Diffusion maps and metrics

The main goal of this subsection is to define a map that transports the data into a particular l^2 space, in which the usual distance corresponds to a distance on Ω. The first thing to do is to define this distance, and we would do that using the spectral decomposition presented in the previous subsection.
Maintaining the previous notation, we introduce the following mapping:

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

in which each eigenfunction is interpreted as a coordinate on the set. This map takes abstract entities and provides a representation of the data as point in an Euclidean space.

Definition 3.2. *(Diffusion metric)* Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, A the diffusion operator associated to a kernel $a(x, y)$. We define the family of diffusion metrics $\{D_m\}_{m\geq 1}$ as

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

Proposition 3.1.7. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $\{\phi_n\}_{n\in\mathbb{N}}$ the orthonormal basis associated to the diffusion operator A. Then,*

$$D_m^2(x, y) = \sum_{i\geq 0} \lambda_i^m (\phi_i(x) - \phi_i(y))^2 = \|x - y\|^2.$$

In other words, the diffusion metric can be computed as a weighted Euclidean distance in the embedding space, the weights being $\lambda_0^m, \lambda_1^m, \ldots$

Proof. The proof is a mere computation, in fact

$$\begin{aligned} D_m^2(x, y) &= a^m(x, x) + a^m(y, y) - 2a^m(x, y) \\ &= \sum_{i\geq 0} \lambda_i^m(\phi_i^2(x)) + \sum_{i\geq 0} \lambda_i^m \phi_i^2(y) - 2\sum_{i\geq 0} \lambda_i^m \phi_i(x)\phi_i(y) = \sum_{i\geq 0} \lambda_i^m(\phi_i(x) - \phi_i(y))^2. \end{aligned}$$

□

Proposition 3.1.8. *The diffusion distance can be expressed using the L^2 norm of the diffusion kernels:*

$$D_{2m}^2(x, y) = \int_\Omega |a^m(x, z) - a^m(y, z)|^2 d\mu(z) = \|a^m(x, \cdot) - a^m(y, \cdot)\|_{L^2}^2. \tag{3.9}$$

Proof. Let us begin with the calculation of the L^2 norm of the functions:

$$\begin{aligned} \int_\Omega |a^m(x, z) - a^m(y, z)|^2 d\mu(z) &= \int_\Omega (a^m(x, z) - a^m(y, z))^2 d\mu(z) \\ &= \int_\Omega \Big(\sum_{i\geq 0} \lambda_i \phi_i(x)\phi_i(z) - \sum_{i\geq 0} \lambda_i^m \phi_i(y)\phi_i(z) \Big)^2 d\mu(z) \\ &= \Big[\sum_{i\geq 0} \lambda_i^{2m} \Big(\int_\Omega \phi_i(x)^2 \phi_i(z)^2 d\mu(z) + \int_\Omega \phi_i(y)^2 \phi_i(z)^2 d\mu(z) \\ &\quad - 2\int_\Omega \phi_i(x)\phi_i(y)\phi_i(z)^2 d\mu(z) \Big) \Big] \\ &= \sum_{i\geq 0} \lambda_i^{2m} \Big(\phi_i^2(x) + \phi_i^2(y) - 2\phi_i(x)\phi_i(y) \Big). \end{aligned}$$

On the other side, we have

$$\begin{aligned} D_{2m}^2(x, y) &= a^{2m}(x, x) + a^{2m}(y, y) - 2a^{2m}(x, y) \\ &= \sum_{i\geq 0} \lambda_i^{2m} \phi_i^2(x) + \sum_{i\geq 0} \lambda_i^{2m} \phi_i^2(y) - \sum_{i\geq 0} \lambda_i^{2m} 2\phi_i(x)\phi_i(y) \\ &= \sum_{i\geq 0} \lambda_i^{2m} \Big(\phi_i^2(x) + \phi_i^2(y) - 2\phi_i(x)\phi_i(y) \Big). \end{aligned}$$

□

Using these results, we can derive that D_m is a semi-metric metric in the classical sense. First of all, the symmetry is satisfied thanks to the symmetry of the kernel $a(x, y)$. Moreover, D_m is non negative and it verifies the triangular inequality:

$$\begin{aligned} D_{2m}^2(x,y) &= \int_\Omega |a^m(x,z) - a^m(y,z)|^2 d\mu(z) \\ &= \int_\Omega |a^m(x,z) - a^m(h,z) + a^m(h,z) - a^m(y,z)|^2 d\mu(z) \\ &\leq \int_\Omega |a^m(x,z) - a^m(h,z)|^2 d\mu(z) + \int_\Omega |a^m(y,z) - a^m(h,z)|^2 d\mu(z) \\ &= D_{2m}^2(x,h) + D_{2m}^2(h,y). \end{aligned}$$

Remark 3.6. We can give a probabilistic interpretation about the diffusion distance, since it is strictly correlated to the number of paths connecting x and y. It is small if there are many high probability paths between two points and it will be large if, on the contrary, the number of connection is small. Recall the equivalent definition

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

We can assume that the kernels have compact support, so that the integral is extended to neighborhood of points x and y. Moreover, the integral is extended on the set when μ is different from 0, which is the set where the density of points of the data set is high. Hence, the distance is small if the density of points is high in a neighborhood containing x and y. It is high if the density is small in a set containing both points. In order to give a probabilistic interpretation, we identify kernels with probabilistic density of a stochastic process. This is depicted in Figure (3.1). We can notice that B and C are well connected because we have many paths that joint them. On the contrary, A and B are very distant because they are connected by few paths. To sum up, we can say that, since we are speaking by a probabilistic point of view, if we have many paths between two points, the probability to joint them is higher respect to the probability to joint two points connected by less paths.

We can therefore introduce the family of *diffusion maps* $\{\Phi_{t,p}\}_{t,p\in\mathbb{N}}$ given by

$$\Phi_{t,p}(x) = \begin{pmatrix} \lambda_1^t \phi_1(x) \\ \lambda_2^t \phi_2(x) \\ \vdots \\ \lambda_p^t \phi_p(x) \end{pmatrix}. \tag{3.10}$$

Each component of $\Phi_{t,p}(x)$ is called *diffusion coordinate* and the map $\Phi_{t,p}\colon \Omega \longrightarrow \mathbb{R}^p$ embeds the dataset into a Euclidean space of dimension p. In other words, we have proved the following Corollary.

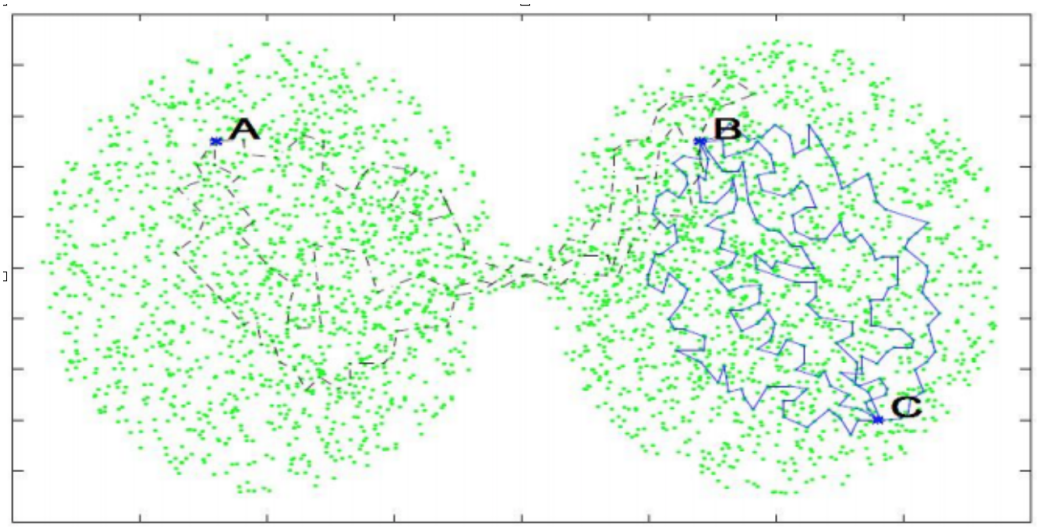

Figure 3.1: Example of diffusion distance in a set of points, created by Lafon in [15]. Points B and C are close in the sense of D_m; on the contrary because of a presence of a bottleneck, point A and B are very distant from each others.

Corollary 3.1.9. *The embedding generated by the eigenfunctions allows a dimensionality reduction of the data. Indeed, for a given accuracy δ, we retain only the eigenvalues $\lambda_1, \ldots, \lambda_p$ that, when raised to the power m, exceed the δ threshold, and we use the corresponding eigenfunctions $\phi_1, \ldots, \phi_p$ to embed the data points into $\mathbb{R}^p$.*

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

In this section we analyse the case that Ω is a subset of $\mathbb{R}^n$. In particular, we assume the data lie on a submanifold of $\mathbb{R}^n$ and we show how different operators can recover the Riemannian geometry. The main goal is to recover the manifold structure regardless the distributions of the sample points. In order to do so, we will introduce a family of diffusion processes, parameterized by a number $\alpha \in \mathbb{R}$ which can be tuned up to specify the influence of the density. Two values of α are of particular interesting:

- $\alpha = 0$, the diffusion reduces to the one defined in the previous section;
- $\alpha = 1$, if the points approximately lie on a submanifold of $\mathbb{R}^n$, we obtain an approximation of the Laplace-Beltrami operator.

As for the notation, let $\mathcal{M}$ be a C^∞ compact manifold of dimension d in $\mathbb{R}^n$ and μ be a measure on $\mathcal{M}$. The metric on $\mathcal{M}$ 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 $\mathcal{M}$, i.e.

$d\mu(x) = q(x)dx$. This density can be thought as the density of the sample points in our data set, thus it does not have to be uniform.

From now on we make the fundamental assumption that the only objects that are observable are defined in relation with the geometry of the ambient space $\mathbb{R}^n$ and the distribution $d\mu(x) = q(x)dx$ of the points. For instance, we have access to the Euclidean distance between two points, or it make sense to compute integral against $d\mu$ but we do not have the knowledge of the geodesic distances on $\mathcal{M}$. Likewise, the action of the Laplace-Beltrami operator cannot be observed as it is an object of the intrinsic geometry of $\mathcal{M}$. Our goal is to show that by using the geometry of the ambient space we can approximate objects whose definitions rely on the intrinsic geometry only.

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

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

where we assume that h is a function infinitely differentiable with an exponential decay at infinity.

3.2.1 Family of diffusions

Let us set up the family of diffusions that will allow us to recover the Riemannian structure regardless the probabilistic terms. We fix $\alpha \in \mathbb{R}$ and a rotation invariant kernel $k_\varepsilon(x, y) = h\left(\frac{\|x-y\|^2}{\varepsilon}\right)$. Let

$$q_\varepsilon(x) = \int_{\mathcal{M}} k_\varepsilon(x, y)q(y)dy$$

and form the new kernel

$$k^\alpha_\varepsilon(x, y) = \frac{k_{\varepsilon(x,y)}}{q^\alpha_\varepsilon(x)q^\alpha_\varepsilon(y)}.$$

We normalize the quantity $k^\alpha_\varepsilon(x, y)$ by setting

$$d^\alpha_\varepsilon(x) = \int_{\mathcal{M}} k^\alpha_\varepsilon(x, y)q(y)dy$$

and by defining the stochastic kernel

$$a_{\varepsilon,\alpha}(x, y) = \frac{k^\alpha_\varepsilon(x, y)}{d^\alpha_\varepsilon(x)}.$$

Let $A_{\varepsilon,\alpha}$ be defined by

$$A_{\varepsilon,\alpha}f(x) = \int_{\mathcal{M}} a_{\varepsilon,\alpha}(x, y)f(y)q(y)dy,$$

even in this case we can think about this operator as a Markov operator. In fact, $A_{\varepsilon,\alpha}1 = 1$ and if $f \geq 0$ then $A_{\varepsilon,\alpha}f \geq 0$.

Our main result will concern the operator

$$L_{\varepsilon,\alpha} \coloneqq \frac{\mathrm{Id} - A_{\varepsilon,\alpha}}{\varepsilon}$$

as $\varepsilon \mapsto 0$ and it can be interpreted as the infinitesimal generator associated to a Markov semigroup. This will be more clear thanks to the results of the next paragraphs.

3.2.2 Technical Results

In this subsection we prove an asymptotic expansion for the operator $L_{\varepsilon,\alpha}$ acting on a submanifold $\mathcal{M} \subset \mathbb{R}^n$. We start by comparing the metric defined on the manifold with that defined on the (local) projection of the submanifold over the tangent space.
We assume $\mathcal{M}$ to be C^∞ and compact. Let $x \in \mathcal{M}$ be a fixed point not on the boundary, $T_x\mathcal{M}$ be the tangent space to $\mathcal{M}$ at x and $\{e_1, \dots, e_d\}$ be a fixed orthonormal basis of $T_x\mathcal{M}$. In what follows, we introduce two systems of local coordinates in the neighborhood of x.

First, we consider normal coordinates: the exponential map $\exp_x$ generates a set of orthogonal geodesics $(\gamma_1, \dots, \gamma_d)$ intersecting at x with initial velocity $(e_1, \dots, e_d)$, and any point y in a sufficiently small neighborhood of x has a set of normal coordinates $(s_1, \dots, s_d)$ along these geodesics. Therefore, any function f defined on $\mathcal{M}$ in the neighborhood of x can be seen as a function $\tilde{f}$ of $(s_1, \dots, s_d)$; in this case if $f \in C^2$ we have

$$\Delta f(x) = -\sum_{i=1}^{d} \frac{\partial^2}{\partial s_i^2} \tilde{f}(0, \dots, 0),$$

where Δ is the Laplace Beltrami operator on $\mathcal{M}$.

Remark 3.7. If x is on the boundary $\partial\mathcal{M}$ of $\mathcal{M}$, and if we choose $e_1, \dots, e_{d-1}$ to be in the tangent space of this boundary at x, while e_d is normal and pointing in, then the normal derivative of a function f at x is defined as

$$\frac{\partial f}{\partial \nu}(x) = -\frac{\partial \tilde{f}}{\partial s_d}(0).$$

The second system of coordinates is given by the orthogonal projection u of y on $T_x\mathcal{M}$. More precisely, the coordinates $(u_1, \dots, u_d)$ are given by $u_i = \langle x - y, e_i \rangle_{\mathbb{R}^n}$. The submanifold is now locally parameterized as $y = (u, g(u))$, where $g \colon \mathbb{R}^d \longrightarrow \mathbb{R}^{n-d}$. Since $u = (u_1, \dots, u_d)$ are tangent coordinates, we must have that $\frac{\partial g}{\partial u_i}(0) = 0$.

Locally we have the following diagram :

$$(s_1, \dots, s_d) \underbrace{\longleftrightarrow}_{\exp_x} y \underbrace{\longleftrightarrow}_{\text{projection}} u.$$

In what follows we will convert all quantities depending on $(s_1, \dots, s_d)$ or y into functions of u. For the notation, $Q_{x,m}(u)$ denotes a generic homogeneous polynomial of degree

m of the variable $u = (u_1, \dots, u_d)$ whose coefficients depend on x. Since these polynomials form an equivalence class, we might abuse the notation and write, for instance $Q_{x,m}(u) + Q_{x,m}(u) = Q_{x,m}(u)$.

Lemma 3.2.1. *If $y \in \mathcal{M}$ is in an Euclidean ball of radius $\varepsilon^{1/2}$ around x, then*

$$s_i = u_i + Q_{x,3}(u) + \mathcal{O}(\varepsilon^2) \tag{3.11}$$

for ε sufficiently small.

