Topological Persistence in Market Micro Structure

Saad Labyad

University of Oxford

A thesis submitted for the degree of

Master of Science

Trinity 2018

This thesis is dedicated to my parents.

Acknowledgements

I would like to express my sincere gratitude to my supervisors for their help and encouragement.

Abstract

Persistent homology is a branch of topological data analysis that aims to identify geometric patterns in metric spaces. This information is encoded in some topological summaries and recent research has provided tools for a statistical approach to some of these summaries. We try to understand how the order flows on limit order books of "similar" stocks trigger each other using this theory.

Contents

List of Figures

Chapter 1

Introduction

More than half of global markets currently operate using a limit order book (LOB) to store the information of the orders on an electronic market [25]. Understanding the dynamics of the book, and notably the order flows is important for a market participant trying to liquidate an important position for instance. Empirical studies have lead to the identification of stylized facts of the LOB: heavy-tailed return distributions, volatility clustering, long memory in order flow and autocorrelation and long memory of returns [25].

1.1 Financial Context

Order-driven markets are trading platforms where all the information about the orders is available to all participants, as opposed to quote-driven markets where the only information comes from the quotes set and made public by market makers. In the latter, traders can only post orders at the prices proposed by the market makers, whereas they have more freedom in order-driven markets [10]. In Chapter 4 of the present work, we use data from the National Association of Securities Dealers Automated Quotations (NASDAQ) which is an American stock exchange located in New-York City.

There exists several asset classes including Equity, Bonds, Foreign exchange or Commodities for example. Market participants such as corporate managers, proprietary traders or regular investors post buy or sell orders at a given size (quantity of shares) and price. These orders are received by the market and then matched by an algorithm according to priority rules. In the case of the NASDAQ, there is a time-priority queue which will be explained in the next section. A limit order book is the structure where the information about the orders received is stored [10]. The smallest number of shares of the asset that can be traded is called the lot size, the smallest price different authorized between two trades is called the tick size. A market participant can only post orders with sizes that are multiples of the lot size and with prices that are multiples of the tick size. An order indicates the willingness of the poster to buy (or sell) shares of the stock for a given price and quantity [25].

In our work, we are interested in the two following types of orders:

- Market Orders: Orders in a given direction (buy or sell) with a size specification only,
- Limit Orders: Orders in a given direction with both a size and price specification.

Market orders seek for an immediate execution for the best available price. As there are not necessarily other market participants selling or buying the asset for any specific price at the time a limit order is posted, a limit order will most likely not be executed immediately. It will be stored in the limit order book until it gets filled totally or partially by a market order or until it gets canceled totally or partially by the market participant who posted the order.

1.2 Mathematical Approach

There exists several modeling approaches to the LOB: economic models, jump diffusion models, agent-based models, etc. In fact, one of the modeling difficulties inherent to the LOB is its complex state space structure. During the last decade, topological data analysis has emerged as an interesting method to study point cloud data. The objective of persistent homology is to identify persistent geometric structures in a metric space. Biology is one of the notable fields of application of this theory, for instance to study the maltose-binding protein in [27] and to extract Molecular Topological Fingerprints in [45] that characterize proteins. In oncology, persistent homology theory has been used to identify a subgroup of breast cancer in [34] and to forecast disease-free survival of patients with Glioblastoma tumors in [18]. It has also been successfully applied in Computer vision for the study of natural images in [8] and in Telecommunications to study the efficiency of sensors network coverage in [20]. In Finance, persistent homology has been applied to correlation networks derived from the time series of closing prices of equities listed in the NASDAQ and New York Stock Exchange in [30] and the Dow Jones Industrial Average in [24].

Current areas of research include stability [17, 12, 14] and convergence results [15, 7], the study of new summary statistics [6, 16], kernel based learning [38], and the use of persistent homology within machine learning models [35,31].

1.3 Work Outline

Our main objectives are to investigate the statical approach to persistent topology and see how it helps us understand the dynamics of the order flows. We present the construction of persistent homology in Chapter 2 as well as the stability results for robustness to model error. We define a statistical framework for topological descriptors in Chapter 3. We survey the recent statistical methods in this area for the construction of Fréchet means, hypothesis tests and confidence sets on persistence diagrams. We also approach the learning problem using Reproducing Kernels, vectorization methods and finally propose an original Perceptron model. In Chapter 4, we study the order flows of the limit order books of 7 stocks listed on NASDAQ. We conclude with a discussion on our results and further work in Chapter 5.

Chapter 2

Persistent Homology

The interpretation and the definition of the topological descriptors that we use in the next chapters is not straightforward. Therefore, this chapter aims to show the formal construction of these objects, while the next one discusses statistical inference in these settings with no specific reference to finance.

In Section 3.1 we construct the key element of this theory, the pth persistent homology. In Section 3.2 we focus on a specific step of this construction, the definition of a filtered simplicial complex because it is crucial from a computational standpoint. In Section 3.3 we define the topological descriptors we will be using, and in Section 3.4 we discuss the stability of these descriptors which is important for practical applications with noisy data. We provide proofs of the propositions when they are not provided in the references we cite.

2.1 Preliminary Definitions

In this section we follow the definitions in [22].

Definition 2.1.1 (k-simplex)**.** *Let $k \in \mathbb{N}$. Let $(x_i)_{i \in [\![0,k]\!]}$ be affinely independent points of $\mathbb{R}^d$. We say that $\sigma = conv(x_0, \dots, x_k)$ is a k-simplex where $conv()$ denotes the convex hull of a set of points.*
For all integers $j \leq k$, and indexes $(i_0, \dots, i_j) \in [\![0, k]\!]$, we say that the j-simplex $\tau = conv(x_{i_0}, \dots, x_{i_j})$ is a face of σ and we write $\tau \leq \sigma$. If $j < k$, we say that τ is a proper face of σ.

A 0-simplex is called a vertex, a 1-simplex is called an edge, a 2-simplex is called a triangle and a 3-simplex is called a tetrahedron.[1]

[1] Since the field of reference in the present work will be $\mathbb{F}_2 = \mathbb{Z}/2\mathbb{Z}$, we do not define oriented simplices.

Definition 2.1.2 (Simplicial complex)**.** *We say that a finite collection*[2] *of simplices K is a simplicial complex if:*

- $\forall \sigma \in K$*,* $\forall \tau \leq \sigma$ *we have* $\tau \in K$*.*
- $\forall \sigma, \sigma_0 \in K$ *we have* $\sigma \cap \sigma_0$ *is either empty or a face of both.*

We say that $|K| = \cup_{\sigma \in K} \sigma$ *is the underlying space of* K*. We denote by Vert(K) the set of vertices of K. For all integers* $p \in [\![0, n]\!]$*, we denote by* K_p *the set of p-simplices of* K*.*

To consider finite subsets of other metric spaces (Appendix B) which are not necessarily Euclidian, it is convenient to define abstract simplicial complexes.

Definition 2.1.3 (Abstract simplicial complex)**.** *We say that a finite collection of sets* E *is an abstract simplicial complex if* $\forall x \in E$*,* $\forall y \subset x$ *we have* $y \in E$*.*

Given a simplicial complex K with vertices $Vert(K) = (x_i)_{i \in [\![0,n]\!]}$, we build the abstract simplicial complex E given by:

$$E = \{(x_{i_0}, \dots, x_{i_k}) : k \leq n,\ (i_0, \dots, i_k) \in [\![0, n]\!],\ conv(x_{i_0}, \dots, x_{i_k}) \in K\}.$$

We say that E is a vertex scheme of K, and that K is a geometric realization of E.

To develop the rest of the theory, it would be useful to define a vector space structure on the underlying space of any simplicial complex and then consider linear maps between these spaces. An approach to construct these maps is to start by defining barycentric coordinates and maps between vertices, then extend them using convex combinations.

Proposition 1 (Barycentric coordinates)**.** *Let* K *be a simplicial complex with vertices* $(x_i)_{i \in [\![0,n]\!]}$*. There exists a unique collection of* n *maps* $(b_i)_{i \in [\![0,n]\!]}$ *such that for all* $x \in |K|$ *we have* $x = \sum_{i=0}^{n} b_i(x) x_i$*. We call the maps* $(b_i)_{i \in [\![0,n]\!]}$ *the barycentric coordinates of* x*.*

Proof. Let K be a simplicial complex with vertices $(x_i)_{i \in [\![0,n]\!]}$. Let $x \in |K|$. By definition, $x \in conv(x_{i0}, \dots, x_n)$. Thus, $\exists (\lambda_j)_{j \in [\![0,n]\!]}$ such that $x = \sum_{j=0}^{n} \lambda_j x_j$ and $\sum_{j=0}^{n} \lambda_j = 1$. This proves the existence of the maps, as $(\lambda_j)_{j \in [\![0,n]\!]}$ will depend on x.

[2] This definition can be extended to infinite sets of simplices but we will not use such complexes in the present work.

Let $(\mu_j)_{j\in[\![0,n]\!]}$ such that $x = \sum_{j=0}^{n} \mu_j x_j$ and $\sum_{j=0}^{n} \mu_j = 1$. Therefore, $\sum_{j=0}^{n}(\lambda_j - \mu_j)x_j = 0$ and $\sum_{j=0}^{n}(\lambda_j - \mu_j)$. Since the vectors $(x_i)_{i\in[\![0,n]\!]}$ are affinely independent, $\lambda_j - \mu_j = 0$ for all integers $j \in [\![0, n]\!]$. Hence the uniqueness of the maps.

Hence, for all $x \in |K|$ there exists a unique $(\lambda_j^{(x)})_{j\in[\![0,n]\!]}$ such that $x = \sum_{j=0}^{n} \lambda_j^{(x)} x_j$. We then define the barycentric coordinates as the maps $b_i : x \mapsto \lambda_i^{(x)}$ for all integers $i \in [\![0, n]\!]$. □

Definition 2.1.4 (Vertex map). *Let K, L be two simplicial complexes. A vertex map is a function $\phi : Vert(K) \mapsto Vert(L)$.*

Definition 2.1.5 (Simplicial map). *Let K, L be two simplicial complexes and ϕ a vertex map. Let n be the dimension of K and $Vert(K) = (x_i)_{i\in[\![0,n]\!]}$. Let $(b_i)_{i\in[\![0,n]\!]}$ be the barycentric coordinates of $|K|$. The simplicial map $f : |K| \mapsto |L|$ induced by ϕ is $f(x) = \sum_{i=0}^{n} b_i(x)\phi(x_i)$. By abuse of notation, we write $f : K \mapsto L$.*

We now define a vector space structure on every set of k-simplices. For the rest of this section, let K be a simplicial complex of dimension n such that $Vert(K) = (x_i)_{i\in[\![0,n]\!]}$.[3]

Definition 2.1.6 (p-chain). *Let $p \in [\![0, n]\!]$. A p-chain is a formal sum of p-simplices weighted by elements of $\mathbb{F}_2$. We denote the set of p-chains by $C_p = \{c = \sum_{i=1}^{card(K_p)} \alpha_i\sigma_i : \forall i\, \sigma_i \in K_p,\ \alpha_i \in \mathbb{F}_2\}$.*
Since $(C_p, +)$ is an abelian group, we can define a $\mathbb{F}_2$-vector space structure on C_p with the external law $. : \mathbb{F}_2 \times C_p \to C_p$ such that for all $c, c' \in C_p$, for all $\lambda, \lambda' \in \mathbb{F}_2$: $\lambda.c + \lambda'.c' = \sum_{i=1}^{card(K_p)}(\lambda.\alpha_i + \lambda'.\alpha_i')\sigma_i$.

For convenience, we denote by $conv(x_0, \ldots, \widehat{x_j}, \ldots, x_p)$ the convex envelope of $\{x_i : i \in [\![0, p]\!],\ i \neq p\}$. We now focus on specific vector subspaces of C_p.

Definition 2.1.7 (Boundary operator). *Let $p \in [\![1, n]\!]$. We define the boundary operator $\partial_p : C_p \to C_{p-1}$ by specifying first its values on K_p : $\partial_p(conv(x_{i_0}, \ldots, x_{i_p})) = \sum_{j=0}^{p} conv(x_{i_0}, \ldots, \widehat{x_{i_j}}, \ldots, x_{i_p})$. We define ∂_p on C_p by linear extension.*
We define $\partial_0 : C_0 \to \{0\}$.

Proposition 2 (Nilpotence of the boundary operator). *$\forall p > 0$, $\partial_p \circ \partial_{p+1} = 0$.*

[3] We only focus on the field $\mathbb{F}_2$ in the present work but the following definitions can be adapted to any field.

Proof. Let $p \in [\![1, n]\!]$ as this result is trivial for $p = 0$. Let $\sigma = conv(x_{i_0}, \ldots, x_{i_p+1})$. We have :

$$\partial_p \circ \partial_{p+1}(\sigma) = \partial_p(\sum_{j=0}^{p+1} conv(x_{i_0}, \ldots, \widehat{x_{i_j}}, \ldots, x_{i_{p+1}}))$$

$$\partial_p \circ \partial_{p+1}(\sigma) = \sum_{j=0}^{p+1} \partial_p(conv(x_{i_0}, \ldots, \widehat{x_{i_j}}, \ldots, x_{i_{p+1}}))$$

$$\partial_p \circ \partial_{p+1}(\sigma) = \sum_{\substack{j,k \in [\![0,p+1]\!] \\ j \neq k}} conv(x_{i_0}, \ldots, \widehat{x_{i_k}}, \ldots, \widehat{x_{i_j}}, \ldots, x_{i_p})$$

$$\partial_p \circ \partial_{p+1}(\sigma) = \sum_{\substack{j,k \in [\![0,p+1]\!] \\ j < k}} conv(x_{i_0}, \ldots, \widehat{x_{i_k}}, \ldots, \widehat{x_{i_j}}, \ldots, x_{i_p}) + \sum_{\substack{j,k \in [\![0,p+1]\!] \\ k < j}} conv(x_{i_0}, \ldots, \widehat{x_{i_k}}, \ldots, \widehat{x_{i_j}}, \ldots, x_{i_p})$$

Since the two terms on the right hand side are obviously equal, then their sum is zero in C_{p-1} because the field of reference is $\mathbb{F}_2$ (Appendix B). □

This property implies that for all $p \in [\![0, n-1]\!]$, $im(\partial_{p+1}) \subset ker(\partial_p)$, which allows us to define the homology group. We then define homology maps by extension of the notion of simplicial maps.

Definition 2.1.8 (Homology group). *For all integers $p \in [\![0, n]\!]$, we define the group of p-cycles by $Z_p = \ker(\partial_p)$.*
For all integers $p \in [\![0, n-1]\!]$, we define the group of p-boundaries by $B_p = \mathrm{im}(\partial_{p+1})$ and the pth homology group as the quotient space $H_p = Z_p/B_p$. The pth Betti number is $\beta_p := rank(H_p) = dim(Z_p) - dim(B_p)$.

Definition 2.1.9 (Induced map on homology). *Let $p \in [\![0, n]\!]$. Let L be another simplicial complex and let $f : K \mapsto L$ be a simplicial map. We define the map $f_\# : K_p \mapsto L_p$ such that :*

$$f_\#(conv(x_{i_0}, \ldots, x_{i_p})) = \begin{cases} f(conv(x_{i_0}, \ldots, x_{i_p})) & if \quad f(conv(x_{i_0}, \ldots, x_{i_p})) \in L_p \\ 0 & if \quad f(conv(x_{i_0}, \ldots, x_{i_p})) \notin L_p \end{cases}$$

By linear extension, we define $f_\# : C_p(K) \mapsto C_p(L)$ the induced map on homology.

Proposition 3 (Functoriality). *Let $p \in [\![0, n]\!]$. Let L be another simplicial complex and let $f_\# : K_p \mapsto L_p$ be a homology map. Let $\partial_{p,K}$ and $\partial_{p,L}$ be the p boundary operators on K and L.*

$$f_\# \circ \partial_{p,K} = \partial_{p,L} \circ f_\# \tag{2.1}$$

Proposition 4. *Let $p \in [\![0, n]\!]$. Let L be another simplicial complex and let $f_\# : K_p \mapsto L_p$ be a homology map.*

1. $f_\#(Z_p(K)) \subseteq Z_p(L)$,
2. $f_\#(B_p(K)) \subseteq B_p(L)$,
3. $rank(f_\#(H_p(K))) \leq min(\beta_p(K), \beta_p(L))$.

Proof. Let $c \in Z_p(K)$. By Proposition 3 we have $\partial_{p,L}(f_\#(c)) = f_\#(\partial_{p,K}(c))$. Since $\partial_{p,K}(c) = 0$ and f is linear, therefore $\partial_{p,L}(f_\#(c)) = 0$ i.e. $f_\#(c) \in Z_p(L)$. This proves (1).
Let $c \in B_p(K)$. By definition, there exists $c' \in C_{p+1}(K)$ such that $c = \partial_{p+1,K}(c')$. Then $f_\#(c) = f_\#(\partial_{p+1,K}(c'))$. By Proposition 3, $f_\#(\partial_{p+1,K}(c')) = \partial_{p+1,L}(f_\#(c'))$ therefore $c = \partial_{p+1,L}(f_\#(c'))$ i.e. $f_\#(c) \in B_p(L)$. This proves (2).
From (1) and (2), we have $f_\#(H_p(K)) \subset H_p(L)$. Hence $f_\#(H_p(K))$ is a subgroup of $H_p(L)$, thus $rank(f_\#(H_p(K))) \leq rank(H_p(L)) = \beta_p(L)$. Since $f_\#$ is a group morphism, $rank(f_\#(H_p(K))) \leq rank(H_p(K)) = \beta_p(K)$. This proves (3). □

Definition 2.1.10 (Filtered simplicial complex). *We say that K is a filtered simplicial complex if there exists an integer $m > 0$ and a family of simplicial complexes $(K_i)_{i \in [\![1,m]\!]}$ such that for all $i \in [\![1, m-1]\!]$ $K_i \subset K_{i+1}$ and $K_m = K$.*

The next section will emphasize some classic ways to construct filtrations.

Definition 2.1.11 (pth persistent homology). *We consider an integer $p \in [\![0, n]\!]$. Let K be a filtered simplicial complex with subcomplexes $(K_i)_{i \in [\![1,m]\!]}$. For all $i, j \in [\![1, m]\!]$ such that $i \leq j$, let $f_{i,j} : K_i \mapsto K_j$ be the inclusion map between K_i and K_j. We let $f_\# : H_p(K_i) \mapsto H_p(K_j)$ be the induced homology map. The pth persistent homology is the set $\{(H_p(K_i))_{i \in [\![1,m]\!]}, (f_{i,j})_{1 \leq i \leq j \leq m}\}$.*

2.2 The Vietoris–Rips Complex

Let $(\mathbb{X}, d_\mathbb{X})$ be a metric space. To obtain the persistent homology sets of $\mathbb{X}$, we have to define a filtered simplicial complex with vertices the elements of $\mathbb{X}$. In practice, we aim to minimize the computational cost of generating this complex.
In [21], the authors propose the following general approach to build filtered simplicial complexes in $\mathbb{R}^n$. We consider the set $\mathcal{F}(\mathbb{R}^n) = \{E \subset \mathbb{R}^n | E \text{ finite}\}$ and a Borel measurable function $f : \mathcal{F}(\mathbb{R}^n) \mapsto \mathbb{R}^+$ such that:

1. $f(\tau) \leq f(\sigma)$ if $\tau \subset \sigma$,

2. $f(\sigma + x) = f(\sigma) \quad \forall \sigma \in \mathcal{F}(\mathbb{R}^n)$, $x \in \mathbb{R}^n$,

3. $\exists \rho : \mathbb{R}^+ \mapsto \mathbb{R}^+$ increasing and such that $\|x - y\|_2 \leq \rho(f(\{x, y\})) \quad \forall x, y \in \mathbb{R}^n$.

Proposition 5. *Let $\sigma \in \mathcal{F}(\mathbb{R}^n)$. $\forall r > 0, \quad F(\sigma, r) = \{\tau \subset \sigma | f(\tau) \leq r\}$ is a simplicial complex.*

We focus on the two following cases and their generalization in any metric space:

- $f(\{x_1 \dots x_p\}) = \inf\limits_{y \in \mathbb{R}^n} \max\limits_{i \in [\![1,p]\!]} \|x_i - y\|_2$. $F(\sigma, r)$ is called the Čech complex.
- $f(\{x_1 \dots x_p\}) = \max\limits_{i,j \in [\![1,p]\!]} \|x_i - x_j\|_2$. $F(\sigma, r)$ is called the V̌ietoris-Rips complex.

We follow the definitions of the two following complexes in [22].

Definition 2.2.1 (Čech complex)**.** *For all values of the filtration parameter $r > 0$, we define the intrinsic Čech complex on the vertex set $\mathbb{X}$ as $\check{C}ech(\mathbb{X}, r) = \{\ [x_0, \dots, x_k] : \cap_{i=0}^{k} B(x_i, r) \neq \emptyset\}$.*

Example 2.2.1. *Consider the metric space $(\mathbb{R}^2, d_2)$ where d_2 is the distance induced by the L_2 norm. We consider a set of points:*

$$\mathbb{X} = \{(0.45, 0.45), (0.4, 0.6), (0.6, 0.7), (0.9, 0.45), (0.75, 0.7), (0.8, 0.9)\}.$$

Let $r = 0.18$. Geometrically, the Čech complex of vertex set $\mathbb{X}$ with filtration parameter r is the simplicial complex which k-simplices are the convex hulls of k points which minimum enclosing ball is of radius r at most. We plot this complex in Figure 2.1.

The transparent gray circles are the circles centered in the points of $\mathbb{X}$ of radius r. They are the vertices (0-simplices) of $\check{C}ech(\mathbb{X}, r)$. Each blue line is an edge (1-simplex) of $\check{C}ech(\mathbb{X}, r)$ (the convex hull of two distinct points is a line). Each yellow triangle is a triangle (2-simplex) of $\check{C}ech(\mathbb{X}, r)$ (the convex hull of three points in general position in the plane is the inside surface of a triangle). We note that there is no combination of 4 or more circles with a common intersection. Therefore there is no simplex of dimension 3 or more in $\check{C}ech(\mathbb{X}, r)$ and its dimension is exactly 2.

Definition 2.2.2 (Vietoris–Rips complex)**.** *For all values of the filtration parameter $r > 0$, we define the Vietoris-Rips complex on the vertex set $\mathbb{X}$ as $Rips(\mathbb{X}, r) = \{\ [x_0, \dots, x_k] : \forall i, j\ d_X(x_i, x_j) \leq r\}$.*

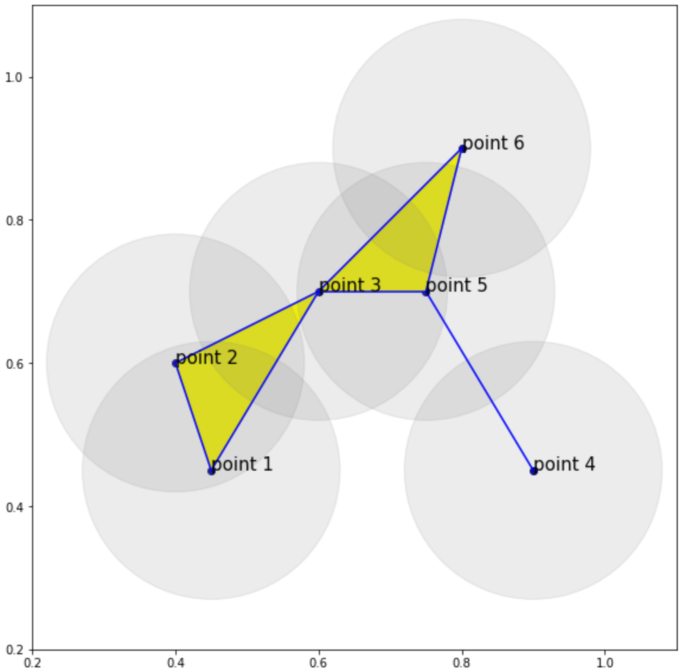

Figure 2.1: Čech complex of $\mathbb{X}$ with filtration parameter r.

Example 2.2.2. *Consider again the metric space* $(\mathbb{R}^2, d_2)$. *We consider another set of points:*

$$\mathbb{X} = \{(0.45, 0.45), (0.5, 0.6), (0.6, 0.6), (0.6, 0.7]), (0.8, 0.45), (0.75, 0.7), (0.6, 0.8), (0.7, 0.8), (0.8, 0.9)$$

Let $r = 0.18$. *Geometrically, the Vietoris–Rips of vertex set* $\mathbb{X}$ *with filtration parameter* r *is the simplicial complex which* k*-simplices are the convex hulls of* k *points with pairwise distances* r *at most. We plot this complex in Figure 2.2.*

The transparent gray circles are the circles centered in the points of $\mathbb{X}$ *of radius* r. *Again, the points are the vertices of Rips*$(\mathbb{X}, r)$, *the blue lines are* 1*-simplices and the yellow triangles are the* 2*-simplices. We note that the points 4,6,7 and 8 have pairwise distances inferior to* r. *Therefore their convex envelope is a tetrahedron (3-simplex) of Rips*$(\mathbb{X}, r)$. *There are no* 4*-simplices in Rips*$(\mathbb{X}, r)$, *its dimension is exactly* 3.

In the present work, we use the software Ripser [4] to compute the Vietoris–Rips complex of our point clouds. Ripser is a C++ code that computes the Vietoris-Rips complex of a given point cloud inputted as a distance matrix, and uses it to output the corresponding barcode, one of the the topological descriptors that we present in the next section. There is no peer-reviewed software paper for Ripser to our knowledge, but we use it because it is presently the fastest code for the computation of the homology induced by the Vietoris–Rips complex [36].

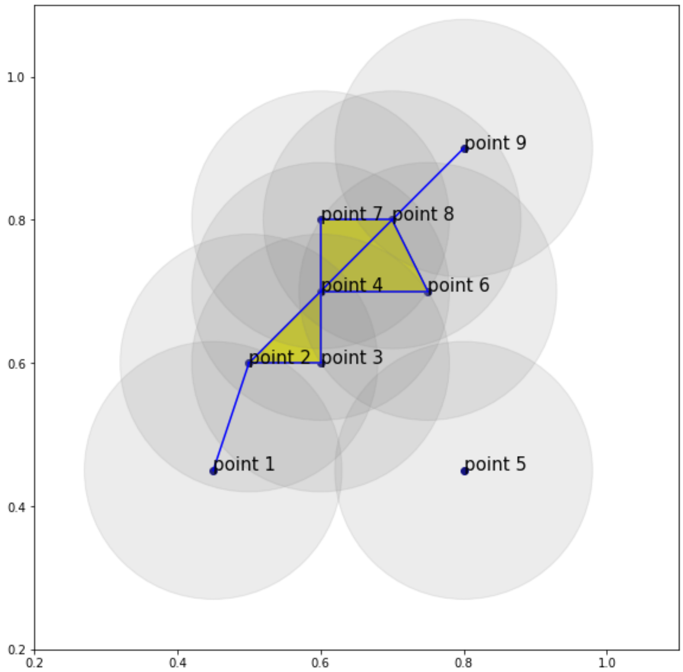

Figure 2.2: Vietoris-Rips complex of $\mathbb{X}$ with filtration parameter r.

2.3 Functional Summaries

Let K be a filtered simplicial complex with subcomplexes $(K_i)_{i\in[\![1,m]\!]}$. In this section, we focus on descriptors of the pth persistent homology of K, $\{(H_p(K_i))_{i\in[\![1,l]\!]},(f_{i,j})_{1\leq i\leq j\leq l}\}$ given an integer $p\in[\![0,n]\!]$. The following definitions are taken from [36].

Definition 2.3.1 (Barcode)**.** *A barcode B_p is a finite multiset of intervals representing the birth and death of a p-homology class.*
Formally, for each filtration step $i\in[\![1,m]\!]$, let $U^{(i)}=(u_b^{(i)})_{b\in[\![1,\beta_p(K_i)]\!]}$ be a basis of $H_p(K_i)$ chosen accordingly to [46 Corollary 4.1] . Let $U=\bigcup_{i=1}^m U^{(i)}$ be the set of all these basis vectors and $\forall u\in U$, $I(u)=\{i:i\in[\![1,m]\!],u\in U^{(i)}\}$. Then :

$$B_p:=\{[\,min(I(u)),\overline{max}(I(u))):u\in U\},$$

where $\overline{\max}(I(u))=\max(I(u))$ if $\max(I(u))<m$, and $\overline{\max}(I(u))=+\infty$ otherwise.

In practice, the usual graphical representation for barcodes is the following diagram : for all $[\,a_i,b_i)\in B_p$, we draw a horizontal line from a_i to b_i if $b_i<+\infty$. If b_i is not finite, we usually stop the line at an arbitary threshold. The y-coordinates of these horizontal lines are arbitrary.

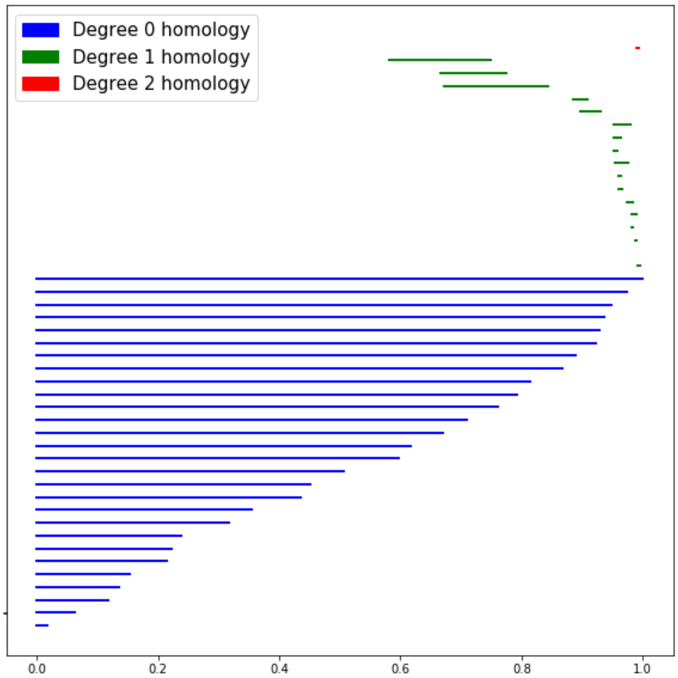

Figure 2.3: Barcode example

Example 2.3.1.

Definition 2.3.2 (Persistence diagram). *Let $\overline{\mathbb{R}} = \{\mathbb{R} \cup \{\pm\infty\}\}$. Let B be the barcode of K. The pth persistence diagram $dgm_p(K)$ induced by K is the finite multiset of points in $\overline{\mathbb{R}}$ such that $dgm_p(K) := \{(a_i, b_i) : [a_i, b_i) \in B_p\}$.*

We denote by $\mathcal{D}$ the set of persistence diagrams.