Proof. Let γ be the geodesic connecting x and y parameterized by arclength. We have $\gamma(0) = x$ and let s be such that $\gamma(s) = y$. If y has normal coordinates $(s_1, \dots, s_d)$, then we have

$$s\gamma'(0) = (s_1, \dots, s_d) \qquad\qquad \gamma'(0) = (v_1, \dots, v_d).$$

A Taylor expansion yields

$$\gamma(s) = \gamma(0) + s\gamma'(0) + \frac{s^2}{2}\gamma''(0) + \frac{s^3}{6}\gamma^{(3)}(0) + \mathcal{O}(s^4),$$

and so

$$u_i = \langle \gamma(s) - \gamma(0), e_i \rangle = \langle s\gamma'(0), e_i \rangle + \frac{s^2}{2}\langle \gamma''(0), e_i \rangle + \frac{s^3}{6}\langle \gamma^{(3)}(0), e_i \rangle + \langle \mathcal{O}(s^4), e_i \rangle.$$

We can notice that $\langle \gamma''(0), e_i \rangle = 0$ by definition of a geodesic, recalling the Proposition 2.7. Moreover, since $\gamma'(0) = (v_1, \dots, v_d)$ and $s\gamma'(0) = (s_1, \dots, s_d) = (sv_1, \dots, sv_d)$, we have that $s = \frac{s_i}{v_i}$ for all $i = 1, \dots, d$. In particular, $s^3 = \frac{s_i^3}{v_i^3} = \frac{u_i^3}{v_i^3}$ and using the expansion we obtain

$$\frac{s^3}{6}\langle \gamma^{(3)}(0), e_i \rangle = Q_{x,3}(u).$$

We can then conclude using the fact that y is in an Euclidean ball of radius $\varepsilon^{1/2}$ around x. □

Clearly, in analogy with the previous proof, we can obtain an expansion of a high order for the formula (3.11), i.e.

$$s_i = u_i + Q_{x,3}(u) + Q_{x,4}(u) + \mathcal{O}(\varepsilon^{5/2}). \tag{3.12}$$

Lemma 3.2.2. *If $y \in \mathcal{M}$ is in an Euclidean ball of radius $\varepsilon^{1/2}$ around x, then for ε sufficiently small we have:*

$$\|x - y\|^2 = \|u\|^2 + Q_{x,4}(u) + Q_{x,5}(u) + \mathcal{O}(\varepsilon^3). \tag{3.13}$$

Proof. The manifold is locally parameterized by $u \mapsto (u, g(u))$, where $g\colon \mathbb{R}^d \longrightarrow \mathbb{R}^{n-d}$. Writing $g = (g_{i+1}, \dots, g_n)$ and applying Pythagore's theorem, we obtain

$$\|x - y\|^2 = \|u\|^2 + \sum_{i=d+1}^{n} g_i(u)^2.$$

Clearly, $g_i(0) = 0$ and $\frac{\partial g}{\partial u_i}(0) = 0$. As a consequence, using the Taylor expansion

$$g_i(u) = \underbrace{g_i(0)}_{=0} + \sum_{j=1}^{d} u_i \underbrace{\partial_{u_i} g_i(0)}_{=0} + \underbrace{\frac{1}{2}\sum_{k,j=1}^{d} u_i u_j \partial^2_{u_k,u_j} g_i(0)}_{=b_{i,x}(u)} + \underbrace{\frac{1}{6}\sum_{k,j,h=1}^{d} u_k u_j u_h \partial^3_{u_k,u_j,u_h} g_i(0)}_{=c_{i,x}(u)} + \mathcal{O}(\|u\|^4)$$
$$= b_{i,x}(u) + c_{i,x}(u) + \mathcal{O}(\|u\|^4),$$

we get that $g_i(u)^2 = b_{i,x}^2(u) + 2b_{i,x}c_{i,x} + \mathcal{O}(\|u\|^6)$. Setting $Q_{x,4}(u) = \sum_{i=1}^{d} b_{i,x}(u)^2$ and $Q_{x,5}(u) = \sum_{i=1}^{d} 2b_{i,x}c_{i,x}$ we proved the (3.13).

□

Lemma 3.2.3. *If $y \in \mathcal{M}$ is in an Euclidean ball of radius $\varepsilon^{1/2}$ around x, then, for ε sufficiently small, we have:*

$$\det\left(\frac{dy}{du}\right) = 1 + Q_{x,2}(u) + Q_{x,3}(u) + \mathcal{O}(\varepsilon^2). \tag{3.14}$$

Proof. Let us denote with $A \coloneqq \left(\frac{dy}{du}\right)$ the matrix obtained by the change of variable $(s_1, \dots, s_d) \mapsto (u_1, \dots, u_d)$. Using (3.12), we know that

$$A = \begin{pmatrix} 1 + Q_{x,2}(u_1) + Q_{x,3}(u_1) & Q_{x,2}(u_2) + Q_{x,3}(u_2) & \dots & Q_{x,2}(u_d) + Q_{x,3}(u_d) \\ Q_{x,2}(u_1) + Q_{x,3}(u_1) & 1 + Q_{x,2}(u_2) + Q_{x,3}(u_2) & \dots & Q_{x,2}(u_d) + Q_{x,3}(u_d) \\ \vdots & \vdots & \vdots & \vdots \\ Q_{x,2}(u_1) + Q_{x,3}(u_1) & Q_{x,2}(u_2) + Q_{x,3}(u_2) & \dots & 1 + Q_{x,2}(u_d) + Q_{x,3}(u_d) \end{pmatrix}.$$

If we call $H(u) = \det\Big(A(u)\Big)$, via the Taylor expansion in 0 we have

$$H(u) = H(0) + \langle \nabla H(0), u \rangle + \frac{1}{2} u^T \operatorname{Hess} H(0) u + \dots$$

and so, to conclude the proof we only need to show that $H(0) = 1$ and $\nabla H(0) = 0$. Clearly, $H(0) = 1$ thanks to the form of the matrix A. To prove that the vector

$$\nabla H(0) = \left(\frac{\partial}{\partial u_1} H_{|u=0}, \dots, \frac{\partial}{\partial u_d} H_{|u=0}\right)$$

is the null vector, let us calculate each term:

$$\frac{\partial}{\partial u_i} H_{|u=0} = \frac{\partial}{\partial u_i} \det A_{|u=0} = \det A \operatorname{tr}(A^{-1} \frac{\partial}{\partial u_i} A)_{|u=0} = 1 \cdot \operatorname{tr} \begin{pmatrix} 0 & \dots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \dots & 0 \end{pmatrix} = 0,$$

where, in the second equality, we used the Jacobi's formula for the derivative of the determinant of an invertible matrix; and in the third equality we used the fact that

$$A^{-1}\frac{\partial}{\partial u_i}A = \begin{pmatrix} 0 & \dots & Q_{x,1}(u_i)+Q_{x,2}(u_i) & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & Q_{x,1}(u_i)+Q_{x,2}(u_i) & \dots & 0 \end{pmatrix}.$$

□

Let $k_\varepsilon(x,y)$ be an isotropic kernel :

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

where h is assumed to have an exponential decay and let G_ε be the corresponding operator

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

Lemma 3.2.4. *Let $f \in C^3(\mathcal{M})$ and let $0 < \gamma < \frac{1}{2}$. Then we have, uniformly for all $x \in \mathcal{M}$ at distance larger than ε^γ from $\partial\mathcal{M}$,*

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

where

$$m_0 = \int_{\mathbb{R}^d} h(\|u\|^2)du \qquad \text{and} \qquad m_2 = \int_{\mathbb{R}^d} u_1^2 h(\|u\|^2)du$$

and ω is a term that depends on $\mathcal{M}$.

Proof. First of all, we notice that the domain of integration can be restricted to the intersection of $\mathcal{M}$ with the ball of radius ε^γ around x up to an error of order $\mathcal{O}(\varepsilon^2)$. In fact,

$$G_\varepsilon f(x) = \frac{1}{\varepsilon^{\frac{d}{2}}}\int_{y\in\mathcal{M};\|x-y\|>\varepsilon^\gamma} k_\varepsilon(x,y)f(y)dy + \frac{1}{\varepsilon^{\frac{d}{2}}}\int_{y\in\mathcal{M};\|x-y\|<\varepsilon^\gamma} k_\varepsilon(x,y)f(y)dy$$

and

$$\begin{aligned}
|\frac{1}{\varepsilon^{\frac{d}{2}}}\int_{y\in\mathcal{M};\|x-y\|>\varepsilon^\gamma} k_\varepsilon(x,y)f(y)dy| &\le \|f\|_\infty \frac{1}{\varepsilon^{\frac{d}{2}}}\int_{\mathcal{M}} |h\Big(\frac{\|x-y\|^2}{\varepsilon}\Big)|dy \\
&\le \|f\|_\infty \int_{z\in\mathcal{M};\|z\|>\varepsilon^{\gamma-1/2}} |h(\|z\|^2)|dz \\
&\le \|f\|_\infty \int_{\|z\|>\varepsilon^{\gamma-1/2}} |h(\|z\|^2)|dz \\
&\le C\|f\|_\infty Q(\varepsilon^{\gamma-\frac{1}{2}})e^{-(\varepsilon(\gamma-1/2))^2} \\
&\le C\|f\|_\infty Q(\varepsilon^{\gamma-\frac{1}{2}})e^{-\varepsilon(\gamma-1/2)},
\end{aligned}$$

where we have used the exponential decay of the kernel and consequently the polar coordinates, in order to obtain Q, a polynomial of the form $Q(x) = \frac{x^{n-2}}{2} + \frac{n-2}{4}x^{n-3} + \ldots$. Moreover, the last inequality follows from the fact that since $0 < \gamma < 1/2$, we have $e^{-\varepsilon^{(\gamma-\frac{1}{2})^2}} \leq e^{-\varepsilon^{\gamma-\frac{1}{2}}}$. Clearly, because this term is exponentially small, we can say that is bounded by $\mathcal{O}(\varepsilon^2)$. Therefore,

$$G_\varepsilon f(x) = \frac{1}{\varepsilon^{\frac{d}{2}}} \int_{y\in\mathcal{M};\|x-y\|<\varepsilon^\gamma} k_\varepsilon(x,y) f(y) dy + \mathcal{O}(\varepsilon^2).$$

Now that things are localized around x, we can Taylor expand the function $(s_1, \ldots, s_d) \mapsto f(y(s_1, \ldots, s_d))$:

$$f(y) = f(x) + \sum_{i=1}^{d} s_i \frac{\partial \tilde{f}(0)}{\partial s_i} + \frac{1}{2}\sum_{i,j=1}^{d} s_i s_j \frac{\partial^2 \tilde{f}(0)}{\partial s_i \partial s_j} + Q_{x,3}(s_1, \ldots, s_d) + \mathcal{O}(\varepsilon^2),$$

where $\tilde{f}(s_1, \ldots, s_d) = f(y(s_1, \ldots, s_d))$. Invoking (3.11), we obtain

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

Likewise, because of equation (3.13), the Taylor expansion of the kernel is

$$h\Big(\frac{\|x-y\|^2}{\varepsilon}\Big) = h\Big(\frac{\|u\|^2}{\varepsilon}\Big) + \Big(\frac{Q_{x,4}(u)}{\varepsilon} + \frac{Q_{x,5}(u)}{\varepsilon}\Big) h'\Big(\frac{\|u\|^2}{\varepsilon}\Big) + \mathcal{O}(\varepsilon^2).$$

Using equation (3.14) to change the variable $s \mapsto u$ in the previous integral defining by $G_\varepsilon f(x)$ yields:

$$\begin{aligned}
\varepsilon^{\frac{d}{2}} G_\varepsilon f(x) = \int_{\|u\|<\varepsilon^\gamma} & \left(h\Big(\frac{\|u\|^2}{\varepsilon}\Big) + \Big(\frac{Q_{x,4}(u)}{\varepsilon} + \frac{Q_{x,5}(u)}{\varepsilon}\Big) h'\Big(\frac{\|u\|^2}{\varepsilon}\Big) \right) \\
& \times \left(\tilde{f}(0) + \sum_{i=1}^{d} u_i \frac{\partial \tilde{f}(0)}{\partial s_i} + \frac{1}{2}\sum_{i,j=1}^{d} u_i u_j \frac{\partial^2 \tilde{f}(0)}{\partial s_i \partial s_j} + Q_{x,3}(u) \right) \\
& \times (1 + Q_{x,2}(u) + Q_{x,3}(u)) du + \mathcal{O}(\varepsilon^{\frac{d}{2}+2}).
\end{aligned}$$

This identity can be dramatically simplified by identifying odd functions and setting their integral to zero. One is left with

$$\begin{aligned}
\varepsilon^{\frac{d}{2}} G_\varepsilon f(x) = \int_{\mathbb{R}^d} & \tilde{f}(0) h\Big(\frac{\|u\|^2}{\varepsilon}\Big) + \frac{1}{2}\Big(\sum_{i,j=1}^{d} \frac{\partial^2 \tilde{f}(0)}{\partial s_i \partial s_j}\Big) u_i^2 h\Big(\frac{\|u\|^2}{\varepsilon}\Big) \\
& + \tilde{f}(0)\frac{Q_{x,4}(u)}{\varepsilon} h'\Big(\frac{\|u\|^2}{\varepsilon}\Big) + \tilde{f}(0) Q_{x,2}(u) h\Big(\frac{\|u\|^2}{\varepsilon}\Big) \\
& + Q_{x,2}(u) h\Big(\frac{\|u\|^2}{\varepsilon}\Big) \frac{1}{2}\Big(\sum_{i,j=1}^{d} \frac{\partial^2 \tilde{f}(0)}{\partial s_i \partial s_j}\Big) + \tilde{f}(0)\frac{Q_{x,4}(u)}{\varepsilon} Q_{x,2}(u) \\
& + \frac{Q_{x,4}(u)}{\varepsilon} Q_{x,2}(u) \frac{1}{2}\Big(\sum_{i,j=1}^{d} \frac{\partial^2 \tilde{f}(0)}{\partial s_i \partial s_j}\Big) du + \mathcal{O}(\varepsilon^{\frac{d}{2}+2}),
\end{aligned}$$

where the domain of integration has been extended to $\mathbb{R}^d$ thanks to the exponentially decay of h. Changing the variable according to $u \mapsto \sqrt{\varepsilon}u$,

$$\begin{aligned} G_\varepsilon f(x) =& \tilde{f}(0) \int_{\mathbb{R}^d} h(\|u\|^2)du + \frac{\varepsilon}{2}\Big(\sum_{i,j=1}^{d} \frac{\partial^2 \tilde{f}(0)}{\partial s_i \partial s_j}\Big) \int_{\mathbb{R}^d} u_1^2 h(\|u\|^2)du \\ &+ \varepsilon \tilde{f}(0) \int_{\mathbb{R}^d} (Q_{x,4}(u)h'(\|u\|^2) + Q_{x,2}(u)h(\|u\|^2))du + \mathcal{O}(\varepsilon^2), \end{aligned}$$

where we have used the homogeneity of $Q_{x,4}$ and $Q_{x,2}$. Finally, observing that

$$\tilde{f}(0) = f(x) \qquad \text{and} \qquad \sum_{i=1}^{d} \frac{\partial^2 \tilde{f}(0)}{\partial s_i \partial s_j} = -\Delta f(x),$$

we end up with

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

where

$$\omega(x) = \frac{2}{m_2} \int_{\mathbb{R}^d} (Q_{x,4}(u)h'(\|u\|^2) + Q_{x,2}(u)h(\|u\|^2))du.$$

The uniformity follows from the compactness and smoothness of $\mathcal{M}$. □

Lemma 3.2.5. *Let $f \in C^3(\mathcal{M})$ and let $0 < \gamma \leq 1/2$. Then we have, uniformly for all $x \in \mathcal{M}$ at distance less than or equal to ε^γ from $\partial\mathcal{M}$,*

$$G_\varepsilon f(x) = m_0^\varepsilon(x) f(x_0) + \sqrt{\varepsilon} m_1^\varepsilon(x) \frac{\partial f}{\partial \nu}(x_0) + \mathcal{O}(\varepsilon),$$

where x_0 is the closest point to x that belongs to the boundary and where $m_0^\varepsilon(x)$ and $m_1^\varepsilon(x)$ are bounded functions of x and ε.