Example 2.3.2. *We plot the corresponding persistence diagram for the previous example.*

2.4 Stability

In this section we define various metrics on $\mathcal{D}$ which will be useful for the statistical methods of the following chapter. In applications of persistent homology with empirical data, it is important to quantify the noise in persistence diagrams induced by the noise of the observations. This is made possible by stability theorems.
Let d be a metric on $\mathbb{R}^2$. For any line $L \subset \mathbb{R}^2$, we denote by $\pi_L(d) : \mathbb{R}^2 \mapsto \mathbb{R}^2$ the projection on the line L with respect to the distance d. If the projection is not unique (it could be the case if d is note induced by a norm), we chose arbitrarily one of the

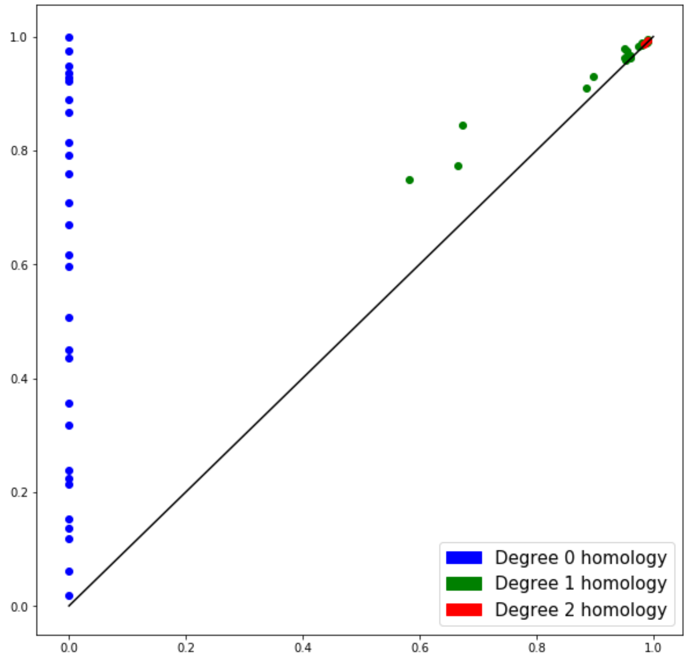

Figure 2.4: Persistence diagram example

vectors realizing the minimum.We denote by Δ the diagonal line of $\mathbb{R}^2$, i.e. the set $\Delta = \{(x, x) : x \in \mathbb{R}^2\}$.

For two finite subsets $X, Y \subset \mathbb{R}^2$ such that $card(X) \leq card(y)$ for instance, we say that $\gamma(d)(X, Y)$ is a matching between X, Y if there exists $\phi : X \mapsto Y$ injective such that $\gamma(d)(X, Y) = \{(x, \phi(x)) \mid x \in X\} \cup \{(y, \pi_\Delta(d)(y)) \mid y \in Y\}$. We denote by $\Gamma(d)(X, Y)$ the set of matchings $\gamma(d)(X, Y)$.

2.4.1 A first metric on $\mathcal{D}$

In the literature, the strongest stability results are the ones for the Bottleneck distance shown in [14].

Definition 2.4.1 (Bottleneck distance)**.** *Let $D_1, D_2 \in \mathcal{D}$. The Bottleneck distance between D_1 and D_2 is defined as :*

$$W_\infty[d](D_1, D_2) = \inf_{\gamma \in \Gamma(d)(D_1, D_2)} \{ \sup_{(x,y) \in \gamma} \{d(x, y)\}\}$$

Proposition 2.4.1. *The Bottleneck distance is a metric on $\mathcal{D}$.*

Proof. Let $D_1, D_2, D_3 \in \mathcal{D}$. $W_\infty[d](D_1, D_2) \geq 0$ because d verifies the separation axiom. $W_\infty(D_1, D_2) = 0$ if and only if there exists $\gamma \in \Gamma(d)(D_1, D_2)$ such that

$d(x_1, x_2) = 0$ for all $x_1, x_2 \in \gamma$. Therefore for all $x_1, x_2 \in \gamma$ we have $x_1 = x_2$, hence $D_1 = D_2$. $W_\infty[d]$ is clearly symmetric and subadditive because d is, therefore it is indeed a metric on $\mathcal{D}$. □

We use the notation $W_\infty = W_\infty[d_\infty]$ where d_∞ is the distance induced by the L_∞ norm, because it is the most used metric for the Bottleneck distance in the literature. Before presenting the stability theorem for the Bottleneck distance, we define a specific class of functions to which it applies to following [22].

Definition 2.4.2 (Triangulable space). *Let $\mathbb{X}$ be a topological space. We say that $\mathbb{X}$ is triangulable if there exists a simplicial complex K such that $\mathbb{X}$ is homeomorphic to $|K|$. We say that $\mathcal{T} = (K, |K|)$ is a triangulation of $\mathbb{X}$.*

Definition 2.4.3 (Tame function). *Let $\mathbb{X}$ be a triangulable topological space. Let $f : \mathbb{X} \to \mathbb{R}$ continuous. We define the sublevel sets of f as $\mathbb{X}_a = f^{-1}(-\infty, a]$ and the homology maps $f_p^{a,b} : H_p(\mathbb{X}_a) \to H_p(\mathbb{X}_b)$. We say that $\mathrm{im}(f_p^{a,b})$ is the pth persistent homology group and define its Betti number $\beta_p^{a,b} = rank(\mathrm{im}(f_p^{a,b}))$. We say that $\alpha \in \mathbb{R}$ is a homological critical value if $\forall \epsilon > 0$, $\exists p \in \mathbb{N}$ such that $f_p^{a-\epsilon,a+\epsilon}$ is not an isomorphism.*
We say that f is a tame function if:

1. *It has finitely many homological critical values,*

2. *$\forall a, b \in \mathbb{R}, a < b$, $\forall p \in \mathbb{N}\, \beta_p^{a,b} < +\infty$.*

Theorem 1 (Stability theorem for the Bottleneck distance). *Let X be a triangulable topological space. For any tame functions $f_1, f_2 : \mathbb{X} \to \mathbb{R}$ with pth persistence diagrams $D_1 = dgm(f_1)$ and $D_2 = dgm(f_2)$, we have:*

$$W_\infty(D_1, D_2) \leq \|f_1 - f_2\|_\infty.$$

There exists a stability inequality for the bottleneck distance using the Gromov-Hausdorff distance which will be useful for statistical inference [13].

Definition 2.4.4 (Hausdorff distance). *Let A, B be two closed subsets of a metric space. The Hausdorff distance between A and B is defined as :*

$$W_H[d](A, B) = \max\{\sup_{a\in A} d(a, B), \sup_{b\in B} d(b, A)\}.$$

Definition 2.4.5 (Gromov–Hausdorff distance). *Let A, B be two closed metric spaces. The Gromov-Hausdorff distance between A and B is defined as :*

$$W_{GH}[d](A, B) = \inf\{W_H[d](f(A), g(B)) : \mathbb{X}\, \textit{metric space}, f : A \to \mathbb{X}, g : B \to \mathbb{X}\, \textit{isometries}\}.$$

2.4.2 Resolution-stability tradeoff

A third metric on $\mathcal{D}$, the Wasserstein distance defined below, captures the difference between point clouds in diagrams in more detail. Nonetheless, the associated stability theorem is applicable to a smaller functional space [22].

Definition 2.4.6 (Wasserstein distance). *Let $D_1, D_2 \in \mathcal{D}$. The degree-q Wasserstein distance between D_1 and D_2 is defined as :*

$$W_q[d](D_1, D_2) = \inf_{\gamma \in \Gamma(d)(D_1,D_2)} \Big\{ \big(\sum_{(x,y)\in\gamma} d(x,y)^q \big)^{\frac{1}{q}} \Big\}.$$

We use the notation $W_q(D_1, D_2) := W_q[d_\infty](D_1, D_2)$. The degree q Wasserstein distance converges to the Bottleneck distance:

Proposition 6. *Let $D_1, D_2 \in \mathcal{D}$.*

$$\lim_{q\to\infty} W_q[d](D_1, D_2) = W_\infty[d](D_1, D_2).$$

Proof. By definition, $W_q[d](D_1, D_2) = \inf_{\gamma\in\Gamma(d)(D_1,D_2)} \{(\sum_{(x,y)\in\gamma} d(x,y)^q)^{\frac{1}{q}}\}$. For all $\gamma \in \Gamma(d)(D_1, D_2)$, we have:

$$\lim_{q\to\infty} \big(\sum_{(x,y)\in\gamma} d(x,y)^q \big)^{\frac{1}{q}} = \sup_{(x,y)\in\gamma} \{d(x,y)\}.$$

Since:

$$\lim_{q\to\infty} \inf_{\gamma\in\Gamma(d)(D_1,D_2)} \Big\{ \big(\sum_{(x,y)\in\gamma} d(x,y)^q \big)^{\frac{1}{q}} \Big\} = \inf_{\gamma\in\Gamma(d)(D_1,D_2)} \Big\{ \lim_{q\to\infty} \big(\sum_{(x,y)\in\gamma} d(x,y)^q \big)^{\frac{1}{q}} \Big\},$$

therefore:

$$\lim_{q\to\infty} \inf_{\gamma\in\Gamma(d)(D_1,D_2)} \Big\{ \big(\sum_{(x,y)\in\gamma} d(x,y)^q \big)^{\frac{1}{q}} \Big\} = \inf_{\gamma\in\Gamma(d)(D_1,D_2)} \Big\{ \sup_{(x,y)\in\gamma} \{d(x,y)\} \Big\}.$$

□

Theorem 2 (Stability theorem for the Wasserstein distance). *Let X be a triangulable metric space whose triangulations grow polynomially with constant exponent j. For any Lipschitz tame functions $f_1, f_2 : \mathbb{X} \to \mathbb{R}$ with pth persistence diagrams $D_1 = dgm(f_1)$ and $D_2 = dgm(f_2)$, $\exists C \in \mathbb{R}, k > j$ such that $\forall k \leq q$:*

$$W_q(D_1, D_2) \leq C.\|f_1 - f_2\|_\infty^{1-\frac{k}{q}}.$$

Chapter 3

Statistical Inference on Persistence Diagrams

The goal of this chapter is to introduce a statistical framework on the set of persistence diagrams $\mathcal{D}$. We define probability measures and a notion of mean on this set in Section 4.1. In Section 4.2, we introduce a notion of Reproducing Kernels on $\mathcal{D}$ which will be useful for hypothesis testing in Section 4.3. In Section 4.4, we present confidence sets on $\mathcal{D}$. In Section 4.5, we present some statistical learning methods on $\mathcal{D}$ and introduce our own perceptron model for persistence diagrams.

3.1 Fréchet Mean

The structure of $\mathcal{D}$ raises important issues to build or estimate fundamental stochastic objects, which is highlighted in this section by the Fréchet mean of persistence diagrams. A statistical approach first requires us to define probability measures on the set of persistence diagrams. We note that $\mathcal{D}$ is not complete for any of the metrics we defined so far. Therefore a classical approach in the literature is to restrict the study to a certain subset of persistence diagrams.
For any metric M on $\mathcal{D}$, we define $\mathcal{D}_M = \{D \in \mathcal{D} : M(D, \Delta) < +\infty\}$ where Δ is the diagonal line in $\mathbb{R}^2$.

3.1.1 Existence of Fréchet Means

The results in [32] show that $(\mathcal{D}_{W_q}, W_q)$ is a complete separable metric space, hence it admits probability measures. Let $B(\mathcal{D}_{W_q})$ be the Borel $\sigma-$algebra on $\mathcal{D}_{W_q}$ and let μ be a probability measure on $(\mathcal{D}_{W_q}, B(\mathcal{D}_{W_q}))$.

Definition 3.1.1 (Fréchet variance and mean). *We define the Fréchet function $F_\mu(D) : \mathcal{D}_{W_q} \to \mathbb{R}$ as $F_\mu(D) = \int_{\mathcal{D}_{W_q}} W_q(D, \delta) d\mu(\delta)$ for all $D \in \mathcal{D}_{W_q}$.*
We define the Fréchet variance as $Var_\mu = \inf\limits_{D \in \mathcal{D}_{W_q}} \{F_\mu(D)\}$.
We define the Fréchet mean as $\mathbb{E}_\mu = \{D \in \mathcal{D}_{W_q} | F_\mu(D) = Var_\mu\}$.

The following result from [32] shows the existence of a Fréchet mean.

Theorem 3.1.1 (Existence of Fréchet mean). *If the probability measure μ has a finite second moment and a compact support, then $\mathbb{E}_\mu \neq \emptyset$.*

3.1.2 Computation: A greedy approach

The approach in [43] focuses on the metric space $(\mathcal{D}_W, W)$. By equivalence of norms in finite dimensional spaces (applied to the L_∞ and L_2 induced metrics on $\mathbb{R}^2$), it is straight forward that $(\mathcal{D}_W, W)$ is also a complete separable metric space. We define the metric $W := W_2[d_2]$ where d_2 is the distance induced by the L_2 norm on $\mathbb{R}^2$. Let μ be a probability measure on $(\mathcal{D}_W, B(\mathcal{D}_W))$. An interesting case of Fréchet mean is the case of a discrete measure $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, $X_i \in \mathcal{D}_W$ as it allows us to define a notion of the barycenter of a finite collection of persistence diagrams $(X_i)_i$. The authors in [43] propose the algorithm below to find local minima of the Fréchet function on $(\mathcal{D}_W, B(\mathcal{D}_W), \mu)$ based on the Kuhn–Munkres algorithm [33].

It is shown in [43 Section 3.1] that Algorithm 1 converges to a local minimum of the Fréchet function. A limit of this algorithm emphasized in [29] is its computational cost due to the combinatorials of the pairings considered.

3.1.3 Computation: Entropic smoothed formulation

An alternative approach to the greedy Algorithm 1 is to formulate the Fréchet Mean problem in $(\mathcal{D}_W, B(\mathcal{D}_W), \mu)$ where $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, $X_i \in \mathcal{D}_W$ as an entropic smoothed optimal transport problem as developed in [29].
We start by defining some notation.

Definition 3.1.2 (Optimal transport problem). *Let $\mathcal{X}$ be a set with a cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$. Let $U_1 = (a_i)_{i \in [\![1,n_1]\!]}$ and $U_2 = (b_i)_{i \in [\![1,n_2]\!]}$ be finite subsets of $\mathcal{X}$. Let $\mu_1 = \sum_{i=1}^{n_1} \alpha_i \delta_{a_i}$ and $\mu_2 = \sum_{i=1}^{n_2} \beta_i \delta_{b_i}$ where $(\alpha_i)_{i \in [\![1,n_1]\!]}$ and $(\beta_i)_{i \in [\![1,n_2]\!]}$ are real valued positive weights such that $\sum_{i=1}^{n_1} \alpha_i = \sum_{i=1}^{n_2} \beta_i$.*

We define the cost matrix $C = (c(a_i, b_j))_{i,j}$ and the transportation polytope $\Pi(\mu_1, \mu_2) = \{P \in M_{n_1,n_2}(\mathbb{R}_+) : \forall i \sum_{j=1}^{n_2} P_{i,j} = \alpha_i, \forall j \sum_{i=1}^{n_1} P_{i,j} = \beta_j\}$. The associated optimal

Algorithm 1: Greedy algorithm for the Fréchet mean.

Data: Persistence diagrams $(X_1, \dots, X_n)$

Result: Fréchet mean Y

```
Draw $k$ uniformly in $[\![1, n]\!]$;
initialization $Y \leftarrow X_k$;
stop ← False;
/* this is a comment to tell you that we will now really start
   code                                                        */
while not stop do
    $K = card(Y)$;
    for $i = 1$ to $n$ do
        $(y^j, x^i_j) \leftarrow$ Kuhn–Munkres$(Y, X_i)$ /* compute the optimal pairings
            between $Y$ and $X_i$ with the Kuhn-Munkres algorithm.     */
        ;
    end
    for $j = 1$ to $K$ do
        $y^j \leftarrow \underset{i \in [\![1,n]\!]}{(\quad} x^i_j)$ /* Assign to each off-diagonal point of $Y$ the
            arithmetic mean of its pairings.                           */
        ;
    end
    if $Hungarian(Y, X_i) = (y^j, x^i_j)$ then
        stop ← True;
    end
end
```

transport problem is :

$$d_C(\mu_1, \mu_2) = \inf_{P \in \Pi(\mu_1, \mu_2)} \langle P, C \rangle, \tag{3.1}$$

where $\langle P, C \rangle$ denotes the standard inner product on M_{n_1,n_2}.

In the case $n_1 = n_2 =: n$ and $\forall i\, \alpha_i = \beta_i$, a result by Choquet (1956) shows $\Pi(\mu_1, \mu_2) = \{P(\sigma) \in M_{n_1,n_2}(\mathbb{R}_+) : \sigma \in \mathfrak{S}_n, \forall i, j P_{i,j} = \delta_{j,\sigma(i)}\}$ where $\mathfrak{S}_n$ is the set of permutations of $[\![1, n]\!]$. The idea developed in [29] is to introduce an entropic regularization in equation 3.1.

Definition 3.1.3 (Regularized optimal transport problem)**.** *We consider an optimal transport problem as defined in equation 3.1. Let $\gamma > 0$ be the regularization weight, and we define the entropy $h(P) = -\sum_{i,j} P_{i,j} log(P_{i,j})$ for $P \in \Pi(\mu_1, \mu_2)$. We define the regularized problem as the minimization program :*

$$\hat{d}_C^\gamma(\mu_1, \mu_2) = \inf_{P \in \Pi(\mu_1, \mu_2)} \langle P, C \rangle - \gamma h(P). \tag{3.2}$$

We call $\hat{d}_C^\gamma$ the Sinkhorn distance. Let $K = \exp(-\frac{1}{\gamma}C)$ (with term-wise exponentiation). We define the Sinkhorn map $S : M_{n_1,1} \times M_{n_2,1} \to M_{n_1,1} \times M_{n_2,1}$ such that $S(u, v) = (\frac{U_1}{Kv}, \frac{U_2}{K^T u})$ (term-wise division).

We can now get back to our initial problem, formulating the Wasserstein distance in terms of optimal transport.

Proposition 7 (Solution of the regularized problem)**.** *The equation 3.2 has a unique solution $P^* = \mathrm{diag}(u^*) K \mathrm{diag}(v^*)$ where (u^*, v^*) is a fixed point of the Sinkhorn map.*

Proposition 8 (Optimal transport formulation for the Wasserstein distance)**.** *Let $D_1, D_2 \in \mathcal{D}$ such that $D_1 = (a_i)_{i \in [\![1,n_1]\!]}$ and $D_2 = (b_i)_{i \in [\![1,n_2]\!]}$. We consider the two following measures on $\mathbb{R}^2$: $\mu_1 = \sum_{i=1}^{n_1} \delta_{a_i}$ and $\mu_2 = \sum_{i=1}^{n_2} \delta_{b_i}$.*
We define a cost matrix $C \in M_{n_1+1,n_2+1}$ between μ_1 and μ_2 by:

$$C = \left[\begin{array}{c|c} & d_2(b_1, \pi_\Delta(b_1)))^2 \\ ((d_2(a_i, b_j))^2)_{i,j} & \vdots \\ & d_2(b_{n_2}, \pi_\Delta(b_{n_2})))^2 \\ \hline (d_2(a_1, \pi_\Delta(a_1)))^2 \ldots d_2(a_{n_1}, \pi_\Delta(a_{n_1})))^2 & 0 \end{array}\right]$$

Then using (insertref) we have:

$$W(D_1, D_2) = d_C(\mu_1 + n_2\delta_\Delta, \mu_2 + n_1\delta_\Delta). \tag{3.3}$$

We approximate the solution to the optimal transport problem in (3.3) with the regularized form in (3.2) which leads to the following dual formulation [19]:

Theorem 3 (Dual formulation)**.** *The regularized formulation in 3.2 can be expressed as follows for the Wasserstein distance problem:*

$$\hat{d}_C^{\gamma}(\mu_1, \mu_2) = \max_{(v^{(1)}, v^{(2)}) \in M_{n_1,1} \times M_{n_2,1}} \langle v^{(1)}, U_1 \rangle + \langle v^{(2)}, U_2 \rangle - \gamma \sum_{i,j} \exp\Big(\frac{v_i^{(1)} + v_j^{(2)} - C_{i,j}}{\gamma} \Big). \tag{3.4}$$

Proposition 9 (Differentiability of the Sinkhorn distance)**.** *For any measure μ_2, the map $\mu_1 \mapsto \hat{d}_C^{\gamma}(\mu_1, \mu_2)$ is differentiable and its differential is $\nabla_{\mu_1} \hat{d}_C^{\gamma}(\mu_1, \mu_2) = A^*$ where (A^*, B^*) is a solution to 3.4.*

Finally, we define the barycenter of persistence diagrams and a method to compute it as in [29].

Definition 3.1.4 (Histogram)**.** *Let $(D_i)_{i \in [\![1,n]\!]}$ be persistence diagrams. We discretize $\mathbb{R}_+^2$ with a grid, and we associate to each persistence diagram D_i a probability measure μ_i which is proportional to the counting measure on each rectangle of the grid. We say that the $(\mu_i)_{i \in [\![1,n]\!]}$ are histograms.*

Definition 3.1.5 (Barycenter of a family of persistence diagrams)**.** *Let $(D_i)_{i \in [\![1,n]\!]}$ be persistence diagrams with histograms $(\mu_i)_{i \in [\![1,n]\!]}$. We note that for all histograms μ, for all $i \in [\![1,n]\!]$, the regularized optimal transport problems for the Wasserstein distance between μ and μ_i have the same cost matrix C. By differentiability of the Sinkhorn distance, The map $\mathcal{E}(\mu) = \frac{1}{n} \sum_{i=1}^{n} \hat{d}_C^{\gamma}(\mu, \mu_i)$ is differentiable and $\nabla_{\mathcal{E}}(\mu) = \frac{1}{n} \sum_{i=1}^{n} A_i^*$. We define the barycenter of $(D_i)_{i \in [\![1,n]\!]}$ as the minimum of $\mathcal{E}$.*

3.2 Distance-induced Reproducing Kernels

Reproducing kernels are a useful tool for inference and learning. In this section, we define reproducing kernels and some of the issues that appear to use them on the space of persistence diagrams $\mathcal{D}$, as well as the sliced Wasserstein kernel introduced in [9]. The algebraic structures needed on the spaces we will consider are defined in Appendix B. For the following definitions in this section, let $\mathbb{X}$ be a topological space.

Definition 3.2.1 (Kernel)**.** *We say that a map $k : \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}$ is a kernel if it is symmetric. We say that k is positive semi-definite if $\forall n \in \mathbb{N}^*$, $\forall (u_i)_{i \in [\![1,n]\!]} \in \mathbb{X}^n$, $\forall (\alpha_i)_{i \in [\![1,n]\!]} \in \mathbb{R}^n$ $\sum_{i=1}^{n} \alpha_i \alpha_j k(u_i, u_j) \geq 0$. We say that k is positive definite if the inequality is strict.*

Definition 3.2.2 (Gram matrix). *Let* $(\mathcal{H}, \langle ., . \rangle_{\mathcal{H}})$ *be a Hilbert space. Let* $n \in \mathbb{N}^*$ *and* $U = (u_i)_{i \in [\![1,n]\!]} \in \mathcal{H}^n$. *The Gram matrix* $G(U)$ *is the matrix with elements* $G(U)_{i,j} = \langle u_i, u_j \rangle_{\mathcal{H}}$, $i, j \in [\![1, n]\!]$.

For the rest of this section, let $(\mathcal{H}, \langle ., . \rangle_{\mathcal{H}})$ be a Hilbert space such that $\mathcal{H} \subseteq \mathbb{R}^{\mathbb{X}}$. For any kernel k on $\mathbb{X}$ and and $U = (u_i)_{i \in [\![1,n]\!]} \in \mathcal{H}^n$, we define the Gram matrix of k as $G_k(U) = (k(u_i, u_j))_{i,j \in [\![1,n]\!]}$.

Definition 3.2.3 (Reproducing Kernel Hilbert Space). *We say that* $\mathcal{H}$ *is a reproducing kernel Hilbert space if* $\exists \Phi : X \mapsto \mathcal{H}$ *verifying the reproducing property :*

$$\forall f \in \mathcal{H}, \quad \forall x \in \mathbb{X}, \quad f(x) = \langle f, \Phi(x) \rangle_{\mathcal{H}}.$$

$\mathcal{H}$ *is called a feature space,* Φ *a feature map and* $\Phi(x)$ *a feature vector.*
We define a reproducing kernel as the kernel

$$\begin{aligned} k \colon \mathbb{X} \times \mathbb{X} &\to \mathbb{R} \\ (x, y) &\mapsto \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}. \end{aligned}$$

We note that for all $U = (u_i)_{i \in [\![1,n]\!]} \in \mathbb{X}^n$, the matrix $(k(u_i, u_j))_{i,j \in [\![1,n]\!]}$ is the gram matrix of $V = (\Phi(u_i))_{i \in [\![1,n]\!]} \in \mathcal{H}^n$. The Moore–Aronszajn Theorem [2] gives a first characterization of kernels.

Theorem 4 (Kernel characterization). *A symmetric map* $k : \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}$ *is a reproducing kernel* $\iff$ k *is positive semi-definite.*

The interest of reproducing kernels for the present work is that they will be useful to build test statistics and to solve a loss minimization program for learning models on $\mathcal{D}$. Therefore, an important question is how to build a reproducing kernel k given $(\mathbb{X}, \mathcal{H})$.

Definition 3.2.4 (Conditionally negative semidefinite map). *We say that a symmetric map* $d : \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}_+$ *is conditionally negative semidefinite (c.n.s.d.) if* $\forall n \in \mathbb{N}^*$, $\forall (u_i)_{i \in [\![1,n]\!]} \in X^n$, $\forall (\alpha_i)_{i \in [\![1,n]\!]} \in \mathbb{R}^n$ *such that* $\sum_{i=1}^n \alpha_i = 0$ *we have* $\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j d(u_i, u_j) \leq 0$.

The Kimeldorf–Wahba Theorem [26] shows how to build reproducing kernels from c.n.s.d distances.

Theorem 5 (Kernels from conditionally negative semidefinite maps). *Let $d : \mathbb{X}\times\mathbb{X} \mapsto \mathbb{R}_+$ be conditionally negative semidefinite. Then $\forall\sigma > 0$, the map*

$$\begin{aligned} k\colon \mathbb{X}\times\mathbb{X} &\to \mathbb{R} \\ (u,v) &\mapsto \exp\Big(-\frac{d(u,v)}{2\sigma^2}\Big) \end{aligned}$$

is positive definite.

We now specify $\mathbb{X} = \mathcal{D}$. We note that the Wasserstein distance is not a conditionally negative semidefinite map, as shown numerically in [38 Appendix A]. However, [9] proposes a map on $\mathcal{D}$, the sliced Wasserstein distance that is c.n.s.d .

Definition 3.2.5 (Sliced Wasserstein metric). *For all measure μ, ν on $\mathbb{R}$ such that $\mu(\mathbb{R}) = \nu(\mathbb{R}) < +\infty$, let $\mathcal{W}(\mu,\nu) = \inf\limits_{P\in\Pi(\mu,\nu)} \iint_{\mathbb{R}\times\mathbb{R}} |x-y|\, P(dx,dy))$ where $\Pi(\mu,\nu)$ is the set of probability measures on $\mathbb{R}$ with marginals μ and ν.*
For all $\theta\in\mathbb{S}_1$ the radius 1 sphere in $(\mathbb{R}^2, \|\,.\,\|_2)$, we define $L(\theta) = \{\lambda\theta : \lambda\in\mathbb{R}\}$ and $\pi_\theta : \mathbb{R}^2 \to L(\theta)$ the orthogonal projection on $L(\theta)$. Let $D_1, D_2 \in \mathcal{D}$, we consider the measures $\mu_1^\theta = \sum_{x\in D_1}\delta_{\pi_\theta(x)}$ and $\mu_{1\Delta}^\theta = \sum_{x\in D_1}\delta_{\pi_\theta\circ\pi_\Delta(x)}$ and similarly for μ_2^θ and $\mu_{2\Delta}^\theta$.
We define the Sliced Wasserstein metric as :

$$\mathcal{SW}(D_1,D_2) = \frac{1}{2\pi}\int_{\mathbb{S}_1} \mathcal{W}(\mu_1^\theta + \mu_{2\Delta}^\theta, \mu_2^\theta + \mu_{1\Delta}^\theta)d\theta$$

The sliced Wasserstein kernel k_{SW} is defined as the Gaussian Kernel induced by $\mathcal{SW}$.

Proposition 3.2.1. *The sliced Wasserstein metric is c.n.s.d .*

Definition 3.2.6 (Sliced Wasserstein kernel). *The sliced Wasserstein kernel k_{SW} is defined as the Gaussian Kernel induced by $\mathcal{SW}$ from Theorem 5:*

$$\begin{aligned} k_{SW}\colon \mathcal{D}\times\mathcal{D} &\to \mathbb{R} \\ (u,v) &\mapsto \exp\Big(-\frac{d(u,v)}{2\sigma^2}\Big) \end{aligned}$$

Other notable Kernels used in the literature are:

- The persistence weighted gaussian kernel k_{PWGK} [28].
- The Persistence Scale Space kernel k_{PSS} [38].

3.3 Hypothesis Testing

The objective of this section is to introduce some methods to build statistics for null hypothesis significance testing on $\mathcal{D}$.

3.3.1 Two-sample test

Let $\mu^{(1)}, \mu^{(2)}$ be two probability measures on $\mathcal{D}$. We consider two families of persistence diagrams $D^{(1)} = (D_i^{(1)})_{i \in [\![1,n_1]\!]}$ and $D^{(2)} = (D_i^{(2)})_{i \in [\![1,n_2]\!]}$ such that $\forall m \in \{1,2\}, (D_1^{(m)}, \dots, D_{n_m}^{(m)}) \underset{i.i.d.}{\sim} \mu^{(m)}$. We consider the test with null hypothesis:

$$H_0 : \mu^{(1)} = \mu^{(2)}$$

A first approach for two-sample tests on persistence diagrams was proposed in [39] using a permutation test. Let T be a test statistic. We define the set of relabellings of the observations $\mathcal{P}(D^{(1)}, D^{(2)}) = \{(E^{(1)}, E^{(2)}) : E^{(1)} \cap E^{(2)} = \emptyset, \forall m \in \{1,2\} \quad E^{(m)} \subset D^{(1)} \cup D^{(2)}, card(E^{(m)}) = n_m\}$. Under H_0, all those labellings have the same probability. The permutation p-value of the test is defined as

$$p = \frac{card(\{(E^{(1)}, E^{(2)}) \in \mathcal{P}(D^{(1)}, D^{(2)}) : T((D^{(1)}, D^{(2)})) \geq T((E^{(1)}, E^{(2)}))\})}{card(\mathcal{P}(D^{(1)}, D^{(2)}))} \tag{3.5}$$

From a computational standpoint, we can estimate p by Monte-Carlo simulation in case $card(\mathcal{P}(D^{(1)}, D^{(2)}) = \binom{n_1+n_2}{n_1}$ is too large. Let $\hat{p}_N$ be this estimate for N Monte-Carlo iterations.

Theorem 3.3.1 (Permutation p-value). *$\hat{p}_N$ is a p-value.*

An extension of this permutation test to m families of persistence diagrams (where $m \geq 3$) is proposed in [11]: the authors use the same statistic and reshuffle the labellings of the m families for a Monte-Carlo estimate of the p-value.

Another approach developed in [42] relies on the RKHS theory. The test statistic here is an empirical estimate of the Maximum Mean Discrepancy (MMD).