Proof. Let x_0 be the closest point to x on the boundary, where closeness is measured with the norm of the ambient space. This point is uniquely defined if the boundary is smooth and ε is small enough. Let us pick a specific orthonormal basis of $T_{x_0}\mathcal{M}$ so that the first $(d-1)$ vectors $e_1, \ldots, e_{d-1}$ belong to the tangent space $T_{x_0}\partial\mathcal{M}$ of the boundary at x_0. As before, we can now consider the projections $u = (v, u_d)$ of points y in the neighborhood of x onto $T_{x_0}\mathcal{M}$, where $v = (u_1, \ldots, u_{d-1}) \in \mathbb{R}^{d-1}$ is the projection over the first $(d-1)$ basis vectors and $u_d \in \mathbb{R}$ is the projection over e_d (pointing in). By definition of x_0, we have $\langle x - x_0, e_i \rangle = 0$ for $i = 1, \ldots, d-1$, and therefore x has coordinates $(0, \eta)$ where $\eta \geq 0$. The proof is very similar to that of the previous lemma, so we can truncate the integral defining $G_\varepsilon f(x)$ by considering only points y that are at most at distance ε^γ from x. The correction term is exponentially small, and therefore can be bounded by $\mathcal{O}(\varepsilon)$. In addition to this truncation, we decompose the domain into slices. More precisely, we define

$$S(u_d) = \{(v, u_d) \in \mathbb{R}^d \ ; \ \|(v, u_d) - (0, \eta)\| \leq \varepsilon^\gamma\}.$$

To compute the integral defining $G_\varepsilon f(x)$ up to a order ε, we can integrate over all $S(u_d)$ for $u_d \in [\eta - \varepsilon^\gamma, \eta + \varepsilon^\gamma]$. Now this is not good enough as we want to take advantage of the symmetries of the kernel. We therefore consider

$$\tilde{S}(u_d) = \bigcap_{i=1}^{d-1} R_i S(u_d),$$

where R_i is the reflection on $\mathbb{R}^d$ defined by

$$R_i(u_1, \dots, u_{i-1}, u_i, u_{i+1}, \dots, u_d) = (u_1, \dots, u_{i-1}, -u_i, u_{i+1}, \dots, u_d).$$

This domain has now all the symmetries that we need. Moreover, up to a term of order ε^2, the projection of $\partial\mathcal{M}$ onto $T_{x_0}\mathcal{M}$ is a hypersurface (in $\mathbb{R}^d$) with equation $u_d = \varphi(u_1, \dots, u_{d-1})$, where φ is a homogeneous polynomial of degree 2. Consequently, up to an error of the same order, it is approximately preserved by all the reflections R_i. In particular, going from the slices $S(u_d)$ to $\tilde{S}(u_d)$ is only generating an error of order ε.

$$G_\varepsilon f(x) = \varepsilon^{-d/2} \int_{\eta-\varepsilon^\gamma}^{\eta+\varepsilon^\gamma} \int_{\tilde{S}(u_d)} h\left(\frac{\|v\|^2 + (\eta - u_d)^2}{\varepsilon}\right) \tilde{f}(u) dv du_d + \mathcal{O}(\varepsilon),$$

where $u = (v, u_d)$. For the same reason, starting the integration from $u_d = 0$ generates an error of order ε:

$$G_\varepsilon f(x) = \varepsilon^{-d/2} \int_{0}^{\eta+\varepsilon^\gamma} \int_{\tilde{S}(u_d)} h\left(\frac{\|v\|^2 + (\eta - u_d)^2}{\varepsilon}\right) \tilde{f}(u) dv du_d + \mathcal{O}(\varepsilon).$$

If we Taylor expand $\tilde{f}$ around $u = 0$, we obtain:

$$f(y) = \tilde{f}(0) + \sum_{i=1}^{d} u_i \frac{\partial \tilde{f}(0)}{\partial s_i} + \mathcal{O}(\varepsilon) = f(x_0) + \sum_{i=1}^{d-1} u_i \frac{\partial \tilde{f}(0)}{\partial s_i} - u_d \frac{\partial f}{\partial \nu}(x_0) + \mathcal{O}(\varepsilon).$$

Now, the symmetry of the kernel implies that for $i = 1, \dots, d-1$,

$$\int_{\tilde{S}(u_d)} h\left(\frac{\|v\|^2 + (\eta - u_d)^2}{\varepsilon}\right) u_i dv = 0.$$

Therefore, the only first order term of the Taylor expansion that survives is the partial derivative along u_d. We can conclude that

$$G_\varepsilon f(x) = m_0^\varepsilon(x) f(x_0) + \sqrt{\varepsilon} m_1^\varepsilon(x) \frac{\partial f}{\partial nu}(x_0) + \mathcal{O}(\varepsilon),$$

with

$$m_0^\varepsilon(x) = \varepsilon^{-d/2} \int_{0}^{\eta+\varepsilon^\gamma} \int_{\tilde{S}(u_d)} h\left(\frac{\|v\|^2 + (\eta - u_d)^2}{\varepsilon}\right) dv du_d = \int_{-\eta\sqrt{\varepsilon}}^{\varepsilon^{\gamma-\frac{1}{2}}} \int_{\frac{1}{\sqrt{\varepsilon}}\tilde{S}(u_d)} h(\|u\|^2) dv du_d$$

and

$$m_1^\varepsilon(x) = -\varepsilon^{-d/2}\int_0^{\eta+\varepsilon^\gamma}\int_{\tilde{S}(u_d)} h\Big(\frac{\|v\|^2+(\eta-u_d)^2}{\varepsilon}\Big)\frac{u_d}{\sqrt{\varepsilon}}dvdu_d = -\int_{-\eta\sqrt{\varepsilon}}^{\varepsilon^{\gamma-\frac{1}{2}}} u_d\int_{\frac{1}{\sqrt{\varepsilon}}\tilde{S}(u_d)} h(\|u\|^2)dvdu_d.$$

Clearly, these functions are well behaved as

$$|m_0^\varepsilon(x)| \leq \int_{\mathbb{R}^d} h(\|u\|^2)du \qquad \text{and} \qquad |m_1^\varepsilon(x)| \leq \int_{\mathbb{R}^d} |u_d h(\|u\|^2)|du.$$

The uniformity follows from the compactness and smoothness of $\mathcal{M}$ and of its boundary $\partial\mathcal{M}$.

□

We now use these results to obtain an asymptotic expansion for the operator $L_{\varepsilon,\alpha}$. To simplify the notations, we can assume that function h is scaled in such a way that $m_0 = 1$ and $m_2 = 2$, where m_0 and m_2 are defined in Lemma 3.2.4. We recall that, the operator Δ has eigenvalues and eigenfunctions on $\mathcal{M}$:

$$\Delta\phi_i = \lambda_i\phi_i,$$

and we suppose that ϕ_i verifies the Neumann condition $\partial\phi_i = 0$ at the boundary $\partial\mathcal{M}$. Moreover, thanks to Theorem 2.5.3 these eigenfunctions form a Hilbert basis of $L^2(\mathcal{M})$. Let

$$E_K = \text{span}\{\phi_i\ ;\ 0 \leq i \leq K\}$$

be the linear span of the first $K+1$ eigenfunctions. We even know that, thanks to Proposition 2.5.4, if we denote with $e(t,x,y) \in C^\infty(\mathbb{R}^+ \times M \times M)$ the heat kernel for functions, that

$$e(t,x,y) = \sum_i e^{-\lambda_i t}\phi_i(x)\phi_i(y).$$

Proposition 3.2.6. *For a fixed $K > 0$, we have on E_K*

$$\lim_{\varepsilon\to 0} L_{\varepsilon,\alpha}\phi = \frac{\Delta(\phi q^{1-\alpha})}{q^{1-\alpha}} - \frac{\Delta(q^{1-\alpha})}{q^{1-\alpha}}\phi.$$

Proof. We fix $0 < \gamma \leq 1/2$ and we start by focusing on the set $\mathcal{M}_\varepsilon$ of points of $\mathcal{M}$ that are at distance larger than ε^γ from $\partial\mathcal{M}$. We now that

$$q_\varepsilon(x) = q + \varepsilon(\omega q - \Delta q)$$

and that, consequently,

$$q_\varepsilon^{-\alpha} = q^{-\alpha}(1 + \alpha\varepsilon(\frac{\Delta q}{q} - \omega)).$$

Let

$$k_\varepsilon^\alpha(x,y) = \frac{k_\varepsilon(x,y)}{q_\varepsilon^\alpha(x)q_\varepsilon^\alpha(y)}$$

and for $\phi \in E_k$, define

$$K_\varepsilon^\alpha \phi(x) = \int_{\mathcal{M}} k_\varepsilon^\alpha(x,y)\phi(y)q(y)dy.$$

Then,

$$\begin{aligned}
K_\varepsilon^\alpha \phi &= q_\varepsilon^{-\alpha}(x)\int_{\mathcal{M}} k_\varepsilon(x,y)q_\varepsilon^{-\alpha}(y)q(y)\phi(y)dy \\
&= q_\varepsilon^{-\alpha}\int_{\mathcal{M}} k_\varepsilon(x,y)q^{-\alpha}(1-\alpha\varepsilon(\frac{\Delta q}{q}-\omega))q(y)\phi(y)dy \\
&= q_\varepsilon^{-\alpha}\Big(\int_{\mathcal{M}} k_\varepsilon(x,y)\phi(y)q^{1-\alpha}(y)dy + \alpha\varepsilon\int_{\mathcal{M}} k_\varepsilon(x,y)\phi(y)(\frac{\Delta q}{q}-\omega)q^{1-\alpha}dy\Big) + \mathcal{O}(\varepsilon^2) \\
&= q_\varepsilon^{-\alpha}\Big(\phi q^{1-\alpha}(1+\varepsilon(\frac{\Delta\phi q^{1-\alpha}}{\phi q^{1-\alpha}}-\omega)) + \alpha\varepsilon(\phi q^{1-\alpha}(\frac{\Delta q}{q}-\omega))\Big) + \mathcal{O}(\varepsilon^2) \\
&= q_\varepsilon^{-\alpha}\Big(\phi q^{1-\alpha} + \varepsilon\Delta(\phi q^{1-\alpha}) - \varepsilon\phi q^{1-\alpha}\omega - \alpha\varepsilon\phi\Delta q q^{1-\alpha} + \alpha\varepsilon\omega\phi q^{1-\alpha}\Big) + \mathcal{O}(\varepsilon^2) \\
&= q_\varepsilon^{-\alpha}\Big(\phi q^{1-\alpha} + \varepsilon\big(\Delta(\phi q^{1-\alpha}) - \phi q^{1-\alpha}\omega - \alpha\phi\Delta q q^{1-\alpha} + \alpha\omega\phi q^{1-\alpha}\big)\Big) + \mathcal{O}(\varepsilon^2) \\
&= q_\varepsilon^{-\alpha}\Big(\phi q^{1-\alpha} + \varepsilon\big(\alpha\phi\frac{\Delta q}{q^{\alpha-1}} + (1-\alpha)\omega\phi q^{1-\alpha} - \Delta(\phi q^{1-\alpha})\big)\Big) + \mathcal{O}(\varepsilon^2) \\
&= q_\varepsilon^{-\alpha}q^{1-\alpha}\Big(\phi + \varepsilon((1-\alpha)\omega\phi + \alpha\phi\frac{\Delta q}{q} - \frac{\Delta(\phi q^{1-\alpha})}{q^{1-\alpha}})\Big) + \mathcal{O}(\varepsilon^2).
\end{aligned}$$

Consequently,

$$d_\varepsilon^\alpha = K_\varepsilon^\alpha 1 = q_\varepsilon^{-\alpha}q^{1-\alpha}\Big(1+\varepsilon((1-\alpha)\omega + \alpha\frac{\Delta q}{q} - \frac{\Delta(q^{1-\alpha})}{q^{1-\alpha}})\Big) + \mathcal{O}(\varepsilon^2).$$

Taking the ratio of the last two equations yields the expansion for the operator

$$\int_{\mathcal{M}} \frac{k_\varepsilon^\alpha(x,y)}{d_\varepsilon^\alpha(x)}\phi(y)q(y)dy = \phi(x) + \varepsilon\left(\phi(x)\frac{\Delta(q^{1-\alpha})(x)}{q^{1-\alpha}(x)} - \frac{\Delta(\phi q^{1-\alpha})(x)}{q^{1-\alpha}(x)}\right) + \mathcal{O}(\varepsilon^2).$$

There, uniformly on $\mathcal{M}_\varepsilon$,

$$L_{\varepsilon,\alpha}\phi(x) = \frac{\Delta(\phi q^{1-\alpha})(x)}{q^{1-\alpha}(x)} - \phi(x)\frac{\Delta(q^{1-\alpha})(x)}{q^{1-\alpha}(x)} + \mathcal{O}(\varepsilon). \tag{3.15}$$

Now, on $\mathcal{M}\setminus\mathcal{M}_\varepsilon$, we have

$$q_\varepsilon(x) = m_0^\varepsilon(x)q(x_0) + \sqrt{\varepsilon}m_1^\varepsilon(x)\frac{\partial q}{\partial\nu}(x_0) + \mathcal{O}(\varepsilon),$$

which implies that

$$q_\varepsilon(x)^{-\alpha} = m_0^\varepsilon(x)^{-\alpha}q(x_0)^{-\alpha}\Big(1-\alpha\sqrt{\varepsilon}\frac{m_1^\varepsilon(x)}{m_0^\varepsilon(x)}\frac{1}{q(x_0)}\frac{\partial q}{\partial\nu}(x_0)\Big) + \mathcal{O}(\varepsilon).$$

As a consequence,

$$\begin{aligned}K_\varepsilon^\alpha\phi(x) =& q_\varepsilon(x)^{-\alpha}\Big(\frac{q(x_0)^{1-\alpha}}{m_0^\varepsilon(x)^{\alpha-1}}\phi(x_0)+\sqrt{\varepsilon}m_1^\varepsilon(x)\frac{\partial(q^{1-\alpha}(m_0^\varepsilon)^{-\alpha}\phi)}{\partial\nu}(x_0)\\&-\alpha\sqrt{\varepsilon}m_0^\varepsilon(x)^{-\alpha}q(x_0)^{1-\alpha}m_1^\varepsilon(x)\frac{1}{q(x_0)}\frac{\partial q}{\partial\nu}(x_0)\phi(x_0)\Big)+\mathcal{O}(\varepsilon)\\=& q_\varepsilon(x)^{-\alpha}\Big(\frac{q(x_0)^{1-\alpha}}{m_0^\varepsilon(x)^{\alpha-1}}\phi(x_0)+\sqrt{\varepsilon}m_1^\varepsilon(x)\frac{\partial(q^{1-\alpha}(m_0^\varepsilon)^{-\alpha})}{\partial\nu}(x_0)\phi(x_0)\\&-\alpha\sqrt{\varepsilon}m_0^\varepsilon(x)^{-\alpha}q(x_0)^{1-\alpha}m_1^\varepsilon(x)\frac{1}{q(x_0)^\alpha}\frac{\partial q}{\partial\nu}(x_0)\phi(x_0)\Big)+\mathcal{O}(\varepsilon),\end{aligned}$$

where we have used the fact that ϕ verifies the Neumann condition at x_0 and therefore can be taken out of any derivative across the boundary. Thus,

$$K_\varepsilon^\alpha\phi(x)=(K_\varepsilon^\alpha 1+\mathcal{O}(\varepsilon))\phi(x_0)$$

and since $d_\varepsilon^\alpha = K_\varepsilon^\alpha 1$, for $x\in\mathcal{M}\setminus\mathcal{M}_\varepsilon$,

$$\frac{K_\varepsilon^\alpha\phi(x)}{d_\varepsilon^\alpha(x)}=\phi(x_0)+\mathcal{O}(\varepsilon)$$

and, therefore, uniformly on $\mathcal{M}\setminus\mathcal{M}_\varepsilon$,

$$L_{\varepsilon,\alpha}\phi(x)=\mathcal{O}(1). \tag{3.16}$$

To summarize the situation:

- uniformly on $\mathcal{M}_\varepsilon$ we have
$$L_{\varepsilon,\alpha}\phi(x)=\frac{\Delta(\phi q^{1-\alpha})(x)}{q^{1-\alpha}(x)}-\phi(x)\frac{\Delta(q^{1-\alpha})(x)}{q^{1-\alpha}(x)}+\mathcal{O}(\varepsilon);$$
- uniformly on $\mathcal{M}\setminus\mathcal{M}_\varepsilon$ we have
$$L_{\varepsilon,\alpha}\phi(x)=\mathcal{O}(1).$$

Since we are interested in the L^2 convergence of this operator on $\mathcal{M}$, we have:

$$L_{\varepsilon,\alpha}\phi(x)=\frac{\Delta(\phi q^{1-\alpha})(x)}{q^{1-\alpha}(x)}-\phi(x)\frac{\Delta(q^{1-\alpha})(x)}{q^{1-\alpha}(x)}+R_\varepsilon,$$

where $R_\varepsilon=\mathcal{O}(\varepsilon)$ if $\mathcal{M}$ has no boundary and $R_\varepsilon=\mathcal{O}(\varepsilon^\gamma)$ for any $\gamma\in(0,\frac{1}{2})$ if $\mathcal{M}$ has a boundary. □

Remark 3.8. If we make no assumption on the normalization constants m_0, m_2, repeating the calculations of Theorem 3.2.6, we obtain

$$\int_{\mathcal{M}}\frac{k_\varepsilon^\alpha(x,y)}{d_\varepsilon^\alpha(x)}\phi(y)q(y)dy=\phi(x)+\varepsilon\frac{m_2}{2m_0}\left(\phi(x)\frac{\Delta(q^{1-\alpha})(x)}{q^{1-\alpha}(x)}-\frac{\Delta(\phi q^{1-\alpha})(x)}{q^{1-\alpha}(x)}\right)+\mathcal{O}(\varepsilon^2).$$

and so, uniformly on $\mathcal{M}_\varepsilon$,

$$L_{\varepsilon,\alpha}\phi(x)=\frac{m_2}{2m_0}\left(\frac{\Delta(\phi q^{1-\alpha})(x)}{q^{1-\alpha}(x)}-\phi(x)\frac{\Delta(q^{1-\alpha})(x)}{q^{1-\alpha}(x)}\right)+\mathcal{O}(\varepsilon).$$

3.2.3 The case of $\alpha = 0$

Setting $\alpha = 0$, from Proposition 3.2.6 we can observe that

$$\lim_{\varepsilon\to 0} L_{\varepsilon,0}\phi(x) = \frac{\Delta\phi q}{q} - \frac{\Delta q}{q}\phi + R_\varepsilon = \Delta\phi + 2\langle\frac{\operatorname{grad} q}{q}, \operatorname{grad}\phi\rangle.$$

This result proves that when the density is uniform, then the operator is equal to the Laplace-Beltrami operator on $\mathcal{M}$.

Remark 3.9. If we put $g = \phi q$, then we have

$$L_{\varepsilon,0}g = \frac{\Delta(gq)}{q} - g\frac{\Delta q}{q}$$

and

$$L_{\varepsilon,0}\Big(\frac{g}{q}\Big) = \Delta g - g\frac{\Delta q}{q}.$$

So, by conjugation with the density, we obtain that $L_{\varepsilon,0}$ has the form Laplacian plus potential. However, since in most application the density is not uniform, this method is clearly inappropriate if the goal is to recover the intrinsic geometry of the manifold.