Definition 3.3.1 (Maximum Mean Discrepancy). *Let $(\mathbb{X}, d_{\mathbb{X}})$ be a metric space and let $\mathcal{H} = \mathbb{R}^{\mathbb{X}}$. For all Borel probability measures P, Q on $\mathbb{X}$, let X, Y be random variables on $\mathbb{X}$ such that $X \sim P$ and $Y \sim Q$. we define the Maximum Mean Discrepancy (MMD) by:*

$$\gamma(D^{(1)}, D^{(2)}) = \sup_{f \in \mathcal{H}} \big(\mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{Y \sim Q}[f(Y)] \big) \tag{3.6}$$

Proposition 3.3.1 (Biased Estimator of the MMD). *Let k be a reproducing kernel on $\mathcal{D}$. A biased estimator of the MMD between the measures induced by $D^{(1)}$ and $D^{(2)}$*

is given by:

$$\begin{aligned}\hat{\gamma}(D^{(1)},D^{(2)}) = \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}k(D_i^{(1)},D_j^{(1)}) + \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{i=1}^{n_2}k(D_i^{(2)},D_j^{(2)}) \\ - \frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}k(D_i^{(1)},D_j^{(2)})\end{aligned} \tag{3.7}$$

Proposition 3.3.2 (Asymptotic distribution of empirical MMD)**.** *Under H_0, the statistic in 3.7 converges to a weighted sum of chi-squares.*

3.3.2 Independence test.

Let μ be a probability measure on $\mathcal{D} \times \mathcal{D}$ with marginals $\mu^{(1)}, \mu^{(2)}$. We consider a family of pairs of persistence diagrams $Z = (Z_i)_{i \in [\![1,n]\!]}$ where $Z_i = (D_i^{(1)}, D_i^{(2)})$, $i \in [\![1,n]\!]$ such that $(Z_1,\dots,Z_n) \underset{i.i.d.}{\sim} \mu$. The following results are based on [42].

Definition 3.3.2 (Hilbert-Schmidt Independence Criterion)**.** *Let $(\mathbb{X}, d_{\mathbb{X}})$ and $(\mathbb{Y}, d_{\mathbb{Y}})$ be metric spaces. For all Borel probability measures P, Q on $\mathbb{X}$ and $\mathbb{Y}$, let X, Y be random variables on $\mathbb{X}$ such that $X \sim P$ and $Y \sim Q$. we define the Hilbert-Schmidt Independence Criterion as:*

$$\gamma(D^{(1)},D^{(2)}) = \sup_{f\in\mathcal{H}}\big(\mathbb{E}_{X\sim P}[\,f(X)] - \mathbb{E}_{Y\sim Q}[\,f(Y)]\,\big) \tag{3.8}$$

Definition 3.3.3 (Biased Estimator of the HSIC)**.** *Let k_1, k_2 be reproducing kernels on $\mathcal{D}$. We define the matrices $K_1 = (k_1(D_i^{(1)}, D_j^{(1)}))_{i,j}$, $K_1 = (k_1(D_i^{(2)}, D_j^{(2)}))_{i,j}$ and $H = (\delta i, j - \frac{1}{n})_{i,j}$. A biased estimator of the HSIC between the measures induced by $D^{(1)}$ and $D^{(2)}$ is given by:*

$$T_{k_1,k_2}(Z) = \frac{1}{n}Tr(K_1 H K_2 H) \tag{3.9}$$

Proposition 3.3.3 (Assymptotic distribution of empirical HSIC)**.** *Under H_0, the statistic in 3.9 converges to a weighted sum of chi-squares.*

3.4 Confidence Sets

We consider a d-dimensional manifold $\mathbb{M}$ with a metric d. Let μ be a probability measure with compact support $X_\mu \in \mathbb{M}$. We observe a point cloud $X_N = (x_i^{(N)})_{i\in[\![1,N]\!]}$ consisting of N points of X_μ sampled from μ. We consider $D_N = dgm(X_N)$ and $D = dgm(X_\mu)$. In applications of persistent homology, it is important to answer the following questions:

1. How to quantify the accuracy of D_N as an estimator of D ?
2. How to distinguish noise from signal in our observations ?

Definition 3.4.1 (Confidence set). *For a given confidence level* $\alpha \in [\,0,1]$ *, we define a* $1-\alpha$ *confidence set for* X_μ *as an interval* $[0, c_N]$ *such that* $\lim_{N\to\infty} \mathbb{P}(W_\infty[d](D_N, D) > c_N) < \alpha$.

Finding a confidence set for X_μ provides answers to both questions 1 and 2, as we then consider the points in the set $\{(a,b) \in \mathbb{R}^2_+ : d((a,b),\Delta) = c_N\}$ where $\Delta = \{(a,a) \in \mathbb{R}^2_+\}$ as noise. In this section, we follow [23] to find these confidence sets.

The starting point proposed in [23] is to note that $W_\infty(D_N, D) \leq W_H(X_N, X_\mu)$ by Theorem 1. Therefore, if c_N is such that $\lim_{N\to\infty} \mathbb{P}((W_H[d](X_N, X_\mu) > c_N) < \alpha$ then:

$$\lim_{N\to\infty} \mathbb{P}(W_\infty(D_N, D) > c_N) < \lim_{N\to\infty} \mathbb{P}(W_H(X_N, X_\mu) > c_N) < \alpha$$

All the methods in [23] find an upper bound for $W_H(X_N, X_\mu)$ which is used to bound $\mathbb{P}(W_\infty(D_N, D) > c_N)$.

3.4.1 Subsampling

Assume we observe X_N. We choose an integer $b_N \leq N$ and consider the $p_N = \binom{N}{b_N}$ subsets of X_N of cardinality b_N, denoted $(Y_i^N)_{i\in[\![1,p_N]\!]}$. Let $T_i = W_H(X_N, Y_i^N), i \in [\![1, p_N]\!]$ and the empirical complementary cumulative distribution function $L_b(r) = \frac{1}{N}\sum_{i=1}^N \mathbb{1}(T_j > r)$.

Theorem 6 (Subsampling confidence intervals). *Let* $c_{b_N} = 2L_{b_N}^{-1}(\alpha)$.

$$\mathbb{P}(W_\infty(D_N, D) > c_{b_N}) \leq \alpha + \mathcal{O}(\frac{b_N}{N}) \tag{3.10}$$

3.4.2 Concentration of measure

An alternative approach is based on the study of the function:

$$\begin{aligned} \rho\colon X_\mu \times \mathbb{R}_+ &\to \mathbb{R}_+ \\ (x,r) &\mapsto \frac{\mu(B(x,\frac{r}{2}))}{r^d} \end{aligned}$$

We assume that $\forall x \in X_\mu$, $r \mapsto \rho(x,r)$ is continuous and bounded. We define the following functions:

- $\rho_R : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\rho_R(r) = \inf\limits_{x \in X_\mu} \rho(x, r)$
- $\rho_X : X_\mu \to \mathbb{R}_+$ such that $\rho_X(x) = \lim_{r \to 0^+} \rho(x, r)$
- $\overline{\rho} = \lim_{r \to 0^+} \rho_R(r)$

We assume that $\exists r_0 > 0$, $\gamma_1 \in \mathbb{R}, \gamma_2 \in \mathbb{R}$ such that :

$$\sup_{x \in X_\mu} \sup_{r \in [\,0, r_0]} |\frac{\partial \rho(x, r)}{\partial r}| \leq \gamma_1, \tag{3.11}$$

$$\sup_{r \in [\,0, r_0]} |\rho'_R(r)| \leq \gamma_2. \tag{3.12}$$

Let $\mu_N(A) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}_A(x_i^{(N)})$ be the empirical measure. For $r_N = \Omega((\frac{\log(N)}{N})^{\frac{1}{d+2}})$ we define:

$$\overline{\rho}_N = \min_{i \in [\![1, N]\!]} \frac{\mu_N(B(x_i^{(N)}, \frac{r_N}{2}))}{r_N^d}.$$

We define $c_N(\alpha)$ as the solution of:

$$\frac{2^{d+1}}{c_N^d(\alpha) \overline{\rho}_N} \exp\Big(- N \frac{c_N^d(\alpha) \overline{\rho}_N}{2} \Big) = \alpha. \tag{3.13}$$

The second method proposed in [23] to derive a confidence set is to randomly split the observations in two families of the same cardinality (assuming N is even) denoted by $X_N = X_N^{(1)} \cup X_N^{(2)}$. From the subset (1), we derive $c_N^{(1)}(\alpha)$ as in Equation 3.13. We then use $c_N^{(1)}(\alpha)$ to construct a confidence set for $X_N^{(2)}$.

Theorem 7 (Concentration of measure confidence intervals)**.**

$$\mathbb{P}\big(W_\infty(D_N^{(2)}, D) > c_N^{(1)}(\alpha)\big) \leq \alpha + \mathcal{O}\Big(\frac{\log(N)}{N}\Big)^{\frac{1}{2+d}}. \tag{3.14}$$

A variant of this method is to introduce a kernel estimator. Let $G : \mathbb{R}_+ \to \mathbb{R}_+$ such that $G(x) = \mathbb{P}(\rho_X(X) \leq x)$ and $g := G'$. We define $\hat{\rho}(x, r_N) = \frac{\mu_N(B(x, \frac{r_N}{2}))}{r_N^d}$ and $V_i = \hat{\rho}(X_i, r_n)$, $\forall i \in [\![1, N]\!]$. Let $r_N = (\frac{\log(N)}{N})^{\frac{1}{2+d}}$ and $b_N = \Omega(r_N^{\frac{1}{4}})$. We estimate g using a kernel k:

$$\hat{g}_N(x) = \frac{1}{N} \sum_{i=1}^N \frac{1}{b_N} K\Big(\frac{v - V_i}{b}\Big).$$

We define $c_N(\alpha)$ as the solution of:

$$\frac{2^{d+1}}{c_N(\alpha)^d} \int_{\overline{\rho}_N}^\infty \frac{\hat{g}_N(x)}{x} \exp\Big(- \frac{N x c_N(\alpha)^d}{2} \Big) dx = \alpha. \tag{3.15}$$

As before we randomly split the observations in two families. From the subset (1), we derive $c_N^{(1)}(\alpha)$ as in equation 3.15. We then use $c_N^{(1)}(\alpha)$ to construct a confidence set for $X_N^{(2)}$.

Theorem 8 (Kernel estimation confidence intervals)**.**

$$\mathbb{P}\big(W_\infty(D_N^{(2)}, D) > c_N^{(1)}(\alpha)\big) \leq \alpha + \mathcal{O}\Big(\frac{\log(N)}{N}\Big)^{\frac{1}{2+d}}. \tag{3.16}$$

3.5 Statistical Learning

3.5.1 Kernel based learning

We observe a finite collection of persistence diagrams $(D_i)_{i\in[\![1,n]\!]}$ associated to real valued outputs $(y_i)_{i\in[\![1,n]\!]}$. We want to find a function $f^* : \mathcal{D} \mapsto \mathbb{R}$ that approximates the observed map with respect to some loss metric. This problem is formalized as follow:

$$f^* = \underset{f\in\mathcal{F}}{\operatorname{argmin}}\Big\{\sum_{i=1}^{n} l(f(x_i), y_i) + \phi(\|f\|)\Big\}, \tag{3.17}$$

where $\mathcal{F} \subset \mathbb{R}^{\mathcal{D}}$ is called the class of predictors, $\phi : \mathbb{R} \mapsto \mathbb{R}$ is the regularization function, and $l : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ is the loss function. The choice of a loss function depends on the learning technique, classic ones include:

- Least squares : $l(f(x), y) = (y - f(x))^2$,
- Logistic regression : $l(f(x), y) = \log(1 + \exp(-yf(x)))$,
- Support Vector Machines : $l(f(x), y) = (1 - yf(x))_+$.

The result by Schôlkopf et al. in [41] allows us to solve Equation 3.17 in a much easier way.

Theorem 9 (Representer Theorem)**.** *Let $\mathcal{H}$ be a Reproducing Kernel Hilbert Space and k the associated reproducing kernel. Then:*

$$f^*\textit{is a solution to Equation } 3.17 \iff \exists(\alpha_i)_{i\in[\![1,n]\!]} \in \mathbb{R}^n \textit{ such that } f^* = \sum_{i=1}^{n} \alpha_i k(D_i, .).$$

3.5.2 Vectorization method

Another approach in the literature is to embed persistence diagrams into vectors of fixed length and use these as inputs for standard learning models [38,1]. Assume $\exists (B_{min}, B_{max}) \in \mathbb{R}_+^2$ such that $(D_i)_{i\in[\![1,n]\!]} \subset ([\,B_{min}, B_{max}] \times [\,B_{min}, B_{max}]\,)^n$. We consider a discretization of $[\,B_{min}, B_{max}] \times [\,B_{min}, B_{max}]\,$. The objective is to encode the information of a given persistence diagram in a histogram derived from the grid: we build a vector by evaluating a given function on the discretization steps and then sorting these values in a chosen order. We illustrate this with the approach in [1] that is based on persistence images.

For a persistence diagram $D = \{(a_i, b_i) : i \in [\![1, n]\!]\}$, we define the persistent image $\rho : \mathbb{R}^2 \to \mathbb{R}$ as:

$$\rho(x, y) = \sum_{i=1}^{n} w(a_i, b_i) \exp(-\frac{(a_i - x)^2 + (b_i - y)^2}{2\sigma^2}),$$

where $\sigma > 0$ and $w : \mathbb{R}^2 \to \mathbb{R}$ a weight function.

3.5.3 A perceptron on $\mathcal{D}^p$

Suppose a finite collection of vectors of persistence diagrams $(X^{(i)})_{i\in[\![1,n]\!]}$ associated to real valued outputs $(y_i)_{i\in[\![1,n]\!]}$. For all $i \in [\![1, n]\!]$, $X^{(i)} = (D_j^i)_{j\in[\![1,p]\!]}$. We want to find a function $f^* : \mathcal{D}^n \mapsto \mathbb{R}$ that approximates the observed map with respect to some loss metric. To do so, we develop a model by analogy with a perceptron [40]. The weighted sum operation in a classical MLP is replaced in our model by the weighted barycenter map developed to compute the Fréchet Mean in Section 3.1. The model is trained by solving a minimization program similar to Equation 3.17, which is made possible by the differentiability of the Sinkhorn distance from Proposition 9.

In particular, the two following questions look of interest:

1. How does this model compare to vectorizing the persistence diagrams and embedding them into a classical Multi Layer Perceptron ?

2. Is the Fréchet mean a good proxy for the addition operator on $\mathcal{D}$?

Definition 3.5.1 (Persistence Diagram Perceptron)**.** *We define a persistence diagram perceptron as a map:*

$$\begin{aligned} f \colon \mathcal{D}^p &\to \mathbb{R} \\ (D_j)_{j\in[\![1,p]\!]} &\mapsto g\big(\mathrm{Bar}\big((\omega_j, D_j)_{j\in[\![1,p]\!]}\big)\big) \end{aligned}$$

Where $(\omega_j)_{j\in[\![1,p]\!]} \subset \mathbb{R}^p$ *and* $g : \mathcal{D} \to \mathbb{R}$.

The function g is fixed a priori. The learning weights $(\omega_j)_{j\in[\![1,p]\!]}$ will be updated during training. To approximate the observed maps with respect to a loss metric l and a regularization term ϕ, the learning weights should solve for:

$$(\omega_j)^*_{j\in[\![1,p]\!]} = \underset{(\omega_j)_{j\in[\![1,p]\!]}\subset\mathbb{R}^p}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} l(g\big(Bar\big((\omega_j, D^i_j)_{j\in[\![1,p]\!]}\big)\big), y_i) + \phi(\|f\|) \right\} \tag{3.18}$$

To train the perceptron, we first consider the general case where g is differentiable. We solve the program 3.18 using the gradient descent method:

$$\Delta\omega_j = -\eta \frac{\partial g\big(Bar\big((\omega_j, D^i_j)_{j\in[\![1,p]\!]}\big)\big)}{\partial \omega_j}$$

$$\omega_j \leftarrow \omega_j + \Delta\omega_j$$

But it is not straightforward to make a suitable choice of function $g : \mathcal{D} \to \mathbb{R}$.

Chapter 4

Persistent Homology of Order Flows

4.1 Dataset

4.1.1 Stocks and period

We consider the limit order books between 3 July 2017 and 1 December 2017 of seven stocks traded on NASDAQ : AAPL (Apple Inc.), CSCO (Cisco Systems, Inc.), FB (Facebook, Inc.), GOOG (Alphabet Inc.), INTC (Intel Corporation), MSFT (Microsoft Corporation) and NVDA (NVIDIA). On the Nasdaq website, they are all listed in the same industry sector (”Technology”) and in the same market capitalization class (”mega caps”, except NVDA which is listed within ”large caps”).

For the choice of the period considered, we tried to avoid important economic events that may influence financial markets too much. We note that among these companies, there were two Mergers and Acquisitions deals over $ 1 billion in the considered period: Alphabet bought HTC on 21 September 2017 (for $ 1.1 billion) and Cisco bought BroadSoft on 25 October 2017 (for $ 1.9 billion). Most importantly, NASDAQ was closed on 4 July 2017, 4 September 2017 and 23 November 2017. The day before each of these days, there is an early market closure at 1:00 p.m EST. We exclude the three market closure days from our study as well the three early market closure days. Therefore, we consider 106 trading days overall.

4.1.2 Limit order book data

The limit order books were obtained via LOBSTER and consist of the message .csv file for every day and every stock considered. This file records all the events that occur on the LOB during the normal trading session of NASDAQ. This trading session lasts from 9:30 a.m EST to 4:00 p.m. EST. There exists 7 types of events on the order book : the submission of a new limit order, the partial deletion of a limit order, the total deletion of a limit order, the execution of a visible limit order, the execution of a hidden limit order, cross trades and an additional event class which is a trading halt indicator. In the present work, we focus solely on three events: the submission of a new limit order, and the execution of a visible or a hidden limit order. We only select the orders posted at the first 5 levels in the book. For all events, we get the following information: at which time it has been received (in seconds, with nanosecond decimal precision), the event's unique order ID, its size, its price and its direction. Overall, this represents 52.9 GB of data (177 378 008 admissible orders). Our main objective is to understand whether certain types of orders trigger other types of orders.

4.2 Model

Let $N_{days} = 106$ be the number of trading days considered. We consider 4 order types : limit buy (lb), limit sell (ls), market buy (mb) and market sell (ms). We denote by $O = (O_i)_{i \in [\![1,4]\!]} = \{lb, ls, mb, ms\}$ the set of order types, and by $S = (S_i)_{i \in [\![1,7]\!]}$ the set of stock names. Let $X = (S_i, O_j)_{i,j}$ be the set of stock-order type combinations. Let $\epsilon > 0$ and $f : \mathbb{R} \to \mathbb{R}$ a continuous map.

There is $T_{day} = 23400s$ in every considered trading session. Let $T_{total} = T_{day} \times N_{days} = 2480400s$ be the total number of seconds in the considered trading sessions.We denote the set of finite unions of closed intervals in $[\,0, T_{total}]$ by:

$$\mathcal{T} = \{\cup_{i=1}^{n}[\,a_i, b_i] \,:\, n \in \mathbb{N}^*, [\,a_i, b_i] \subset [\,0, T_{total}]\ \forall i \in [\![1, n]\!], [\,a_i, b_i] \cap [\,a_j, b_j] = \emptyset\, \forall i \neq j\}.$$

For all $T \in \mathcal{T}$ where $T = \cup_{i=1}^{n}[\,a_i, b_i]$, we define $length(T) = \sum_{i=1}^{n}(b_i - a_i)$. Assume in this section that we define the time at which a given order is received as the cumulative active trading time in seconds since the opening of the first trading session considered in this work. Hence, the times at which orders are received lie between 0 and T_{total}.

For each time period $T \in \mathcal{T}$, we construct $d[\,\epsilon, f, T]$ a symmetric map on $X \times X$ as follows:

- We decompose T as the union of $n_{\epsilon,T} = \frac{length(T)}{\epsilon}$ disjoint time intervals of length ϵ. This decomposition is of course unique and we write $T = \cup_{h=1}^{n_{\epsilon,T}} T_h$.

- For all integers $h \in [\![1, n_{\epsilon,T}]\!]$, let $Y_h^{i,j}$ (respectively $Y_h^{k,l}$) be the number of orders of type O_j (respectively O_l) on stock S_i (respectively S_j) received during T_h.

- Let $\rho_{ij,kl}$ be the empirical correlation between $(Y_h^{i,j})_{h \in [\![1, n_{\epsilon,T}]\!]}$ and $(Y_h^{k,l})_{h \in [\![1, n_{\epsilon,T}]\!]}$.

- We define $d[\,\epsilon, f, T]((S_i, O_j), (S_k, O_l)) := f(\rho_{ij,kl})$.

We now define the symmetric maps we will be using and the notations for the associated spaces:

- $d_{\mathbb{X}}[\,\epsilon, f]\ := d\big[\,\epsilon, f, [\,0, T_{total}]\,\big]$ and the associated space $\mathbb{X} := \Big(X, f\big(d_{\mathbb{X}}[\,\epsilon, f]\,\big)\Big)$. This case corresponds to computing correlations over the whole period considered.

- For all integers $i \in [\![1, N_{days}]\!]$, define:

  $$d_i[\,\epsilon, f]\ := d\Big[\,\epsilon, f, \big[\,(i-1) \times T_{day}, (i) \times T_{day}\big]\,\Big],$$

  and the associated space:

  $$\mathbb{X}_i := \Big(X, f\big(d_{\mathbb{X}_i}[\,\epsilon, f]\,\big)\Big).$$

  This case corresponds to N_{days} spaces where we compute correlations over one day for each one.

- For all integers $i \in [\![1, N_{days}]\!]$, define:

  $$d_{\mathbb{H}_i^{opening}}[\,\epsilon, f]\ := d\Big[\,\epsilon, f, \big[\,(i-1) \times T_{day}, (i-1) \times T_{day} + 3600\big]\,\Big],$$

  and the associated space:

  $$\mathbb{H}_i^{opening} = \Big(X, f\big(d_{\mathbb{H}_i^{opening}}[\,\epsilon, f]\,\big)\Big).$$

  This case corresponds to N_{days} spaces where we compute correlations over the first hour of the day for each one.

- For all integers $i \in [\![1, N_{days}]\!]$, define:

  $$d_{\mathbb{H}_i^{closure}}[\,\epsilon, f]\ := d\Big[\,\epsilon, f, \big[\,(i) \times T_{day} - 3600, (i) \times T_{day}\big]\,\Big],$$

  and the associated space:

  $$\mathbb{H}_i^{closure} = \Big(X, f\big(d_{\mathbb{H}_i^{closing}}[\,\epsilon, f]\,\big)\Big).$$

  This case corresponds to N_{days} spaces where we compute correlations over the last hour of the day for each one.

- For all integers $i \in [\![1, N_{days}]\!]$, define:

$$d_{\mathbb{H}_i^{middle}}[\,\epsilon, f] \,:= d\Big[\,\epsilon, f, \big[\,(i-1)\times T_{day} + 3600, (i)\times T_{day} - 3600\big]\,\Big],$$

  and the associated space:

$$\mathbb{H}_i^{middle} = \Big(X, f\big(d_{\mathbb{H}_i^{middle}}[\,\epsilon, f]\,\big)\Big).$$

  This case corresponds to N_{days} spaces where we compute correlations between the first and last hour of the day for each one.

Given a point cloud, the construction of the Vietoris–Rips filtration in Section 2.2 doesn't require a metric space structure and is applicable to any space E with a dissimilarity map [36] $d_E : E \times E \to \mathbb{R}_+$, i.e. a symmetric map such that for all $x \in E$, $d_E(x,x) = 0$. To make the symmetric maps we defined above dissimilarity maps between elements of X, we consider functions f of the form $f(x) = g((1-x^2))$ where $g : [\,0,1] \to [\,0,1]$ is a continuous increasing function. The role of g is to rescale the dissimilarity matrices because of the small difference between many of their coefficients. Empirically, choosing $g(x) = x^{10}$ seems reasonable for our dataset.

4.3 Results

We wrote our code in Python. Apart from standard Python libraries, the scripts we used but didn't write are the C++ code Ripser [4] (see Section 2.2) for the computation of homology which we execute from Python using the subprocess library and the manifold library from sklearn to apply multi-dimensional scaling to a point cloud.

Given $\mathbb{E} = (E, d_E)$ a finite set with a dissimilarity map, we denote by $Rips(\mathbb{E})$ the filtered simplicial complex obtained with a sequence of nested Vietoris-Rips simplicial complexes with set of vertices E.

The computation of the homology of $Rips(\mathbb{X})$ leads to the results summarized in Table 4.1. We see that there exists persistence intervals up to degree 3 homology. We want to investigate, for each homology degree $p \leq 3$, what are the generators of the pth homology groups for a specific value of the filtration parameter. This will tell us which stock-order type combinations trigger each other and for how long. A natural descriptor to look at for this information is the barcode of $Rips(\mathbb{X})$ (Figure 4.1).

We want to see, for a given value of the filtration parameter r, which vertices of $Rips(\mathbb{X})$ are connected i.e. which ones have pairwise distance (pariwise dissimilarity) less than r. We use the barcode to determine which values of r might be of interest.

Homology degree	Number of persistence intervals
0	28
1	18
2	12
3	3
4 and more	0

Table 4.1: Number of persistence intervals of $Rips(\mathbb{X})$

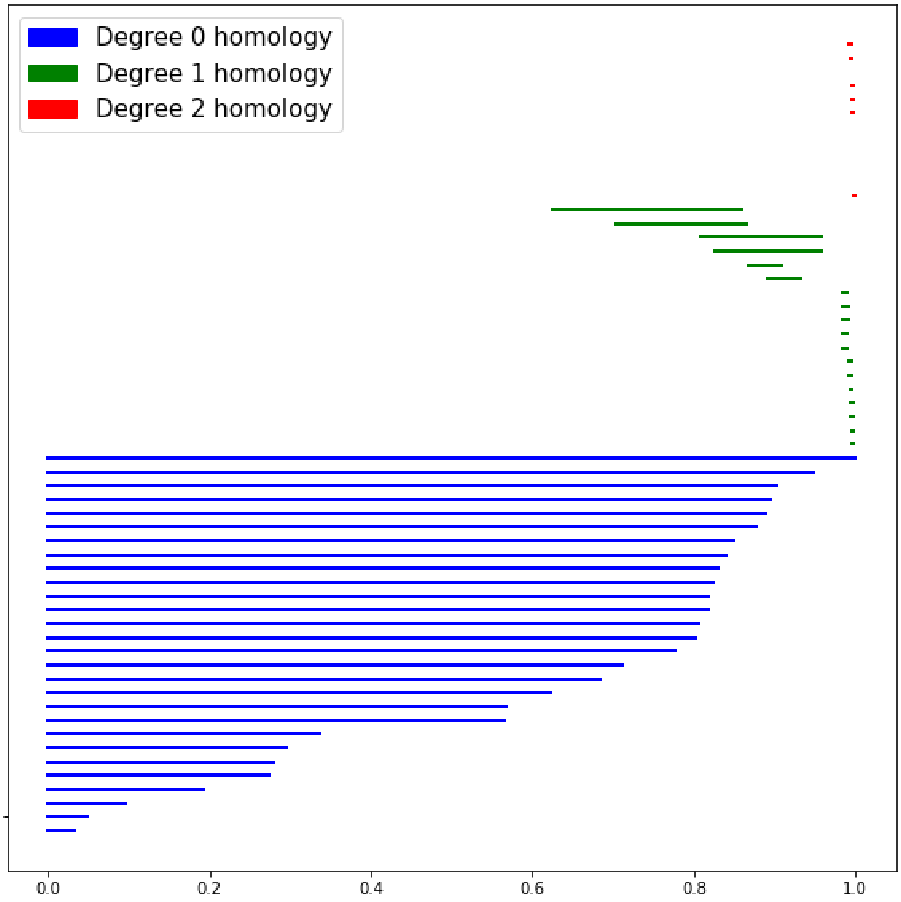

Figure 4.1: Barcode of $Rips(\mathbb{X})$ for degree 0, 1 and 2 homology.

4.3.1 Edges

As suggested by the barcode, there are very few connections for small values of r: the first three edges appearing are $\{(AAPL, lb), (AAPL, ls)\}$ $(r = 0.03389)$, $\{(NVDA, lb), (NVDA, ls)\}$ $(r = 0.0486783)$, and $\{(FB, lb), (FB, ls)\}$ $(r = 0.0972022)$. They involve exclusively pairs of limit orders on the same stock. The next two edges appearing are the pairs of limit orders of CSCO then MSFT. It is interesting to note that the first point of market order type to be connected to any other point appears for $r = 0.279402 : \{(NVDA, mb), (NVDA, ls)\}$. On the stock NVDA, the pair of limit orders was already an edge, which means that for $r \geq 0.279402$ the sell limit order on NVDA is connected to both order types on the buy direction for NVDA.

The consecutive death events at $r = 0.565898$ then $r = 0.568468$ are particularly interesting. The death at $r = 0.565898$ corresponds to the first connection between two points related to different stocks: $\{(MSFT, ls), (AAPL, ls)\}$. We not that both points are related to limit sells. The death at $r = 0.568468$ corresponds to the second point with market order type to be part of an edge: it is again a NVDA point, in $\{(NVDA, ms), (NVDA, lb)\}$. NVDA points are involved in three edges: one between limit orders, and the two market/limit combinations involving orders in opposite directions.

We can make two important remarks so far for $r \leq 0.57$:

- limit orders are more connected than market orders: this comes from the fact that they are more correlated than market orders on this data set. We note the particular behavior of points related to NVDA;

- Connections between points related to the same stock happen more often than connections between points from different stocks.

4.3.2 Higher dimensional simplices

The first triangle appears for $r \simeq 0.5977$. It is interesting to note that this does not correspond to any birth/death event of any degree on Ripser's output: this could come from a numerical error of the software. Indeed, Ripser indicates a birth event in the degree 1 homology for $r = 0.624653$. This triangle corresponds to three NVDA points: both limit orders and buy market order.

The second triangle appears for $r \simeq 0.648$. It corresponds to the limit sells of AAPL, FB and MSFT. We recall that AAPL and MSFT limit sells were the first edge with points from different stocks. The third triangle appears for $r = 0.703819$. This corresponds to a birth event in degree 1 homology. It is interesting to note that this triangle is formed by the limits buys of (again) AAPL, FB and MSFT. There are no edges involving any couple of points from different stocks that do not belong to one of these triangles. The fourth triangle corresponds to GOOG points ($r = 0.808748$): like the first triangle (NVDA points), it is composed by both limit orders and the buy market order. There is no market sell order in any of the triangles constructed so far. Like the third triangle, it corresponds to a birth event in degree 1 homology. There are two more triangles for $r \simeq 0.86$ MSFT limit buy and sell and AAPL limit sell and AAPL limit buy and sell and MSFT limit buy. We note that thee are Very few market sells in higher dimensional simplices and there is a concentration of higher dimensional simplices near 1.

4.3.3 Visualization

To represent visually these connections, we would like to embed our observations in $\mathbb{R}^2$ by multi-dimensional scaling. Even though the coefficients of the dissimilarity matrix we consider here satisfy the identity of indiscernibles, some of them violate the triangle inequality. Since the connections would still be plotted according to the initial dissimilarity matrix, this might not be visually convenient because we could see close points on the diagram without connection, and far points with connection. To avoid this inconvenience, we transform the dissimilarity matrix using Djikstra's Algorithm, then apply multi-dimensional scaling to this new matrix. The plots we obtained can be found in Appendix C.

To analyze the results above, it is important to look at the connection diagrams as dynamic objects with respect to the filtration parameter. in fact, we note on the persistence diagram of $Rips(\mathbb{X})$ (Figure 4.2) a concentration of some of the degree 1 homology points nearby the diagonal, but it is not relevant to compute confidence sets in this case given the small number of points.