3.2.4 The case of $\alpha = 1$

Let us suppose now that $\alpha = 1$. We have:

$$k^1_\varepsilon(x,y) = \frac{k_\varepsilon(x,y)}{q_\varepsilon(x)q_\varepsilon(y)}\ ,\ d^1_\varepsilon(x) = \int_{\mathcal{M}} k^1_\varepsilon(x,y)q(y)dy\ ,\ A^1_\varepsilon\phi(x) = \int_{\mathcal{M}} \frac{k^1_\varepsilon(x,y)}{d^1_\varepsilon(x)}\phi(y)q(y)dy.$$

Applying Proposition 3.2.6 to the operator $L_{\varepsilon,1}$, this leads to

$$\lim_{\varepsilon\to 0} L_{\varepsilon,1}\phi = \lim_{\varepsilon\to 0}\Delta\phi + R_\varepsilon = \Delta\phi,$$

and so we are able to recover the Laplace-Beltrami operator even if the density is not uniform. Moreover, as a byproduct, it is possible to recover the Neumann heat kernel $e^{-t\Delta}$ on $L^2(\mathcal{M})$, using $A_{\varepsilon,1}$.

Proposition 3.2.7. *For any $t > 0$, the Neumann heat kernel $e^{-t\Delta}$ can be approximated on $L^2(\mathcal{M})$ by $A^{\frac{t}{\varepsilon}}_{\varepsilon,1}$:*

$$\lim_{\varepsilon\to 0} A^{\frac{t}{\varepsilon}}_{\varepsilon,1} = e^{-t\Delta}.$$

Proof. In the Proposition 3.2.6 we showed that on E_K

$$L_{\varepsilon,1} = \Delta + R_\varepsilon \quad \text{or equivalently} \quad A_{\varepsilon,1} = I - \varepsilon\Delta - \varepsilon R_\varepsilon.$$

To obtain the result on the heat kernel, we note that

- $\overline{\bigcup_{K>0} E_K} = L^2(\mathcal{M})$

- $(A_{\varepsilon,1})_{\varepsilon>0}$ is uniformly bounded in $L^2(\mathcal{M}, dx)$ by 1

and therefore the result needs only to be proven on E_K for any fixed value of $K > 0$. We also remark that if B is a bounded operator with $\|B\| < 1$ and nonnegative spectrum, then for any $\beta > 0$

$$\begin{aligned}\|(I+B)^\beta - I\| =& \|\sum_{l\geq 1} \frac{\beta(\beta-1)\cdots(\beta-l+1)}{l!} B^l\| \leq \sum_{l\geq 1} \frac{\beta(\beta-1)\cdots(\beta-l+1)}{l!}\|B\|^l \\ \leq& (1+\|B\|)^\beta - 1 = (1+\|B\|)^\beta - \|B\| \\ \leq& (1+\|B\|)(1+\|B\|)^{\beta-1} - \|B\| \leq \beta(1+\|B\|)^{\beta-1}\|B\|.\end{aligned}$$

For a fixed $K > 0$, if ε is small enough, then $(I - \varepsilon\Delta)$ is invertible, and has norm less than 1, in which case :

$$\begin{aligned}\|A_{\varepsilon,1}^{\frac{t}{\varepsilon}} - (I-\varepsilon\Delta)^{\frac{t}{\varepsilon}}\| =& \|(I-\varepsilon\Delta-\varepsilon R_\varepsilon)^{\frac{t}{\varepsilon}} - (I-\varepsilon\Delta)^{\frac{t}{\varepsilon}}\| \\ =& \|(I-\varepsilon\Delta)^{\frac{t}{\varepsilon}}(I-\varepsilon\Delta)^{-\frac{t}{\varepsilon}}(I-\varepsilon\Delta-\varepsilon R_\varepsilon)^{\frac{t}{\varepsilon}} - (I-\varepsilon\Delta)^{\frac{t}{\varepsilon}}\| \\ =& \|(I-\varepsilon\Delta)^{\frac{t}{\varepsilon}}\left[(I-\varepsilon\Delta)^{-\frac{t}{\varepsilon}}(I-\varepsilon\Delta-\varepsilon R_\varepsilon)^{\frac{t}{\varepsilon}} - I\right]\| \\ \leq& \|(I-\varepsilon\Delta-\varepsilon R_\varepsilon)^{\frac{t}{\varepsilon}} - I\| \\ \leq& \|(I-\varepsilon R_\varepsilon)^{\frac{t}{\varepsilon}} - I\| \\ \leq& t(1+\|\varepsilon R_\varepsilon\|)^{\frac{t}{\varepsilon}-1}\|R_\varepsilon\| = \mathcal{O}(\|R_\varepsilon\|).\end{aligned}$$

Now, on E_K, one has $(I-\varepsilon\Delta)^{t/\varepsilon} = e^{-t\Delta} + \mathcal{O}(\varepsilon)$. To prove this statement we need to use the Theorem 10.33 of [9, pp 266]. In fact, if we define $f(x) = (1-\varepsilon x)^{t/\varepsilon}$, and we put $\tilde{f}(\Delta) = (I-\varepsilon\Delta)^{t/\varepsilon}$ we have

$$(I-\varepsilon\Delta)^{t/\varepsilon}\phi_k = \tilde{f}(\Delta)\phi_k = f(\lambda_k)\phi_k = (1-\varepsilon\lambda_k)^{t/\varepsilon}\phi_k$$

with $\{\lambda_k\}$ and $\{\phi_k\}$ eigenvalues and eigenfunctions of Δ. Now, since we are interested in the limit $\varepsilon \to 0$, changing the variable according to $\varepsilon \mapsto t/n$,

$$\lim_{n\to\infty}(1-\frac{t\lambda_k}{n})^n\phi_k = e^{-t\lambda_k}\phi_k.$$

This is actually how the heat operator act behaves on the eigenfunctions of the Laplacian, allowing us to affirm

$$(I-\varepsilon\Delta)^{t/\varepsilon} = e^{-t\Delta} + \mathcal{O}(\varepsilon).$$

Finally, we can conclude that

$$\lim_{\varepsilon\to 0} A_{\varepsilon,1}^{\frac{t}{\varepsilon}} = e^{-t\Delta}.$$

□

Chapter 4

Numerical experiments

This chapter is subdivided in two principal parts. The first one focuses on the discretization of the quantities involved in the diffusion process. In the second one, we will applicate the algorithm to different sets, in order to do a geometric analysis of the data.

4.1 Discretization of the continuous case

In applications, we usually have to deal with finite dataset, so we would suppose our set is composed by N points, $X \coloneqq \{x_1, x_2, \dots, x_N\}$. In particular:

Normalized Graph Laplacian

	Continuous	Discretization
Density	$d(x) = \int_X k_\varepsilon(x,y)q(y)dy$	$d(x_i) = \sum_{x_j \in X} k_\varepsilon(x_i, x_j)$
Diff operator	$A_d f(x) = \int_X \frac{k_\varepsilon(x,y)}{d(x)} f(y)q(y)dy$	$A_d f(x_i) = \sum_{x_j \in X} \frac{k_\varepsilon(x_i,x_j)}{d(x_i)} f(x_j)$
Sym operator	$A_{sym} f(x) = \int_X \frac{k_\varepsilon(x,y)}{\sqrt{d(x)}\sqrt{d(y)}} f(y)q(y)dy$	$A_{sym} f(x_i) = \sum_{x_j \in X} \frac{k_\varepsilon(x_i,x_j)}{\sqrt{d(x_i)}\sqrt{d(x_j)}} f(x_j)$

Laplace Beltrami Normalization

	Continuous	Discretization
Density	$d(x) = \int_X k_\varepsilon(x,y)q(y)dy$	$d(x_i) = \sum_{x_j \in X} k_\varepsilon(x_i, x_j)$
1^{st} norm	$\tilde{k}_\varepsilon(x,y) = \frac{k_\varepsilon(x,y)}{d(x)d(y)}$	$\tilde{k}_\varepsilon(x_i,x_j) = \frac{k_\varepsilon(x_i,x_j)}{d(x_i)d(x_j)}$
2^{nd} norm	$\tilde{d}(x) = \int_X \tilde{k}_\varepsilon(x,y)q(y)dy$	$\tilde{d}(x_i) = \sum_{x_j \in X} \tilde{k}_\varepsilon(x_i,x_j)$
Diff operator	$A_d f(x) = \int_X \frac{\tilde{k}_\varepsilon(x,y)}{\tilde{d}(x)} f(y)q(y)dy$	$A_d f(x_i) = \sum_{x_j \in X} \frac{\tilde{k}_\varepsilon(x_i,x_j)}{\tilde{d}(x_i)} f(x_j)$
Sym operator	$A_{sym} f(x) = \int_X \frac{\tilde{k}_\varepsilon(x,y)}{\sqrt{\tilde{d}(x)}\sqrt{\tilde{d}(y)}} f(y)q(y)dy$	$A_{sym} f(x_i) = \sum_{x_j \in X} \frac{\tilde{k}_\varepsilon(x_i,x_j)}{\sqrt{\tilde{d}(x_i)}\sqrt{\tilde{d}(x_j)}} f(x_j)$

4.1.1 Some details on discretization

The procedure to obtain the diffusion operator on this set is linked to the one of recovering a weighted graph on the data. In fact, a popular way to describe the affinities between data points is using a weighted graph, see for example [18],[26], [27], which vertices correspond to the data points $X = \{x_1, x_2, \dots, x_N\}$, the edges are formed between every pair of nodes, and the weights $k(x_i, x_j)$ for $i, j \in \mathbb{N}$ quantify the affinities between the nodes. In this way the function k has the twofold interpretation of kernel and affinity function. Since the quantities are finite, we define the pairwise affinity matrix K

$$K_{ij} = k(x_i, x_j)$$

for some $k(\cdot, \cdot)$. Actually, in this thesis, we use the Gaussian kernel

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\varepsilon}\right).$$

Remark 4.1. The Gaussian kernel describes well the affinities between points, in fact the smaller is the parameter ε, the faster the exponential decreases and hence the weight function k becomes numerically insignificant as we move away from the center. Furthermore, it is easy to check that this kernel satisfies the admissibility conditions required from the diffusion map's method. So, from this kernel, we construct the diffusion matrix in relation to the normalization we consider.

In order to recover the graph Laplacian normalization, we define D as the diagonal matrix

$$D_{ii} = \sum_{j=1}^{N} K_{ij},$$

then $A_d \coloneqq D^{-1}K$ is the diffusion matrix.

Remark 4.2. Thanks to the connection with the graph theory we can understand why this normalization is called graph Laplacian. This notion is linked to the definition of the Laplacian on a graph which complete construction can be found in [17]. We only point out that Laplacian matrix $\mathcal{L}$ is defined as

$$\mathcal{L} = D - K.$$

We can apply another normalization to recover the operator with which we are actually working:

$$L = D^{-1}\mathcal{L} = I - D^{-1}K.$$

L is commonly called *random walk Laplacian* on a graph.

Concerning the Laplace-Beltrami normalization, we apply the graph Laplacian normalization to a different weighted matrix:

$$\tilde{K} = D^{-1}KD^{-1}.$$

From $\tilde{K}$ we can define

$$\tilde{D}_{ii} = \sum_{j=1}^{N} \tilde{K}_{ij}$$

to recover the diffusion operator $\tilde{A}_d \coloneqq \tilde{D}^{-1}\tilde{K}$.
Our objective is to calculate the eigenvalues and eigenvectors of the diffusion operator A_d, and in order to do so we use the symmetric operator associated to it:

$$Asym = D^{1/2}A_dD^{-1/2} = D^{-1/2}KD^{-1/2}.$$

The matrix $Asym$ is known in [28] as *normalized affinity matrix*. It is easy to recover the eigenvalues and eigenvectors of $Asym$ thanks to the properties of symmetric matrices. Then we can switch to the ones of A_d thanks to the following Lemma.

Lemma 4.1.1 (Normalization Lemma)**.** *Let v be an eigenvector of Asym with eigenvalue λ, then $D^{-1/2}v$ is an eigenvector of A_d with eigenvalue λ.*

Proof. Consider the matrix $A_{sym} = D^{-1/2}KD^{-1/2}$ and let $\{\lambda_i, v_i\}_{i\geq 0}$ be the corresponding eigenvalues and eigenvectors:

$$\begin{aligned} A_{sym}v_i &= \lambda_i v_i \\ D^{-1/2}KD^{-1/2}v_i &= \lambda_i v_i \\ D^{-1/2}(D^{-1/2}KD^{-1/2}v_i) &= D^{-1/2}(\lambda_i v_i) \\ D^{-1}K\Phi_i &= \lambda_i \Phi_i \\ A_d\Phi_i &= \lambda_i \Phi_i, \end{aligned}$$

with $\Phi_i = D^{-1/2}v_i$. □

4.1.2 Algorithm and implementation

In this subsection we develop the algorithms for constructing diffusion maps on a data set $X = \{x_1, x_2, \ldots, x_N\}$, using the Gaussian kernel. We give the codes to implement the affinity matrix K and for recovering the diffusion maps.

4.1.2.1 Weighted affinity matrix

Some comments need to be done in relation to the free parameters that define the weighted affinity matrix.
First, we illustrate two different ways to chose the ε parameter of the Gaussian kernel, suggested in [15] and [29].
Lafon suggests to use ε to be of the order of the average smallest non-zero value of $\|x_i - x_j\|^2$, that is,

$$\varepsilon_{min} = \frac{1}{N}\sum_{i=1}^{N} \min_{x_i \neq x_j} \|x_i - _j\|^2.$$

In [29] the author proposes ε to be the order of the mean value of the matrix K, namely,

$$\varepsilon_{mean} = 4 * mean(K)/N.$$

There is not a priori better choice for this parameter, and for each case we need to chose it in relation to the feature we want to study.
Furthermore, besides the standard construction of the affinity matrix K

$$(K_{standard})_{ij} = k(x_i, x_j),$$

there is another construction suggested in [29]. The advice is to take into account only the points that are at distance smaller then ε, in a way that reduces the neighborhood of the point. In other words, starting from the matrix K, it is possibile to define K_ε as

$$(K_\varepsilon)_{ij} = \begin{cases} k(x_i, x_j) & \text{if } k(x_i, x_j) \leq \varepsilon \\ 0 & \text{otherwise.} \end{cases}$$

In the following we give Matlab code for the computation of the affinity matrix.

```
function [K, epsilon]=WeightMatrix(X, Params)
%INPUT: X       = n x m matrix, with n number of points and m is the
%                 dimension of the points;
%       Params  = structure composed by the following parameters
%                   - .epsilon = 'mean' or 'min';
%                   - .wm      = 'standard' or 'eps_nbr';
%OUTPUT: K      = n x n weighted matrix of a Gaussian kernel with
%                   parameter epsilon;
%       epsilon = width parameter of the kernel.
```

```
10
11 n=size(X,1);
12 sd_X=zeros(n,n);
13 for i = 1: n
14     for j = 1:n
15         sd_X(i,j)=norm(X(i,:)-X(j,:),2)^2;
16     end
17 end
18
19 if strcmp(Params.epsilon, 'mean')
20     m=4*mean(nonzeros(sd_X));
21     epsilon=m/n;
22 elseif strcmp(Params.epsilon,'min')
23     h=0;
24     for i = 1:n
25         h=h+min(nonzeros(sd_X(i,:)));
26     end
27     epsilon=h/n;
28 end
29
30 if strcmp(Params.wm, 'standard')
31     K=exp((-1/epsilon)*sd_X);
32 elseif strcmp(Params.wm, 'eps_nbr')
33     K=exp((-1/epsilon)*sd_X);
34     K(K>epsilon)=0;
35 end
```

4.1.2.2 Diffusion Maps

We notice that the diffusion map algorithm depends a priori from a large number of parameters, for example:

- number of eigenvalues to consider to recover the diffusion map;
- time of diffusion t;
- type of normalization used.

All these choices are incorporated in the code thanks to a structure that collects all this parameters.