4.4 Model Robustness

Stability results for persistence diagrams in Chapter 2 show that for some functional classes, the divergence of persistence diagrams after a small perturbation is well controlled. In order to assess the robustness of our model, we investigate the two following problems:

- How sensitive is the model to the definition of a dissimilarity metric ?
- What can we infer about the persistence of $\mathbb{X}_i$ from the persistence of $\mathbb{X}$?

4.4.1 Sensitivity with respect to the dissimilarity measure

We have to ensure that our topological descriptors are not too sensitive to the definition of our dissimilarity map i.e. to the parameter ϵ, otherwise the choice of a

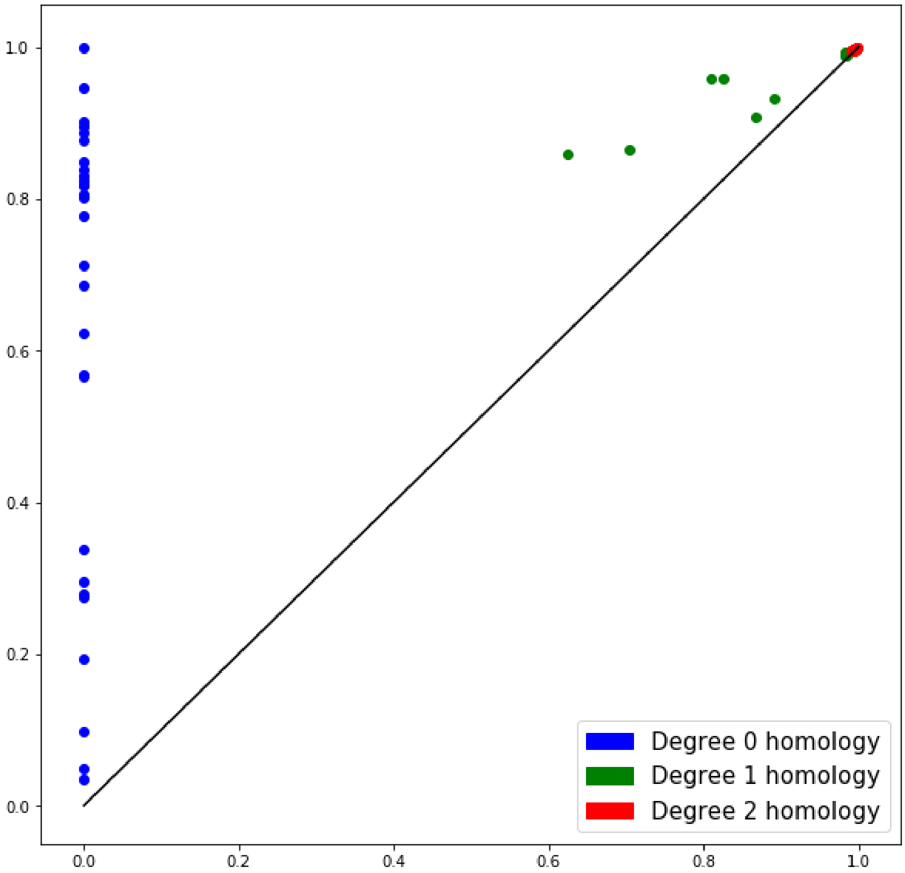

Figure 4.2: Persistence diagram of $Rips(\mathbb{X})$ for degree 0, 1 and 2 homology.

specific value for ϵ would be arbitrary and difficult to generalize. We propose a way to visualize this on Figure [4.3] where we plot the graph of:

$$\epsilon \mapsto W(dgm((X, d_{\mathbb{X}}[\,\epsilon, f\,])), dgm((X, d_{\mathbb{X}}[\,0.1, f\,]))).$$

4.4.2 The time behavior of persistence diagrams

The dissimilarity matrix in the previous subsection is obtained by averaging values over the whole period analyzed, but is this representative of the daily behavior of the LOB ?

To answer this question, we compute for all integers $i \in [\![1, N_{days}]\!]$ the persistence diagram D_i of $\mathbb{X}_i$. We would compute the 2-Wasserstein distance matrix of the set of persistence diagrams $(D_i)_{i \in [\![1, N_{days}]\!]}$ and then compute the associated Vietoris–Rips filtration $Rips((D_i)_{i \in [\![1, N_{days}]\!]})$ and plot the associated persistence diagram for degree 0, 1 and 2 homology. We leave this for future work.

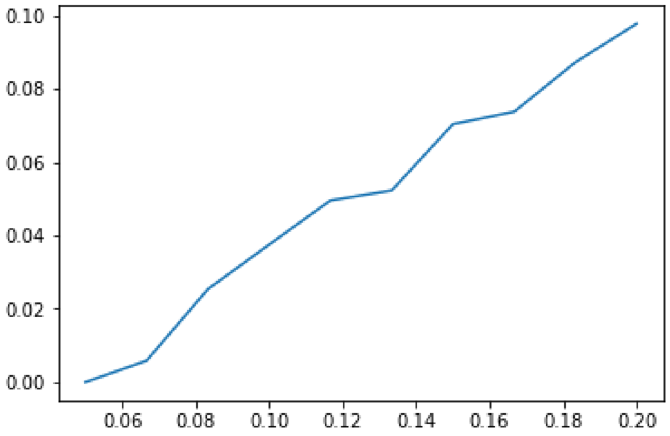

Figure 4.3: Sensitivty with respect to time slot duration

To explain this variability, we now split our data in two and we would like to know whether the two resulting sets follow the same distribution. The permutation test described in Section returns a p value superior to 0.1 which leads to reject the null hypothesis of independence of our data.

Chapter 5

Conclusion

5.1 Discussion

We started the present work by presenting the construction of persistent homology and its main descriptors, as well as some theoretical guarantees of robustness. We then reviewed the recent statistical approaches to this theory and finally applied it to limit order book data. Our main findings are:

- Limit orders tend to form significantly more simplices than Market orders.
- Oders on the same stock tend to form significantly more simplices than orders on different stocks.
- Points related to market sell orders are marginal in 3-simplices or higher.
- AAPL, MSFT and FB limit orders showed interesting correlation.

5.2 Future Extensions

5.2.1 Trading Period Classification

An extension to our work would be to understand the differences in topological persistence between the first hour of trading, the last hour of trading and the rest of the day. Our objective would be to classify persistence diagrams by trading period with the labels $\{openning, middle, closure\}$. For all $i \in [\![1, N_{days}]\!]$ and for all labels m, let $D_i^{(m)}$ be the persistence diagram of $\mathbb{H}_i^{(m)}$. Consider the classification problem: given a persistence diagram, from what trading period was the original point cloud sampled from ? As a first approach to the classification problem, we would vectorize our data similarly to [1] using persistence images. We expect that the discretization of our diagrams to be particularly sparse. We then perform a logistic regression on the vectorized dataset. A second approach would be to use the Sliced Wasserstein Kernel from Section 3.2 to implement a SVM model as described in Subsection 3.5.1. A third use case would be to assume that each observation is a collection of p persistence diagrams that all correspond to the same trading period (for instance the persistence diagrams of the first hour of trading of each day of a week). The question we ask ourselves is: given a vector of persistence diagrams, from what trading period were the p original point clouds sampled from ? We would implement the Perceptron described in Subsection 3.5.3 to answer this question. Standard comparison criteria for these models could be cross validation scores, running times and AUC. Furthermore, it would be interesting to study the persistent homology of order flows defining metrics that incorporate information about the price and size of orders.

5.2.2 Neural Network on $\mathcal{D}^p$

From a theoretical standpoint, the generalization of the Perceptron to build a full neural network seems challenging. One of the problems we encountered for this extension is the computational cost in the computation of the Fréchet mean at every step of the learning process. We are not aware of any heuristics to avoid restarting the whole scheme at every step. Nonetheless, the optimal transport approach to the Fréchet mean is a concept of interest and an extension for Bottleneck distance is an open question. Another interesting question for a deep neural network model on $\mathcal{D}$ would be: shall we embed the barycenter of the diagrams in $\mathbb{R}$ from the first layer or is it preferable to us a map on $\mathcal{D}$ and do the embedding at the last layer ? We did not implement such a model because of the theoretical difficulties coming from the notion of differentiability on $\mathcal{D}^{\mathcal{D}}$.

5.2.3 Other Topological Summaries

Finally we note that the usual topological summaries do not encode any information to identify the generators. But this is important our problem: we would like to know for instance if the generators of a tetrahedron are from the same stock or if they are all distinct immediately from the descriptor. Defining a descriptor encoding this information could be helpful to define metrics that compare instances of the descriptor according to their generators. We couldn't present the statistical approaches to other descriptors here by lack of space: the persistence landscape [7,6] and the persistence entropy [3] are examples of topological summaries suitable for a statistical approach.

Appendix A

Notation and acronyms

For all sets $\mathbb{E}, \mathbb{F} \subset \mathbb{R}$, we use the notation:

- $\mathbb{E}^* := \{x \in \mathbb{E} : x \neq 0\}$,
- $\mathbb{E}_+ := \{x \in \mathbb{E} : x \geq 0\}$,
- $\mathbb{E}_- := \{x \in \mathbb{E} : x \leq 0\}$,
- $\mathbb{E}_+^* := \{x \in \mathbb{E} : x > 0\}$,
- $\mathbb{E}_-^* := \{x \in \mathbb{E} : x < 0\}$,
- $\mathbb{F}^{\mathbb{E}} := \{f : \quad f : \mathbb{E} \to \mathbb{F}\}$.

For all real numbers $a, b \in \mathbb{R}$ such that $a < b$, we use the notation:

- $[\![a, b]\!] := [\,a, b]\, \cap \mathbb{Z}$,

Let $n, p \in \mathbb{N}$ and let $\mathbb{E}$ be a set:

- We denote by $M_{n,p}(\mathbb{E})$ the set of matrices with coefficients in $\mathbb{E}$ with n rows and p columns.
- For all matrices $M \in M_{n,p}(\mathbb{E})$, we denote by $M_{i,j}$ the entry in row i and column j. By abuse of notation we write $M = (M_{i,j})_{(i,j) \in [\![1,n]\!] \times [\![1,p]\!]}$.

We use th following notation for some specific sets:

- We denote by Δ the diagonal line in $\mathbb{R}^2$,
- We denote by $\mathcal{D}$ the set of persistence diagrams. For all filtered simplicial complex X, and by $dgm(X)$ the persistence diagram of X,

- We denote by $\mathbb{F}_2 = \mathbb{Z}/2\mathbb{Z}$ the field with cardinality 2.

We use the acronyms:

- R.K.H.S. : Reproducing Kernel Hilbert Space,
- C.N.S.D. : Conditionally Negative Semidefinite.

Appendix B

Topological definitions

Definition B.0.1 (Field). *Let E be a set. Let $+ : E \times E \to E$ and $\times : E \times E \to E$ be internal composition laws on E. We say that $(E, +, \times)$ is a field if:*

1. *$+$ and $\times$ are associative,*

2. *$+$ and $\times$ are commutative,*

3. *$+$ and $\times$ both have identity elements denoted respectively 0_E and 1_E,*

4. *$\times$ is distributive over $+$,*

5. *All the elements of E have an inverse for $+$,*

6. *All the non zero elements of E have an inverse for $\times$.*

We denote by $\mathbb{F}_2$ the field $(\{0,1\}, +, \times)$ where $\times$ is the usual multiplication law on $\mathbb{R}$ and $+$ is defined by:

$$a + b = \begin{cases} 1 & \text{if} \quad a + b \text{ is odd} \\ 0 & \text{if} \quad a + b \text{ is even} \end{cases}$$

Definition B.0.2 (Metric space). *Let E be a set. We say that a function $d : E \times E \to \mathbb{R}$ is a metric if:*

1. *$\forall x, y \in E$, $d(x, y) \geq 0$,*

2. *$\forall x, y \in E$, $d(x, y) = 0 \iff x = y$,*

3. *$\forall x, y \in E$, $d(x, y) = d(y, x)$,*

4. *$\forall x, y, z \in E$, $d(x, z) \leq d(x, y) + d(y, z)$.*

We say that (E, d) is a metric space.

Definition B.0.3 (Cauchy sequence)**.** *Let (E, d) be a metric space. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in E. We say that $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence if* $\lim_{n\to\infty} \lim_{p\to\infty} d(x_n, x_{n+p}) = 0$.

Definition B.0.4 (Complete metric space)**.** *Let (E, d) be a metric space. We say that E is complete if all Cauchy sequences of E converge in E with respect to d.*

Definition B.0.5 (Inner product space)**.** *Let E be a vector space on $\mathbb{C}$. Let $\langle ., .\rangle_E : E \times E \mapsto \mathbb{R}$. We say that $\langle ., .\rangle_E$ is an inner product in E if:*

1. $\forall x \in E,\ \langle x, x\rangle_E \geq 0$

2. $\forall x \in E,\ \langle x, x\rangle_E = 0 \iff x = 0$

3. $\forall x, y \in E,\ \langle x, y\rangle_E = \overline{\langle y, x\rangle}_E$

4. $\forall x, y, z \in E, \forall \lambda, \mu \in \mathbb{C},\ \langle \lambda x + \mu y, z\rangle_E = \lambda\langle x, z\rangle_E + \mu\langle y, z\rangle_E$

We say that $(E, \langle ., .\rangle_E)$ is an inner product space.

Definition B.0.6 (Hilbert space)**.** *We say that an inner product space $(\mathcal{H}, \langle ., .\rangle_{\mathcal{H}})$ is a Hilbert space if $\mathcal{H}$ is complete.*

Appendix C

Connection diagrams of $Rips(\mathbb{X})$

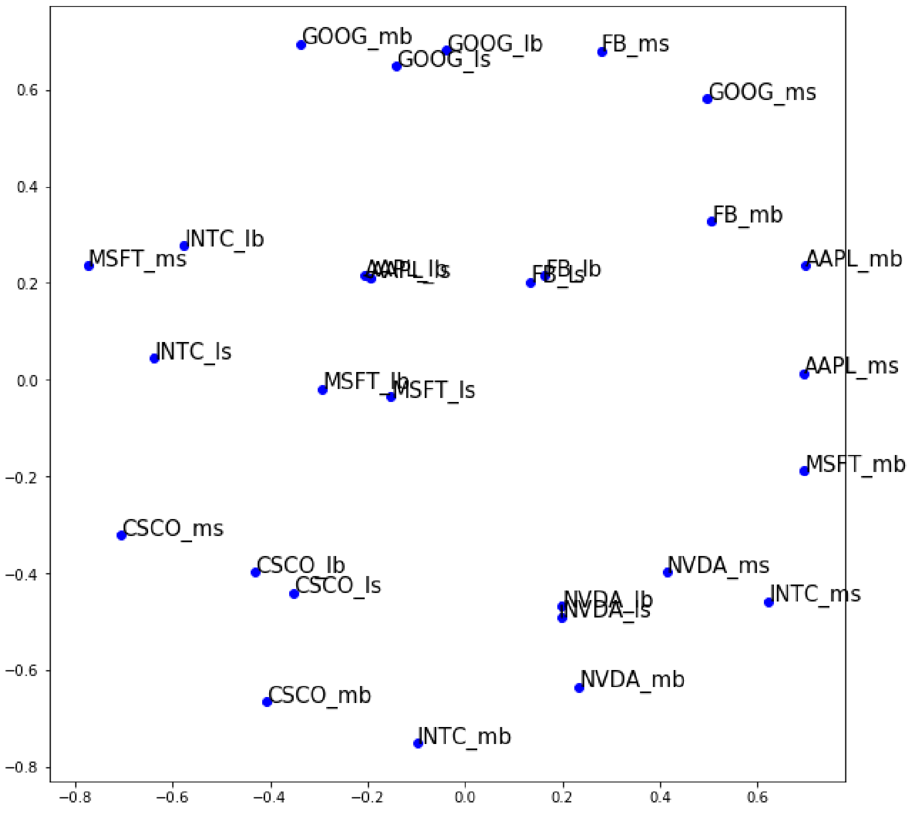

Figure C.1: Connection diagram of $\mathbb{X}$ for $r = 0.0972022$.