Another remark is about the use of the Matlab computation of eigenvalues and eigenvectors of the symmetric matrix $Asym$. In literature, two Matlab functions are used to achieve this result, *eig* and *svd*. The first function is the common one to effectuate this task, the second one exploits the Singular Value Decomposition of a matrix. Some details about the singular value decomposition are given in Appendix B. We simply recall the decomposition for a symmetric matrix:

$$Asym = USV^T,$$

where, S is a diagonal matrix composed by the eigenvalues, $U = V$ in which the columns contain the eigenvectors of A.

```
function [diffusion_map, Lambda, Psi, Asym, A_diff] =
    DiffusionMap(K,Params)
%INPUT: K      = n x n affinity matrix of the data set X of
    dimension n ;
%       Params = structure composed by the following parameters
%                  - .t      = time of diffusion ;
%                  - .maxInd = dimension of the diffusion maps;
%                  - .normalization = 'lb' or 'gl' ;
%                  - .eig = 'svd' or 'standard' ;

%OUTPUT: diffusion_map = maxInd-1 x n matrix representing the
    diffusion map;
%       Lambda        = first maxInd eigenvalues;
%       Psi           = n x maxInd matrix of first maxInd
    eigenvectors;
%       Asym          = n x n symmetric matrix of diffusion;
%       A_diff        = n x n matrix of diffusion.

if exist('Params','var')&& ~isfield(Params,'t')
    Params.t = 1; % by default the time of diffusion is equal to
        1.
end

if exist('Params','var')&& ~isfield(Params,'maxInd')
    Params.maxInd = 6; %by default we only consider the first
        six eigenvalues.
end

```

```
23 D = sum(K,2);
24
25 if strcmp(Params.normalization, 'lb') % laplace-beltrami, by
       default is the graph laplacian type.
26     inverse_D = spdiags(1./D,0,size(K,1),size(K,2));
27     K = inverse_D*K*inverse_D;
28     D = sum(K,2);
29 end
30
31 inverse_sqrt_D = spdiags(sqrt(1./D),0,size(K,1),size(K,2));
32 Asym = inverse_sqrt_D * K * inverse_sqrt_D;
33
34 if exist('Params', 'var') && ~isfield(Params, 'eig')
35
36     [v,lambda] = eig(Asym);
37     [lambda,I] = sort(diag(lambda),'descend'); %by default, eig
           doesn't return the values sorted.
38     Lambda = lambda(1:Params.maxInd);
39     v      = v(:,I(1:Params.maxInd));
40 else
41     [U,S,~]= svd(Asym);
42     lambda = diag(S); %by default, svd returns the values sorted
           .
43     Lambda = lambda(1:Params.maxInd);
44     v      = U(:, 1:Params.maxInd);
45 end
46 %normalize the eigenvectors
47 v       = v./ repmat(sqrt(sum(v.^2)),size(v,1),1);
48 %pick the eigenvectors for Adiff
49 Psi         = inverse_sqrt_D * v;
50 Psi         = Psi./ repmat(sqrt(sum(Psi.^2)),size(Psi,1),1);
51 ind        = 2:length(Lambda);
52
53 clear inverse_sqrt_D
54
55 % diffusion map
56 diffusion_map = (Psi.*(repmat(Lambda'.^ Params.t,size(Psi,1),1))
       );
```

```
diffusion_map = diffusion_map(:,ind)';

inverse_D = spdiags(1./D,0,size(K,1),size(K,2));
%diffusion operator
A_diff = inverse_D*K;
```

4.1.3 Factors of Error

Since the diffusion map method focuses on discover the underlying manifold upon which the data are embedded, we can suppose that our set $X = \{x_1, \dots, x_N\}$ consists of finitely many points on $\mathcal{M}$. Then we have to concentrate on two type of errors, the one due to the fact that the points do not lie exactly on $\mathcal{M}$ and the error caused by the discretization.

About the discretization, we can consider the points as realizations of i.i.d. random variables $\{X_1, \dots, X_N\}$ with density q (supported on $\mathcal{M}$) and than it follows from the law of large numbers that as N goes to infinity, all of the discrete sums above converge at least in some weak sense and modulo a renormalization by $1/N$ to a continuous integrals (Monte Carlo integration):

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} k_\varepsilon(x, x_j) = \int_{\mathcal{M}} k_\varepsilon(x, y) q(y) dy.$$

For a finite value of N, the relative error is expected to be of the order of $\mathcal{O}(N^{-\frac{1}{2}}\varepsilon^{-d/4})$ and the same estimate should apply to the error of approximating $A_{\varepsilon,\alpha}f(x_i)$ by $\bar{A}_{\varepsilon,\alpha}f(x_i)$, where with the bar we indicate the discretization version of the continuos case. A rigorous estimates for the accuracy of the approximation is shown in [22], where the error of approximation of $A_{\varepsilon,\alpha}f(x_i)$ by $\bar{A}_{\varepsilon,\alpha}f(x_i)$ verifies

$$|\bar{A}_{\varepsilon,\alpha}f(x_i) - A_{\varepsilon,\alpha}f(x_i)| = \mathcal{O}(N^{-\frac{1}{2}}\varepsilon^{-d/4}),$$

with high probability. This bound can be further refined [23] by noticing that the numerator and denominator of the expression defining $\bar{A}_{\varepsilon,\alpha}f(x_i)$ are correlated random variables and so it is possible to derive the following estimate:

$$|\bar{A}_{\varepsilon,\alpha}f(x_i) - A_{\varepsilon,\alpha}f(x_i)| = \mathcal{O}(N^{-\frac{1}{2}}\varepsilon^{-d/4+1/2}),$$

with high probability.

Corollary 4.1.2. *In order to achieve a given precision with high probability, the number N of sample points must grow faster than $\varepsilon^{-\frac{d}{4}-\frac{1}{2}}$, where d is the dimension of $\mathcal{M}$.*

Next, we wish to explore the fact that the data points of X might not lie exactly on $\mathcal{M}$. Precisely, suppose that X is a perturbed version of $\mathcal{M}$, that is, there exists a perturbation function $\eta\colon \mathcal{M} \longrightarrow X$ with a small norm (the size of the perturbation) such that every point in X can be written as $x + \eta(x)$ for some $x \in \mathcal{M}$. The function η plays the role of some additive noise on the data. Then, assuming that the kernel k_ε is smooth, we can linearize the effect of the perturbation:

$$k_\varepsilon(x + \eta(x), y + \eta(y)) = k_\varepsilon(x, y) + \mathcal{O}(\frac{\|\eta\|}{\sqrt{\varepsilon}})$$

and, as a consequence, the perturbation of $A_{\varepsilon,\alpha}$ is the same order

$$\tilde{A}_{\varepsilon,\alpha} = A_{\varepsilon,\alpha} + \mathcal{O}(\frac{\|\eta\|}{\sqrt{\varepsilon}}) \tag{4.1}$$

where $\tilde{A}_{\varepsilon,\alpha}$ is the perturbed version of $A_{\varepsilon,\alpha}$. To obtain the effect of the perturbation on the eigenvalues and eigenfunctions, we refer to classical theorems of perturbation theory, like Weyl's theorem.

Theorem 4.1.3 (Weyl's Theorem)**.** *Let A and E be a $N \times N$ symmetric matrices. Let $\lambda_1 \geq \ldots \geq \lambda_N$ be the eigenvalues of A and $\tilde{\lambda}_1 \geq \ldots \geq \tilde{\lambda}_N$ be the eigenvalues of $\tilde{A} = A + E$. Then $|\lambda_i - \tilde{\lambda}_i| \leq \|E\|_2$.*

In this way we can use Weyl's theorem to get error bounds for the eigenvalues computed. For a complete proof of this theorem, see [20, pp 198]. We then can clearly say that, in our case

$$\sup_l |\tilde{\lambda}_l - \lambda_l| \leq \|\tilde{A}_{\varepsilon,\alpha} - A_{\varepsilon,\alpha}\|.$$

So, the bound on the error in equation (4.1) shows that

Corollary 4.1.4. *The approximation is valid as long as the scale parameter $\sqrt{\varepsilon}$ remains larger than the size of the perturbation.*

4.1.4 Information carried by the first eigenvector

In literature, [26][27][28], it is known that the dominant eigenvectors of diffusion matrix, i.e. the ones corresponding to largest eigenvalues, are supposed to extract some important information on the dataset. In this subsection we would try to understand how the first non trivial eigenvector of the diffusion matrix is able to extract important features, thanks to the connection with graph theory.

In [26], the justification follows from discussing the task of finding clusters in a weighted graph G of N nodes. The author represents a cluster using a column vector x whose i^{th} entry captures the participation of node i in that cluster. If a node does not participate

in a cluster, the corresponding entry is zero, furthermore the restriction that the norm of x is one is imposed. Then, based on the link weights of the graph k_{ij} the quantity

$$\sum_{i,j} k_{ij} x_i x_j = x^T K x$$

is defined as the measure for the cohesiveness of the cluster in G. A maximally cohesive cluster x can be found by maximizing the above expression. The Rayleigh-Ritz theorem [25, pp 176] states that the maximum value of the above expression will be λ_{max} , the maximum eigenvalue of K, and the corresponding eigenvector will be the optimal x.

Another interesting approach follows from [27]. Here the authors try to give a correct segmentation of an image I, turned into a weighted graph G, searching the optimal partition of the graph by defining the *normalized cut* disassociation measure. This leads to consider the second eigenvectors of the generalized eigenvalues system

$$(D - W)y = \lambda D y$$

with opportunity constraint, to solve, theoretically, our problem by a continuous point of view. With this method, a reiteration is recommended for further extraction.

4.2 Analysis of datasets

In this section we would like to develop the ideas illustrated so far by numerical examples: we generate sets X and we compute the eigenfunctions and eigenvalues of the diffusion operator. Then, we plot the embedding that is obtained.

4.2.1 Closed curve

We recall that, on a closed curve of length l parametrized by arc lenght, the Laplacian is merely the second derivative. In fact, let us suppose our curve $\Gamma \subset \mathbb{R}^n$ is parametrized by $s : [0, 2\pi] \longrightarrow \Gamma$, then

$$\Delta_\Gamma f(s) = -\frac{\partial^2}{\partial s^2} f(s) = -f^{''}(s).$$

Since the curve is closed, we can consider the following eigenvalue problem

$$\begin{cases} f''(x) = -\lambda f(x), \quad x \in \Gamma, \ \lambda > 0 \\ f(0) = f(2\pi) \\ f'(0) = f'(2\pi) \end{cases}$$

to find an orthonormal basis of eigenfunctions given by

$$\{1, \cos(\frac{2\pi m x}{l}), \sin(\frac{2\pi m x}{l})\}_{m=1}^{\infty}.$$

Embedding the curve Γ using the first two non trivial eigenfunctions results in a circle in the plane:

$$\Gamma(s) \mapsto \{\cos(\frac{2\pi s}{l}), \sin(\frac{2\pi s}{l})\}. \tag{4.2}$$

We applied the diffusion maps algorithm to a toroidal spiral, a trefoil knot, an epitrochoid. The results are shown in Figure (4.1). We obtain the points reorganized on a closed curve in a coherent ways with the organization of the points following the curve. Moreover, when we use the Laplace-Beltrami normalization, we recover a circle in the plane, accordingly to 4.2. 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 peek, this embedding tends to map all points around this peak to a single point, creating a corner.

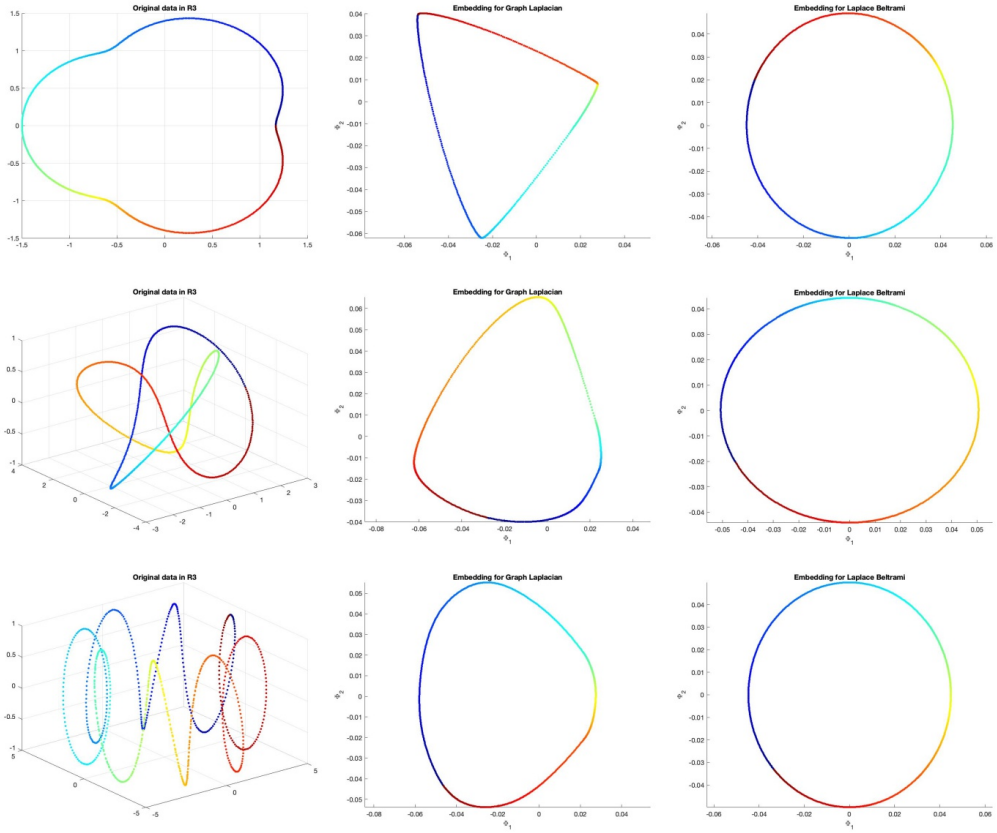

Figure 4.1: 900 randomly sampled from a toroidal spiral, a trefoil knot and an epitrochoid; their embedding using the graph Laplacian (2^{nd} column); their embedding using the Laplace-Beltrami normalization (3^{rd} column).

4.2.2 Curves with endpoints

Similarly, for a curve with two endpoints and of length l, i. e. $\Gamma \subset \mathbb{R}^n$ parametrized by $s : [0,1] \longrightarrow \Gamma$ such that $\Gamma(0) = a$ and $\Gamma(1) = b$, the Neumann eigenfunctions recover from the eigenvalue problem

$$\begin{cases} \Delta f(x) = -\lambda f(x), & x \in \Gamma \ , \ \lambda > 0 \\ f'(a) = f'(b) \end{cases}$$

are $\{1, cos(j\frac{s}{l})\}_{j=1}^{\infty}$.
Figure (4.2) shows us the embedding of a spiral helix obtained by the two normalizations and the eigenvalues and eigenvectors corresponding to the different operators. Even in this case, we can notice the ability of the algorithm to recover the organization of the points following the curve.

4.2.3 Surfaces

We now consider the spectral embedding of the Swiss Roll dataset. The Swiss Roll is a 2-D manifold embedded in $\mathbb{R}^3$, described by the following equation:

$$\begin{cases} x = \frac{1}{2} t \cos t & t \in [0, 3\pi]; \\ y = \frac{1}{2} t \sin t & t \in [0, 3\pi]; \\ z = 20\pi s & s \in [0, 1]. \end{cases}$$

One of the characteristics of the diffusion map method is the one to recover the underlying manifold on which the data lie. So, we expect to recover a rectangle in the diffusion space. Figure (4.3) presents the results obtained by the diffusion map algorithm. We can observe that both normalizations try to recover the underlying manifold, but the result obtained by graph Laplacian normalization is affected by the density of the points.

4.2.4 Image dataset

In this context we can observe that the diffusion map can have a twofold interpretation: each column of the diffusion map present the coordinates of the embedding, each row represents a feature function on the data. Some examples illustrate this idea.

4.2.4.1 Images parameterized by one real number

We study a sequence of face images from the UMIST Face Database 1-a 1-e. Each picture is a pre-cropped 112×92 pixels gray image. The diffusion map method of Laplace-Beltrami type is applied as follows:

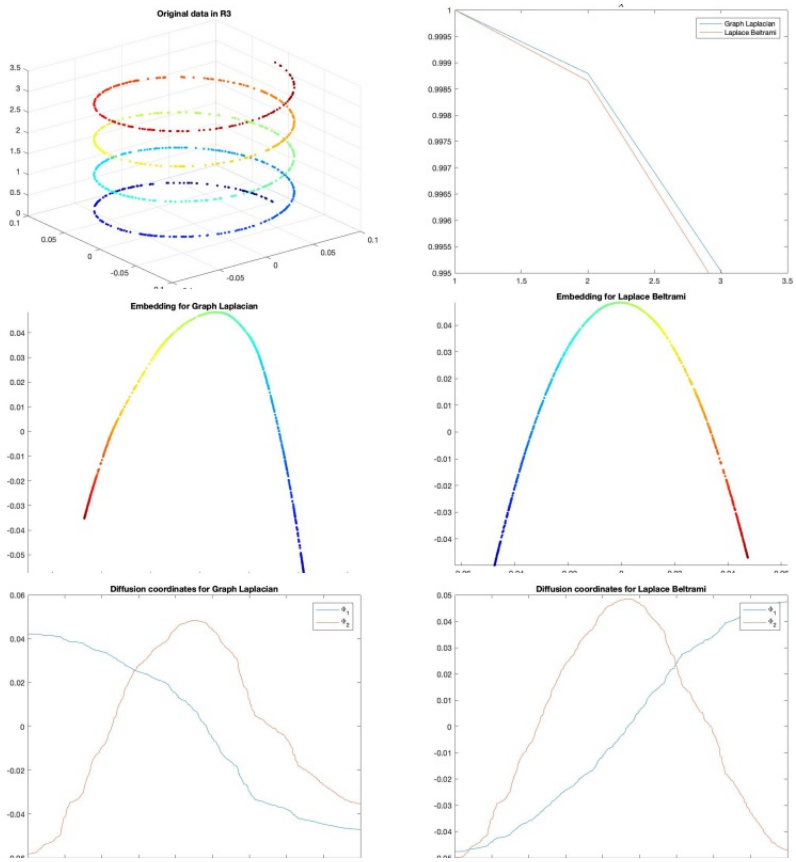

Figure 4.2: From top to bottom, from left to right: 900 randomly sampled points from a spiral helix, first eigenvalues of the diffusion operator corresponding to different normalization, embedding via graph Laplacian, embedding via Laplace Beltrami normalization, first two eigenvectors for the diffusion operator via graph Laplacian, first two eigenvectors for the diffusion operators via Laplace Beltrami type.

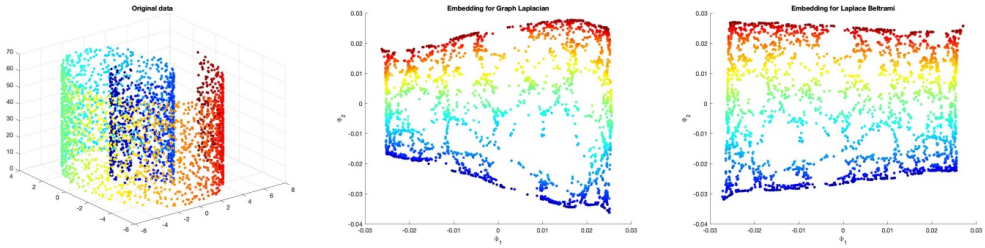

Figure 4.3: 3000 randomly sampled points from a swiss roll manifold. From left to right: original data set in $\mathbb{R}^3$, embedding via graph Laplacian, embedding via Laplace Beltrami normalization.

- Initially, the pictures are indexed by the time parameter, or equivalently, by the angle. To illustrate the capability of reorganization of the method, the pictures are randomly arranged so that they appear unordered.

- The Euclidean distances between the images in the set are measured. The eigenfunctions of the Laplace Beltrami operator in this structure are computed. Finally, the pictures are ordered according to their values.

To begin, we chose a set of 24 faces of the same person turning his head. The result about the organization is illustrated in Figure (4.4), where we can see that reparameterizing our dataset according to the reordered values of the first eigenvector, let us recover the more important parameter of our set. In fact, we can say that it is the angle of rotation of the head.

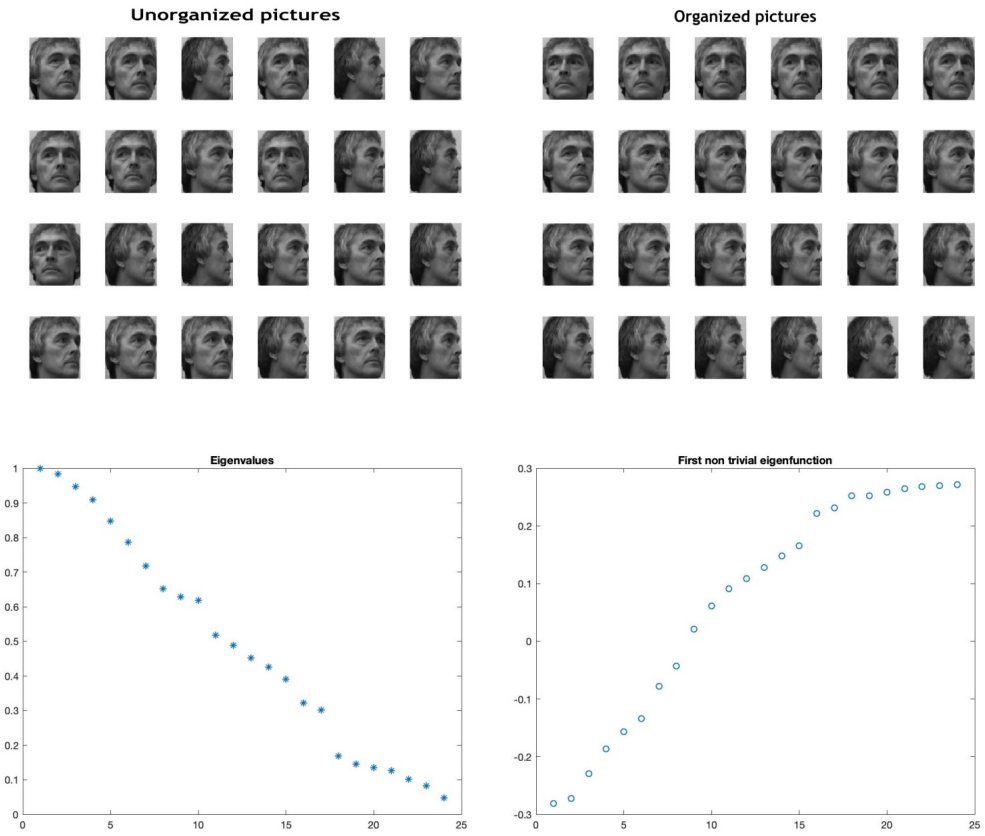

Figure 4.4: Results about images of the same person turning his head and spectral analysis.

Regarding the spectral analysis of this diffusion process, we can observe the last two pictures in Figure (4.4). The first non trivial eigenfunction associates a real number to

each image, this graph looks like that of half a period of cosine, which is the first non-trivial Neumann eigenfunction of the Laplace Beltrami on a non closed curve, as we have seen previously. Therefore, the data seem to be approximately lying along a curve in $\mathbb{R}^{112\times92}$, and the eigenfunction allows to recover the organization of the data with respect to the angle of rotation of the face. So, the method allows us to recover the fundamental feature of the set and moreover, to achieve dimensionality reduction.
Repeating the same process to a different dataset, composed by twelve images, we obtain coherent results with the previous experiment, illustrate in Figure (4.5).

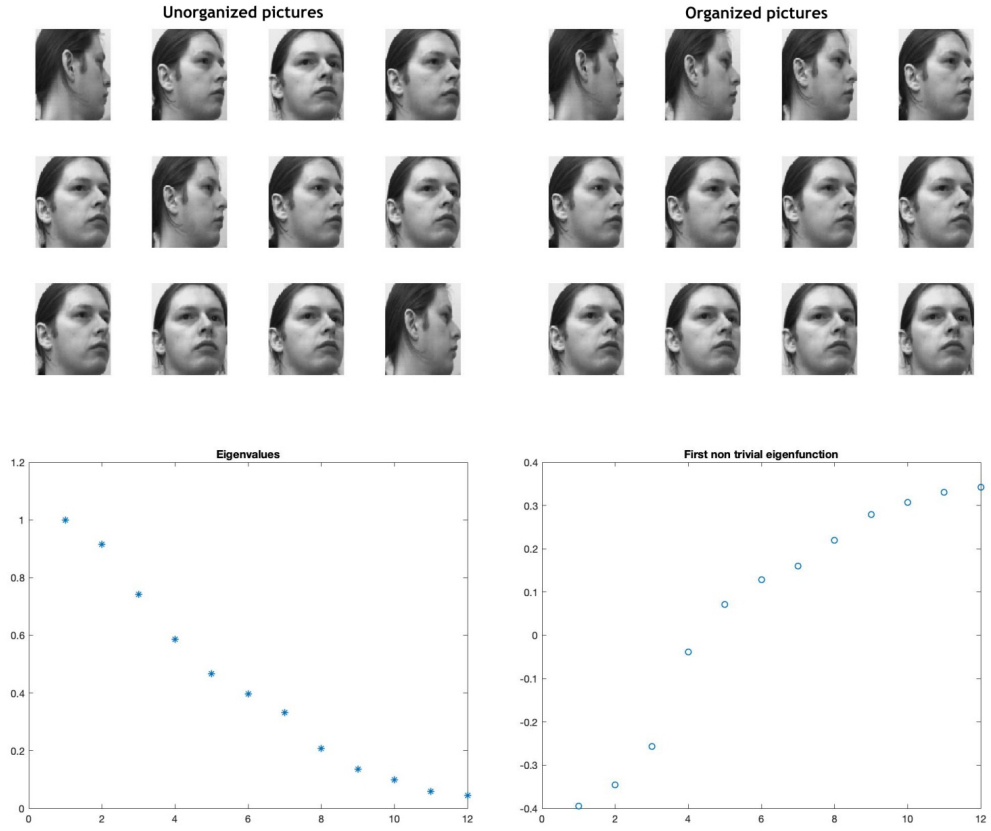

Figure 4.5: Results about images of the same person turning his head and spectral analysis.

4.2.4.2 Images parameterized by two real numbers

We study a database of images parameterized by two real numbers. More precisely, the set is composed of a sequence of 420 images (75×81 pixel) of the word "3D" viewed under different angles. The images are generated using *Blender 2.79*, a tridimensional computer graphics software. In particular, we create a three dimensional model of the

two characters 3 and D. Then, the object is rotated along the vertical axis (angle θ) and horizontal axis (angle φ), like shown on Figure (4.6). The data are sampled considering θ and φ, uniformly distributed from -20 to 20 degrees with a step of 2 degrees.

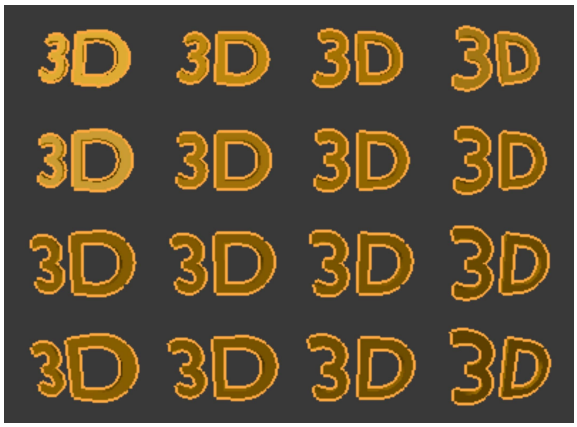

Figure 4.6: Sample of original set. The angle θ, x-axis, is discretized 20 times between -20 and 20 degrees. The angle φ, y-axis, is discretized 20 times between -20 and 20 degrees.

We apply to the data the diffusion maps algorithm, using the Laplace-Beltrami normalization with Gaussian kernel, ε_{min} and the standard construction of the matrix. We plot the image of the set by the mapping (ϕ_1, ϕ_2), as we can see in Figure (4.7).

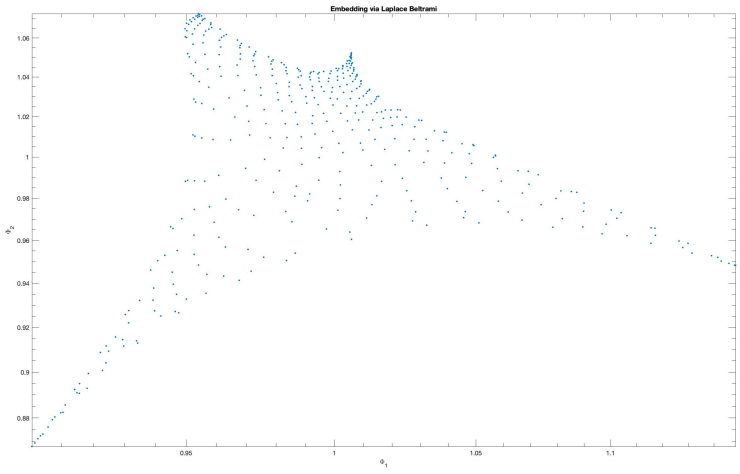

Figure 4.7: The set is mapped into $\mathbb{R}^2$ via (ϕ_1, ϕ_2).

The result is that, the orientation of the object can be controlled by the two coordinates ϕ_1 and ϕ_2, so the natural parameters of the dataset has been recovered by the algorithm. Clearly, even in this case we achieve dimensionality reduction.

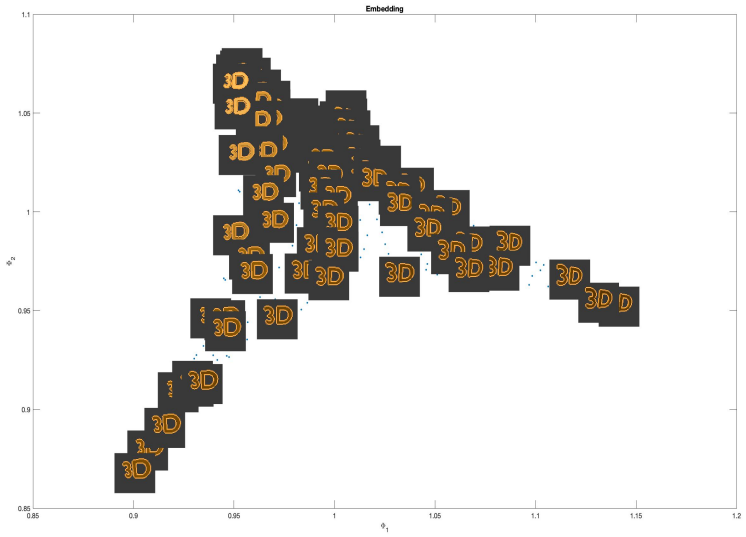

Figure 4.8: Right: some images are plotted in (ϕ_1, ϕ_2). The natural parameters θ and φ are recovered.

4.2.5 Variation of parameters

4.2.5.1 Epsilon parameter

When we implement the function to obtain the weight affinity matrix, we add the choice for the ε parameter. Let us see how this parameter is going to affect the performance of the algorithm. We consider two cases:

- $\varepsilon_{min} = \frac{1}{N}\sum_{i=1}^{N} \min_{x_i \neq x_j} \|x_i - _j\|^2$
- $\varepsilon_{mean} = 4 * mean(K)/N$

in the case of the epitrochoid, the trefoil knot, the toroidal spiral respectively. In Figure (4.9) we can recover the embedding of the curves with these different values. We can remark that there is not a priori best choice, in fact, while for the embedding via graph Laplacian normalization the choice of the parameter change mostly the sign, even if in the toroidal spiral is the same, in the embedding of the Laplace Beltrami type we see that for the trefoil knot the choice of ε_{min} recover the perfect embedding, while for the toroidal spiral the choice of ε_{mean} recover it.

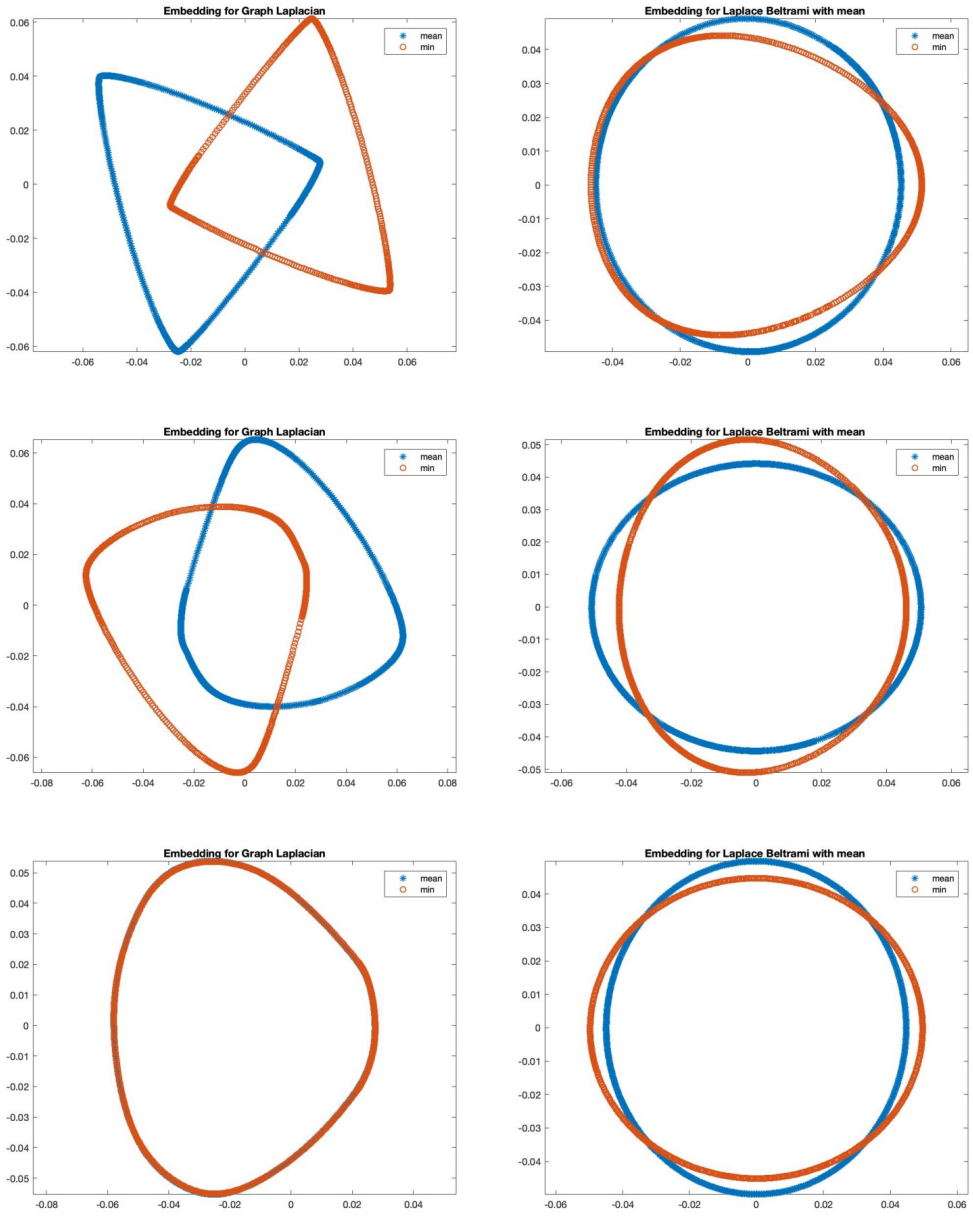

Figure 4.9: Embedding comparison between ε_{min} and ε_{mean}. Epitrochoid (1^{st} row), trefoil knot (2^{nd} row), toroidal spiral (3^{rd} row). Embedding via graph Laplacian normalization (1^{st} column), embedding via Laplace Beltrami normalization (2^{nd} column).