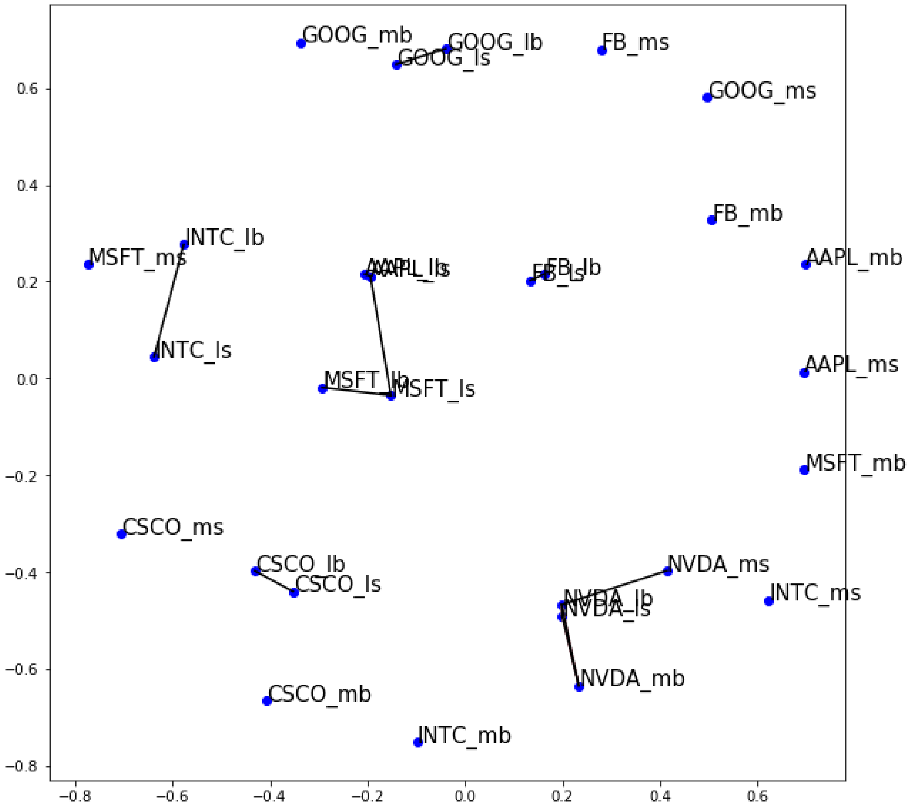

Figure C.2: Connection diagram of $\mathbb{X}$ for $r = 0.568468$.

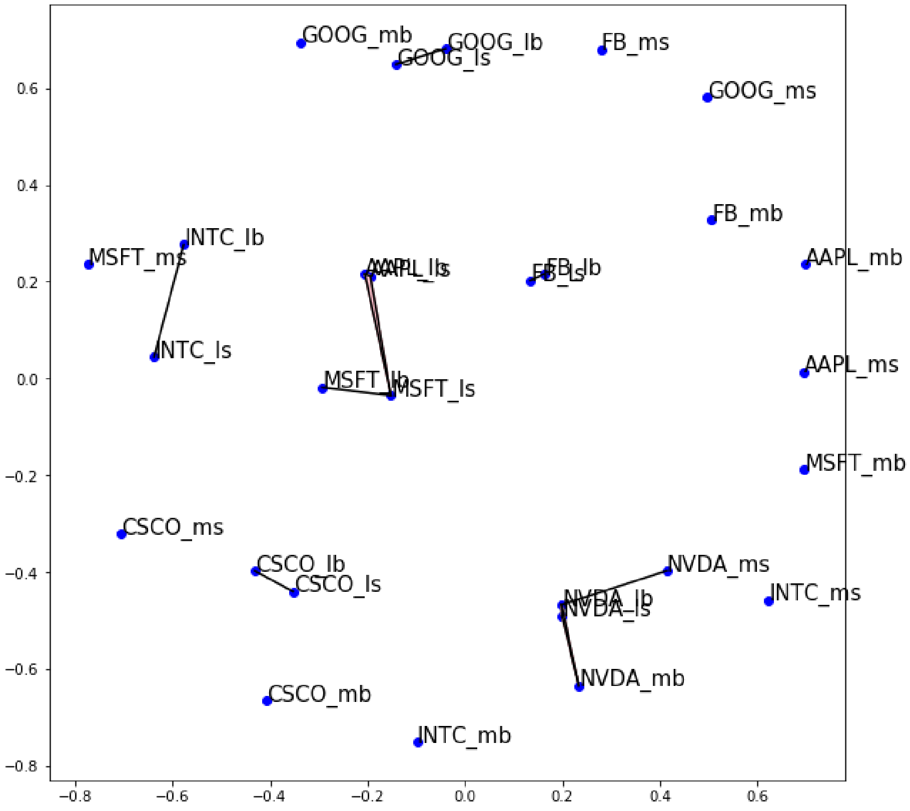

Figure C.3: Connection diagram of $\mathbb{X}$ for $r = 0.6$.

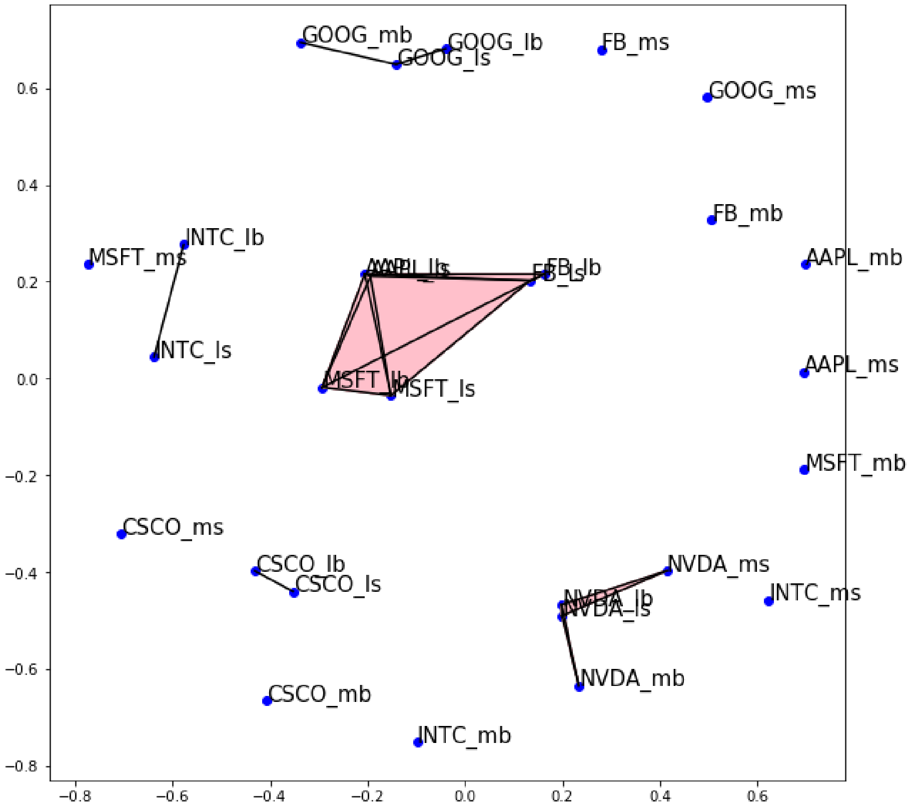

Figure C.4: Connection diagram of $\mathbb{X}$ for $r = 0.71$.

Bibliography

[1] Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, P., ... and Ziegelmeier, L. (2017). *Persistence images: A stable vector representation of persistent homology.* The Journal of Machine Learning Research, 18(1), 218-252.

[2] Aronszajn, N. (1950). *Theory of reproducing kernels.* Transactions of the American mathematical society, 68(3), 337-404.

[3] Atienza, N., Gonzalez-Diaz, R., and Soriano-Trigueros, M. (2018). *On the stability of persistent entropy and new summary functions for Topological Data Analysis.* arXiv preprint arXiv:1803.08304.

[4] Bauer, U., 2016. *Ripser.* GitHub repository, https://github.com/Ripser/ripser.

[5] Bouchaud, J. P., Farmer, J. D., and Lillo, F. (2008). *How markets slowly digest changes in supply and demand.* arXiv preprint arXiv:0809.0822.

[6] Bubenik, P. (2015). *Statistical topological data analysis using persistence landscapes.* The Journal of Machine Learning Research, 16(1), 77-102.

[7] Bubenik, P., and Kim, P. T. (2007). *A statistical approach to persistent homology.* Homology, Homotopy and Applications, 9(2), 337-362.

[8] Carlsson, G., Ishkhanov, T., De Silva, V., and Zomorodian, A. (2008). *On the local behavior of spaces of natural images.* International journal of computer vision, 76(1), 1-12.

[9] Carriere, M., Cuturi, M., and Oudot, S. (2017). *Sliced wasserstein kernel for persistence diagrams.* arXiv preprint arXiv:1706.03358.

[10] Cartea, Á., Jaimungal, S., and Penalva, J. (2015). *Algorithmic and high-frequency trading.* Cambridge University Press.

[11] Cericola, C., Johnson, I. J., Kiers, J., Krock, M., Purdy, J., and Torrence, J. (2017). *Extending hypothesis testing with persistent homology to three or more groups.* Involve, a Journal of Mathematics, 11(1), 27-51.

[12] Chazal, F., Cohen-Steiner, D., Glisse, M., Guibas, L. J., and Oudot, S. Y. (2009, June). *Proximity of persistence modules and their diagrams.* In Proceedings of the twenty-fifth annual symposium on Computational geometry (pp. 237-246). ACM.

[13] Chazal, F., De Silva, V., and Oudot, S. (2014). *Persistence stability for geometric complexes.* Geometriae Dedicata, 173(1), 193-214.

[14] Chazal, F., De Silva, V., Glisse, M., and Oudot, S. (2012). *The structure and stability of persistence modules.* arXiv preprint arXiv:1207.3674.

[15] Chazal, F., Glisse, M., Labruere, C., and Michel, B. (2013). Optimal rates of convergence for persistence diagrams in Topological Data Analysis. arXiv preprint arXiv:1305.6239.

[16] Chintakunta, H., Gentimis, T., Gonzalez-Diaz, R., Jimenez, M. J., and Krim, H. (2015). *An entropy-based persistence barcode.* Pattern Recognition, 48(2), 391-401.

[17] Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2007). *Stability of persistence diagrams.* Discrete and Computational Geometry, 37(1), 103-120.

[18] Crawford, L., Monod, A., Chen, A. X., Mukherjee, S., and Rabadán, R. (2016). *Topological summaries of tumor images improve prediction of disease free survival in glioblastoma multiforme.* arXiv preprint arXiv:1611.06818.

[19] Cuturi, M. (2013). *Sinkhorn distances: Lightspeed computation of optimal transport.* In Advances in neural information processing systems (pp. 2292-2300).

[20] De Silva, V., and Ghrist, R. (2007). *Coverage in sensor networks via persistent homology.* Algebraic and Geometric Topology, 7(1), 339-358.

[21] Duy, T. K., Hiraoka, Y., and Shirai, T. (2016). *Limit theorems for persistence diagrams.* arXiv preprint arXiv:1612.08371.

[22] Edelsbrunner, H., and Harer, J. (2010). *Computational topology: an introduction.* American Mathematical Society.

[23] Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. (2014). *Confidence sets for persistence diagrams.* The Annals of Statistics, 42(6), 2301-2339.

[24] Gidea, M. (2017). *Topology data analysis of critical transitions in financial networks.* 3rd International Winter School and Conference on Network Science. NetSci-X 2017. Springer Proceedings in Complexity. Springer, Cham

[25] Gould, M. D., Porter, M. A., Williams, S., McDonald, M., Fenn, D. J., and Howison, S. D. (2013). *Limit order books.* Quantitative Finance, 13(11), 1709-1742.

[26] Kimeldorf, G., and Wahba, G. (1971). *Some results on Tchebycheffian spline functions.* Journal of mathematical analysis and applications, 33(1), 82-95.

[27] Kovacev-Nikolic, V., Bubenik, P., Nikolić, D., and Heo, G. (2016). *Using persistent homology and dynamical distances to analyze protein binding.* Statistical applications in genetics and molecular biology, 15(1), 19-38.

[28] Kusano, G., Hiraoka, Y., and Fukumizu, K. (2016, June). *Persistence weighted Gaussian kernel for topological data analysis.* In International Conference on Machine Learning (pp. 2004-2013).

[29] Lacombe, T., Cuturi, M., and Oudot, S. (2018). *Large Scale computation of Means and Clusters for Persistence Diagrams using Optimal Transport.* arXiv preprint arXiv:1805.08331.

[30] Leibon, G., Pauls, S., Rockmore, D., and Savell, R. (2008). *Topological structures in the equities market network.* Proceedings of the National Academy of Sciences, 105(52), 20589-20594.

[31] Liu, J. Y., Jeng, S. K., and Yang, Y. H. (2016). *Applying topological persistence in convolutional neural network for music audio signals.* arXiv preprint arXiv:1608.07373.

[32] Mileyko, Y., Mukherjee, S., and Harer, J. (2011). *Probability measures on the space of persistence diagrams.* Inverse Problems, 27(12), 124007.

[33] Munkres, J. (1957). *Algorithms for the assignment and transportation problems.* Journal of the society for industrial and applied mathematics, 5(1), 32-38.

[34] Nicolau, M., Levine, A. J., and Carlsson, G. (2011). *Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival.* Proceedings of the National Academy of Sciences, 108(17), 7265-7270.

[35] Obayashi, I., and Hiraoka, Y. (2017). *Persistence Diagrams with Linear Machine Learning Models.* arXiv preprint arXiv:1706.10082.

[36] Otter, N., Porter, M. A., Tillmann, U., Grindrod, P., and Harrington, H. A. (2017). *A roadmap for the computation of persistent homology.* EPJ Data Science, 6(1), 17.

[37] Phipson, B., and Smyth, G. K. (2010). *Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn.* Statistical applications in genetics and molecular biology, 9(1).

[38] Reininghaus, J., Huber, S., Bauer, U., and Kwitt, R. (2015). *A stable multi-scale kernel for topological machine learning.* In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4741-4748).

[39] Robinson, A., and Turner, K. (2017). *Hypothesis testing for topological data analysis.* Journal of Applied and Computational Topology, 1(2), 241-261.

[40] Rosenblatt, F. (1958). *The perceptron: a probabilistic model for information storage and organization in the brain.* Psychological review, 65(6), 386.

[41] Schölkopf, B., Herbrich, R., and Smola, A. J. (2001, July). *A generalized representer theorem.* In International conference on computational learning theory (pp. 416-426). Springer, Berlin, Heidelberg.

[42] Sejdinovic, D., Gretton, A., Sriperumbudur, B., and Fukumizu, K. (2012). *Hypothesis testing using pairwise distances and associated kernels (with appendix).* arXiv preprint arXiv:1205.0411

[43] Turner, K., Mileyko, Y., Mukherjee, S., and Harer, J. (2014). *Féchet means for distributions of persistence diagrams.* Discrete and Computational Geometry, 52(1), 44-70.

[44] Wu, J., Huang, Z., Li, W., Thoma, J., and Van Gool, L. (2017). *Sliced Wasserstein Generative Models.* arXiv preprint arXiv:1706.02631.

[45] Xia, K., and Wei, G. W. (2014). *Persistent homology analysis of protein structure, flexibility, and folding.* International journal for numerical methods in biomedical engineering, 30(8), 814-844.

[46] Zomorodian, A., and Carlsson, G. (2005). *Computing persistent homology.* Discrete and Computational Geometry, 33(2), 249-274.