Figure (4.10), shows the eigenvalues that are recovered from the different type of normalization. In general, we can remark that the eigenvalues recovered from the diffusion operator obtained by the use of the parameter ε_{min} decrease rapidly towards zero, while the others have a law rate of convergence.

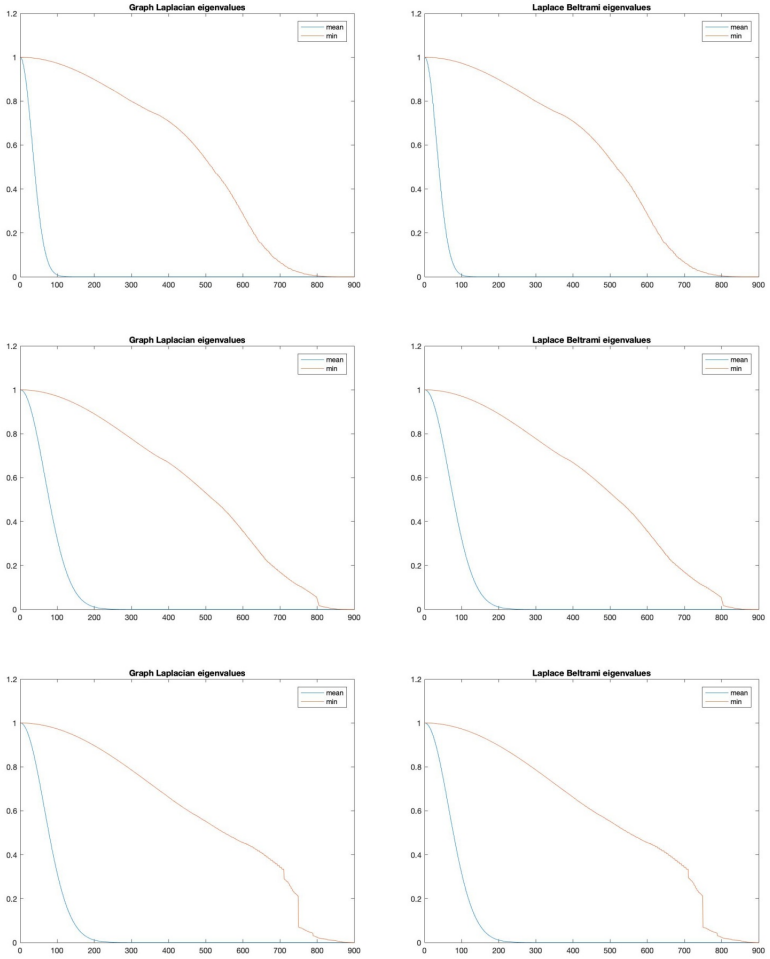

Figure 4.10: Eigenvalues comparison between ε_{min} and ε_{mean}. Epitrochoid (1^{st} row), trefoil knot (2^{nd} row), toroidal spiral (3^{rd} row). Eigenvalues of graph Laplacian normalization (1^{st} column), eigenvalues of Laplace Beltrami normalization (2^{nd} column).

4.2.5.2 Construction of affinity matrix

Here we point out the main difference of the construction of K, considering $K_{standard,\varepsilon_{mean}}$, $K_{standard,\varepsilon_{min}}$, K_{min} and K_{mean}.
Figure (4.11) illustrates the various embeddings. The first thing we notice is that the worst case to recover the geometry is the one via K_{mean}. The others methods can be compared.

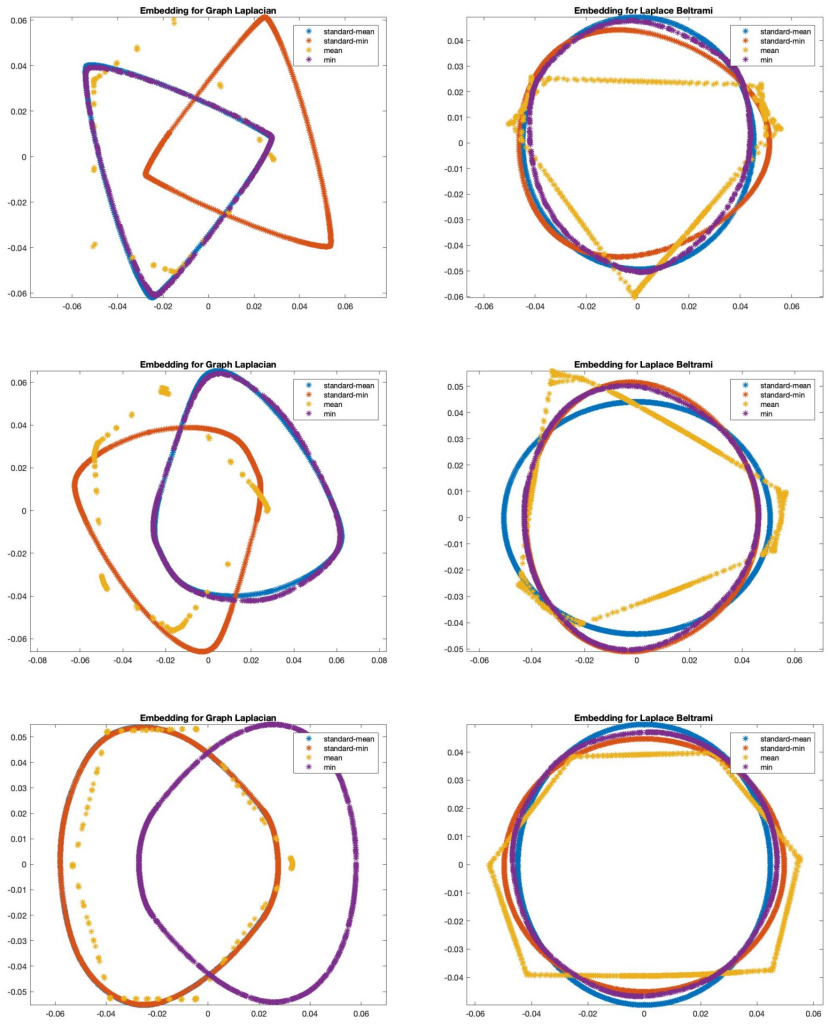

Figure 4.11: Embedding with different affinity matrix. Epitrochoid (1^{st} row), trefoil knot (2^{nd} row), toroidal spiral (3^{rd} row). Embedding via graph Laplacian normalization (1^{st} column), embedding via Laplace Beltrami normalization (2^{nd} column).

Figure (4.12) allows us to make an analysis of the eigenvalues recovered from different con-

structions. The standard construction with ε_{mean} decrease rapidly towards zero and this can make us think that it is the most useful method to achieve dimensionality reduction. The worst case is the one with the K_{mean} as confirmed by the embedding. (4.12).

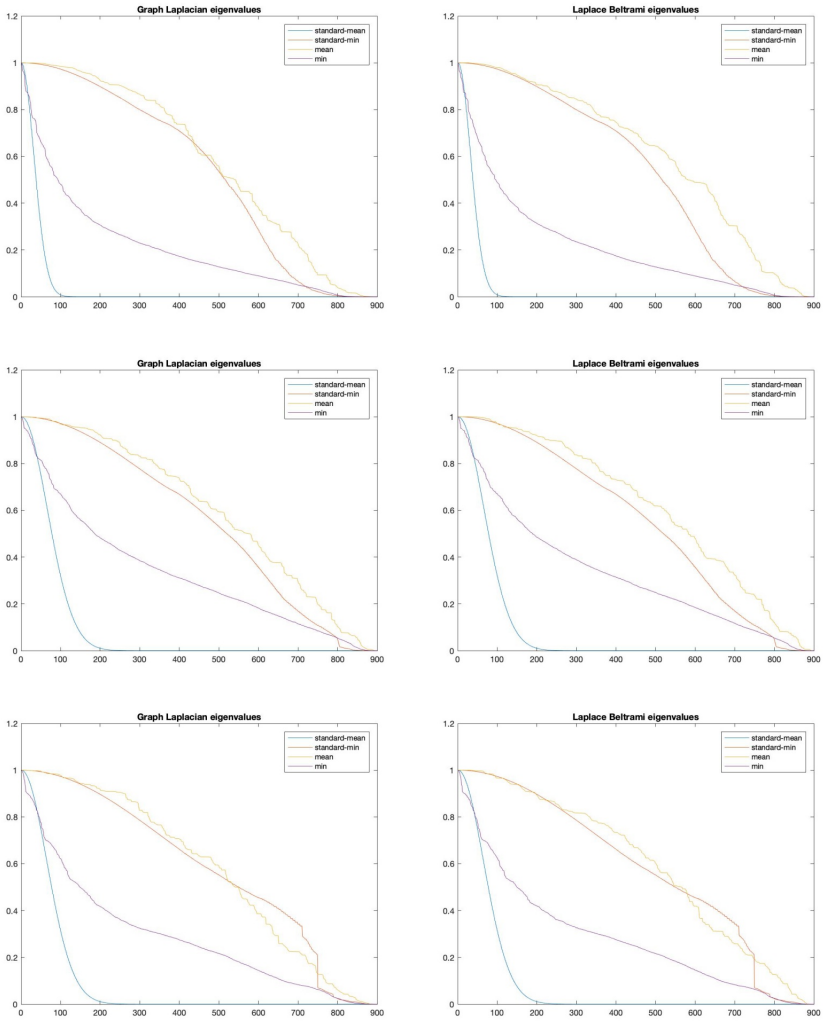

Figure 4.12: Eigenvalues of different constructions of affinity matrix. Epitrochoid (1^{st} row), trefoil knot (2^{nd} row), toroidal spiral (3^{rd} row). Eigenvalues of graph Laplacian normalization (1^{st} column), eigenvalues of Laplace Beltrami normalization (2^{nd} column).

4.2.6 Powers of Ad and multiscale geometric analysis of dataset

We have mentioned that, the kernel represents the probability of transition in one time step from node x to node y. So, for $t \geq 0$, the probability transition from x to y in t time steps is given by the kernel of the t^{th} power A_d^t of A_d. Here, the idea is that taking different powers of A_d, will reveal relevant geometric structure of the set at different scales. This will allow us to underline the fact that t plays the role of a scale parameter.
In order to illustrate this idea, we generate a set X of 900 points in the plane that is actually a union of three clusters, as shown in Figure (4.13). From this set, we build a

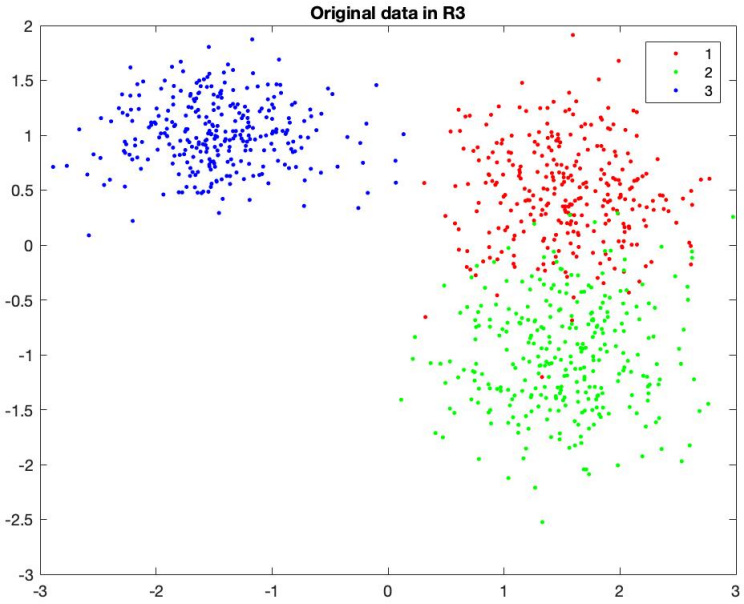

Figure 4.13: Set containing 3 clusters.

graph with Gaussian weights $e^{-\frac{\|x-y\|^2}{\varepsilon}}$ with $\varepsilon = 0.7$. On Figure (4.14) we plot several powers of the matrix A_d, namely for $t = 6$, $t = 30$ and $t = 1000$. The block structure of this powers reveals the multiscale structure of the data: at $t = 6$ the set appear to be made of 3 distinct clusters. For $t = 30$ the two closest clusters have merged. Last, at $t = 1000$, all clusters have merged. Clearly, the block structure does not depend on the specific ordering of the points, since this problem is overcome by the introduction of the diffusion coordinates that reorganize the data regardless the given ordering of the set.

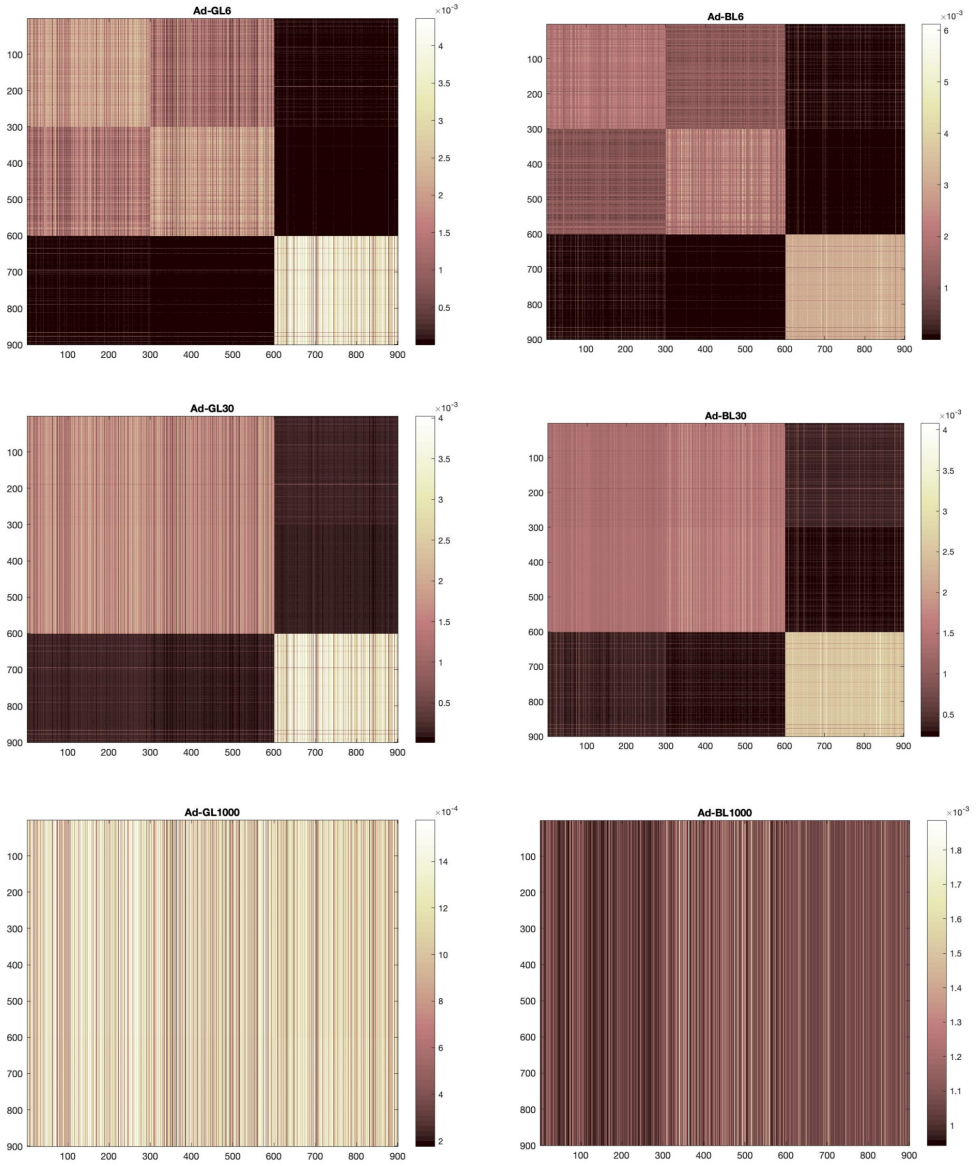

Figure 4.14: Powers of diffusion matrix. Diffusion matrix via Graph Laplacian (1^{st} column), diffusion matrix via Laplace-Beltrami (2^{nd} column). Powers $t = 6$ (1^{st} row), $t = 30$ (2^{nd} row), $t = 1000$ (3^{rd} row).

Actually, if we compare the powers of the diffusion operators obtained by the two different normalization, we can see that the Laplace-Beltrami type is faster than the graph Laplacian's. In fact, while for $t = 500$ the graph Laplacian diffusion presents a subdivision of the set, the Laplace-Beltrami type presents the set as if the three clusters have already merged.

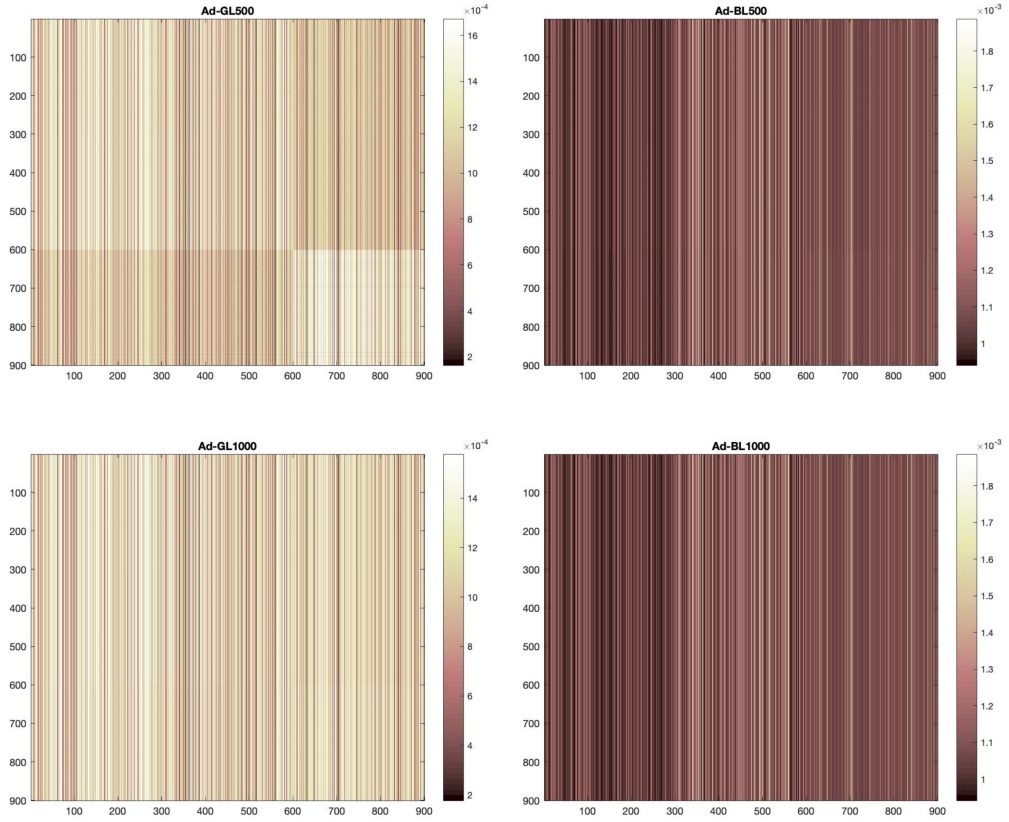

Figure 4.15: Comparison between graph Laplacian and Laplace-Beltrami normalization. Diffusion matrix via Graph Laplacian (1^{st} column), diffusion matrix via Laplace-Beltrami (2^{nd} column). Powers $t = 500$ (1^{st} row), $t = 1000$ (2^{nd} row).

4.2.7 Robustness to noise

We now illustrate the robustness of the method affected by perturbations. In Figure (4.16) we consider a perturbed version of the epitrochoid used in Figure (4.1). We represent the embedding obtained by (ϕ_1, ϕ_2) and, compared to the organization of data points that we would have obtained from a perfect epitrochoid, the results obtained are not affected by the noise, being actually unalterated.

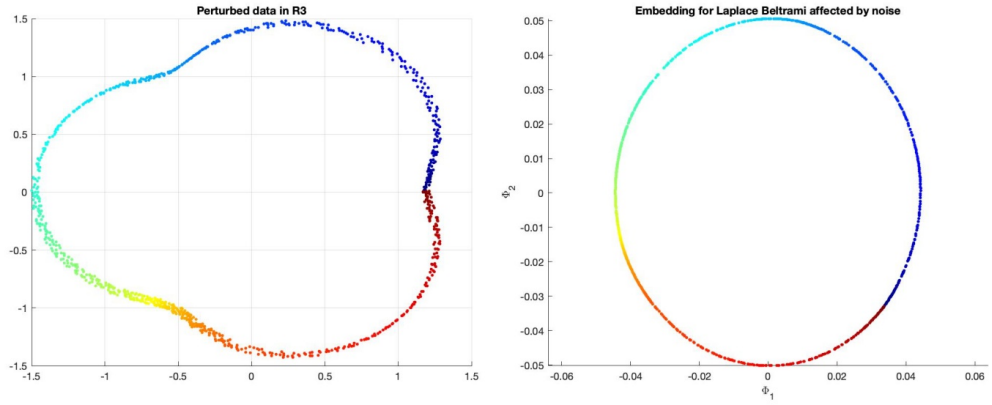

Figure 4.16: Left: the perturbed epitrochoid of Figure (4.1). Right: embedding of the curve using the firs two non trivial eigenfunctions.

Appendices

Appendix A

Conditional expectation

Consider $(\Omega, \mathcal{F}, P)$ a probability space and a random variable $X : \Omega \longrightarrow \mathbb{R}^n$.

Definition A.1. Suppose $\mathcal{G} \subset \mathcal{F}$ is a sub σ-algebra of $\mathcal{F}$. Then $E[X|\mathcal{G}]$ is the unique random variable $Y : \Omega \longrightarrow \mathbb{R}^n$ such that:

- $E[X|\mathcal{G}]$ is $\mathcal{G}$-measurable;
- $\int_G E[X|\mathcal{G}]dP = \int_G XdP$ for all $G \in \mathcal{G}$.

We list now some of the basic properties of conditional expectation:

Proposition A.0.1. *Suppose that $Z : \Omega \longrightarrow \mathbb{R}^n$ is another random variable such that $E[|Z|] < \infty$, $a, b \in \mathbb{R}^n$. Then the following states hold:*

1) $E[aX + bZ|\mathcal{G}] = aE[X|\mathcal{G}] + bE[Z|\mathcal{G}]$;

2) $E[E[X|\mathcal{G}]] = E[X]$;

3) $E[X|\mathcal{G}] = X$ *if* X *is* $\mathcal{G}$*-measurable;*

4) $E[X|\mathcal{G}] = E[X]$ *if* X *is independent of* $\mathcal{G}$;

5) $E[Z \cdot X|\mathcal{G}] = Z \cdot E[X|\mathcal{G}]$ *if* Z *is* $\mathcal{G}$*-measurable,* $\cdot$ *inner product in* $\mathbb{R}^n$.

Proof. 1) $\int_G E[aX + bZ|\mathcal{G}]dP = \int_G aX + bZdP = a\int_G XdP + b\int_G ZdP = aE[X|\mathcal{G}] + bE[Z|\mathcal{G}]$.

2) $E[E[X|\mathcal{G}]] = \int_\Omega E[X|\mathcal{G}]dP = \int_\Omega XdP = E[X]$.

3) Suppose $B \in \mathbb{R}^n$ be a Borel set, let $G = X^{-1}(B) \in \mathcal{G}$, then

$$P(X \in B) = \int_G XdP = \int_G E[X|\mathcal{G}]dP = P(E[X|\mathcal{G}] \in B).$$

4) Let $G \in \mathcal{G}$, then

$$\int_G X dP = \int_\Omega \mathbb{1}_G X dP = \int_\Omega X dP \int_\Omega \mathbb{1}_G dP = E[X]P(G) = E[X] \int_\Omega \mathbb{1}_G dP = \int_G E[X] dP;$$

furthermore $E[X]$ is $\mathcal{G}-$measurable, then $E[X|\mathcal{G}] = E[X]$ by uniqueness.

5) We need to prove the statement only for indicatrice functions and then extend the result for all measurable functions. Let $Y = \mathbb{1}_H$, $H \in \mathcal{G}$ and $G \in \mathcal{G}$. Then,

$$\int_G E[\mathbb{1}_H X|\mathcal{G}]dP = \int_G \mathbb{1}_H X dP = \int_{G\cap H} X dP = \int_G \mathbb{1}_H X dP = \int_G \mathbb{1}_H E[X|\mathcal{G}]dP.$$

□

Theorem A.0.2. *Suppose $\mathcal{G}, \mathcal{H}$ σ-algebras such that $\mathcal{H} \subset \mathcal{G}$, then $E[X|\mathcal{G}] = E[E[X|\mathcal{G}]|\mathcal{H}]$.*

Proof. Let $H \in \mathcal{H}$, since $\mathcal{H} \subset \mathcal{G}$, we have $H \in \mathcal{G}$. Then,

$$\int_H E[E[X|\mathcal{G}]|\mathcal{H}]dP = \int_H E[X|\mathcal{G}]dP = \int_H X dP.$$

It follows from unicity that $E[X|\mathcal{G}] = E[E[X|\mathcal{G}]|\mathcal{H}]$. □

Theorem A.0.3 (The Jensen Inequality)**.** *Let be $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ s.t. $E[|\phi(X)|] < \infty$, then $\phi(E[X|\mathcal{G}]) = E[\phi(X)|\mathcal{G}]$.*

Corollary A.0.4. *The following states hold:*

- $|E[X|\mathcal{G}]| \leq E[|X||\mathcal{G}]$;
- $|E[X|\mathcal{G}]|^2 \leq E[|X|^2|\mathcal{G}]$;
- *if $X_n \longrightarrow X$ in L^2, then $E[X_n|\mathcal{G}] \longrightarrow E[X|\mathcal{G}]$ in L^2.*

Appendix B

Singular Value Decomposition

The SVD is a very important decomposition which is used for many purposes. In this section $\mathbb{K}$ can be either $\mathbb{R}$ or $\mathbb{C}$, furthermore suppose that $m \geq n$.

Theorem B.0.1 (Singular values decomposition)**.** *For any $A \in \mathbb{K}^{m\times n}$, there exist two orthonormal (or unitary) matrices $U \in \mathbb{K}^{m\times m}$ and $V \in \mathbb{K}^{n\times n}$, such that*

$$A = U\Sigma V^T \ (or\ A = U\Sigma V^H), \tag{B.1}$$

where,

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix}$$

and $\Sigma_1 = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$, its diagonal elements are arranged in the order:

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r \geq 0, \qquad r = \mathrm{rank}(A).$$

The elements $\sigma_1, \sigma_2, \ldots, \sigma_r$ together with $\sigma_{r+1} = \ldots = \sigma_n = 0$ are called the singular values of matrix A. The column vector u_i of matrix U is called the left singular vector of A, and the matrix U is called the left singular matrix. The column vector v_i of matrix V is called the right singular vector of A, and the matrix V is called the right singular matrix.

The proof of the Theorem (B.0.1) can be seen in [19, 20].
The SVD of matrix A can also be written as:

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^H. \tag{B.2}$$

Recall that if A is a square $n \times n$ matrix with n linearly independent eigenvectors x_i , $i = 1, \ldots, n$, then A can be factorized as

$$A = X\Lambda X^{-1} \tag{B.3}$$

where X is the square $n \times n$ matrix whose i^{th} column is the eigenvector x_i of A and Λ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues, i.e. $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$. Note that only diagonalizable matrices can be factorized in this way.

For symmetric and Hermitian matrices, the eigenvalues and singular values are obviously closely related. A nonnegative eigenvalue, $\lambda \geq 0$, is also a singular value, $\sigma = \lambda$. The corresponding vectors are equal to each other, $u = v = x$. A negative eigenvalue, $\lambda < 0$, must reverse its sign to become a singular value, $\sigma = |\lambda|$. One of the corresponding singular vectors is the negative of the other, $u = -v = x$. This follows from the fact that all the eigenvectors of an Hermitian matrix are linearly independent, and they are mutual orthogonal, namely the eigen-matrix $X = [x_1, \ldots, x_n]$ is a unitary matrix and $X^{-1} = X^H$. So, it holds that $X^H A X = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$, or $A = X \Lambda X^H$, which can be rewritten as

$$A = \sum_{i=1}^{n} \lambda_i x_i x_i^H \tag{B.4}$$

In other words, we prove the following theorem.

Theorem B.0.2. *Let $A = U \Sigma V^T$ be the SVD of the $m \times n$ matrix A. Suppose A is symmetric, with eigenvalues λ_i and orthonormal eigenvectors x_i. In other words $A = X \Lambda X^{-1}$ is an eigendecomposition of A with $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$, $X = [x_1, \ldots, x_n]$ and $X X^T = I_d$. Then an SVD of A is $A = X \Sigma V^T$, where $\sigma_i = |\lambda_i|$ and $v_i = sgn(\lambda_i) x_i$, with $sgn(0) = 1$.*

Bibliography

[1] Arnold L. , 1974, *Stochastic Differential Equations, Theory and Applications*, John Wiley & Sons, Inc.

[2] Doob J. L. , 1953, *Stochastic Process*, John Wiley & Sons, Inc.

[3] Oksendal B. 2000, *Stochastic Dfferential Equations, An Introduction with Applications*, Fifth Edition, Corrected Printing Springer-Verlag Heidelberg New York.

[4] Bakry D. , Gentil I. , Ledoux M. , 2014, *Analysis and Geometry of Markov Diffusion Operators*, Springer International Publishing Switzerland.

[5] Nummelin E. , 1984, *General irreducible Markov chains and non-negative operators*, Cambridge University Press.

[6] Krickeberg, K. , 1963, *Wahrscheinlichkeitstheorie*, Stuttgart, Teubner.

[7] Conway J.B. , 1990, *A course in functional analysis*, Second edition, Springer-Verlag New York, Inc.

[8] Braun M.L. , 2005, *Spectral Properties of the Kernel Matrix and their Relation to Kernel Methods in Machine Learning*, Ph.D. dissertation, Fakultat der Rheinischen Friedrich-Wilhelms-Universitat Bonn.

[9] Rudin W. , 1991, *Functional Analysis*, International series in pure and applied mathematics, Second Edition, McGraw-Hill, USA.

[10] Do Carmo M. , 1992, *Riemannian Geometry*, Mathematics: theory and applications, Birkhauser Boston.

[11] Lee, J. M. , 1997, *Riemannian Manifolds, an introduction to curvature*, Springer Verlag New York, Inc.

[12] Gallot S. , Hulin D. , Lafontaine J. , 1990, *Riemannian Geometry*, Second Edition, Springer Verlag Berlin Heidelberg.

[13] Rosenberg S. , 1997, *The Laplacian on a Riemannian Manifold*, Cambridge University Press.

[14] Coifman R.R. , Lafon S. , 2006, *Diffusion maps*, Mathematics Department, Yale University, New Haven, CT 06520, USA.

[15] Lafon S. , 2004, *Diffusion Maps and Geometric Harmonics*, Ph.D. dissertation, Yale University.

[16] Coifman R.R. , Lafon S., Lee A:B. , Maggioni M. , Nadler B., Warner F. , Zucker S. , 2005, *Geometric Diffusion as a tool for harmonic analysis and structure definition of data, part I: Diffusion Maps*, Proceedings of the National Academy of Sciences.

[17] Chung F.R.K. , 1997, *Spectral Graph Theory*, American Mathematical Society.

[18] de la Porte J. , Herbst B.M. , Herman W. , van der Walt S.J. , 2008, *An introduction to Diffusion Maps*, Proceedings Nineteenth Annuals Symposium of Pattern Recognition Society of South Africa, pages 125-130.

[19] Klema, V. C., Laub, A. J. , 1980, *The singular value decomposition: Its computation and some application. IEEE Transactions on Automatic Control*, 25(2), pages 164-176.

[20] Demmel, J. W. 1997, *Applied numerical linear algebra*, Society for industrial and applied mathematics, SIAM, Philadelphia, PA, USA.

[21] Kong X. , Hu C. , Duan Z. , 2017, *Principal Component Analysis Networks and Algorithms*, Springer.

[22] Hein M. , Audibert J. , von Luxburg U. , 2005, *From graphs to manifolds- Weak and strong pointwise consistency of graph Laplacians*, in: *Conference On Learning Theory*, pages 470-485.

[23] Singer A. , 2005, *From graph to manifold Laplacian: The convergence rate*, Technical report, Yale University.

[24] Gorban Alexander N. ,Kégl Balázs , Wunsch Donald C. , Zinovyev Andrei, 2008, *Principal Manifolds for Data Visualization and Dimension Reduction*, Springer, pages 238-258.

[25] Horn R.A. , Johnson C.R. , 1990, *Matrix Analysis*, Cambridge University Press.

[26] Sarkar S. , Boyer KL. , 1996, *Quantitative measures of change based on feature organization eigenvalues and eigenvectors* , In Proc. IEEE Conf. Computer Vision and Pattern Recognition.

[27] Shi J. , Malix J. , 1997, *Normalized cuts and image segmentation*, In Proc. IEE Conf. Computer Vision and Pattern recognition, pages 731-737.

[28] Weiss Y. , 1999 *Segmentation using eigenvectors: A unifying view*, In International Conference on Computer Vision.

[29] Wang J. , 2012, *Geometric Structure of High-Dimensional Data and Dimensionality Reduction*, Higher Education Press, Beijing and Springer-Verlag Berlin Heidelberg, Chapter 14.