The Christoffel-Darboux Kernel for Topological Data Analysis

Pepijn Roos Hoefgeest ✉
Vrije Universiteit (VU) Amsterdam, The Netherlands

Lucas Slot ✉
ETH Zürich, Switzerland

Abstract

Persistent homology has been widely used to study the topology of point clouds in $\mathbb{R}^n$. Standard approaches are very sensitive to outliers, and their computational complexity depends badly on the number of data points. In this paper we introduce a novel persistence module for a point cloud using the theory of Christoffel-Darboux kernels. This module is robust to (statistical) outliers in the data, and can be computed in time linear in the number of data points. We illustrate the benefits and limitations of our new module with various numerical examples in $\mathbb{R}^n$, for $n = 1, 2, 3$. Our work expands upon recent applications of Christoffel-Darboux kernels in the context of statistical data analysis and geometric inference [13]. There, these kernels are used to construct a polynomial whose level sets capture the geometry of a point cloud in a precise sense. We show that the persistent homology associated to the sublevel set filtration of this polynomial is stable with respect to the Wasserstein distance. Moreover, we show that the persistent homology of this filtration can be computed in singly exponential time in the ambient dimension n, using a recent algorithm of Basu & Karisani [1].

2012 ACM Subject Classification Mathematics of computing → Algebraic topology

Keywords and phrases Topological Data Analysis, Geometric Inference, Christoffel-Darboux Kernels, Persistent Homology, Wasserstein Distance, Semi-Algebraic Sets

Digital Object Identifier 10.4230/LIPIcs.SoCG.2023.38

Related Version *Full Version*: `https://arxiv.org/abs/2211.15489`

1 Introduction

Persistent homology is a central tool in the field of topological data analysis. It was developed in the early 2000s in order to extract topological and geometric information out of point-cloud data. Since discrete points in $\mathbb{R}^n$ do not have any meaningful topological features in and of themselves, one needs to find a way to construct an "interesting" topological space out of them. An obvious approach is to look at the collection of balls of radius r centered around the data points. When the radius r is chosen correctly, these balls will intersect in ways that reflect the topology of the set the data is sampled from. However, it is not clear a priori which radius should be chosen, and in fact, a *single* "correct" choice need not even exist. The solution is to look at *all* radii $r \geq 0$, and to track which topological features *persist* over time as r increases. More concretely: if $r \leq r'$, then the balls of radius r include into those of radius r', and this collection of inclusions forms what is called a *filtration.* Persistent homology tracks "birth-death events" of homology classes in such a filtration, see Figure 1. Classical approaches to obtain a filtration of topological spaces out of a point cloud, such as the *Čech filtration* (outlined above) and the *Vietoris-Rips filtration* [23] suffer from two main problems:

1. The complexity of computing the persistent homology depends badly on the number of data points.
2. The persistent homology of these filtrations is very sensitive to outliers in the data.

39th International Symposium on Computational Geometry (SoCG 2023).
Editors: Erin W. Chambers and Joachim Gudmundsson; Article No. 38; pp. 38:1–38:20
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

These issues have been addressed in the literature in several ways. *Alpha complexes* [9] are used to compute the persistent homology of the Čech filtration efficiently by first intersecting the metric balls with a Voronoi diagram. *Witness complexes* [18] build a small simplicial complex based on a subsample of the data, thus reducing computational complexity. Heuristically, these subsamples may be chosen to reduce sensitivity to outliers in the full data set, although this effect remains hard to quantify [19]. Chazal et al. [4] introduce the *distance-to-measure* function, which they apply to perform geometric inference of point clouds in $\mathbb{R}^n$. The key feature of this function is that it is stable with respect to the *Wasserstein distance*, implying robustness to (statistical) outliers in the data. Buchet et al. [3] use this property to construct a filtration which is also provably stable in this sense. However, it is hard to compute the associated persistent homology, and they therefore employ an approximation scheme.

In this paper, we propose a novel filtration based on so-called *Christoffel-Darboux kernels.* As we explain in more detail below, the resulting persistent homology can be computed in *linear time* in the number of data points, and is provably robust to statistical outliers. Christoffel-Darboux (CD) kernels have a long history in fundamental mathematics, with applications to orthogonal polynomials and in approximation theory (see [13] for an overview). They are the reproducing kernels $K_d^\mu : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ for the Hilbert space $\mathbb{R}[\mathbf{x}]_d$ of n-variate, real polynomials of degree $d \in \mathbb{N}$ with respect to the inner product $\langle p, q \rangle = \int pq d\mu$ induced by a finite measure μ on $\mathbb{R}^n$. Such reproducing kernels completely describe the inner product $\langle \cdot, \cdot \rangle$, and in our setting the kernels K_d^μ thus capture information about the underlying measure μ. For instance, the *Christoffel polynomial* $P_d^\mu(\mathbf{x}) := K_d^\mu(\mathbf{x}, \mathbf{x})$ can be used to estimate the support $\mathrm{supp}(\mu) \subseteq \mathbb{R}^n$ of μ. Roughly speaking, $P_d^\mu(\mathbf{x})$ is small when $\mathbf{x} \in \mathrm{supp}(\mu)$ and large when $\mathbf{x} \notin \mathrm{supp}(\mu)$ (see Proposition 10 below). This property has recently been applied to perform geometric inference in a statistical setting by Lasserre, Pauwels and Putinar [12, 13, 15]. In these works, the authors consider CD kernels for the *empirical* measure $\mu_\mathcal{X}$ associated to a set of samples $\mathcal{X}$ drawn according to some unknown measure μ. For fixed $d \in \mathbb{N}$, the polynomial $P_d^{\mu_\mathcal{X}}$ associated to $\mathcal{X}$ is straightforward to compute. Moreover, the *sublevel set* $\{\mathbf{x} \in \mathbb{R}^n : P_d^{\mu_\mathcal{X}}(\mathbf{x}) \leq t\}$ captures the support of μ well for suitably selected $t \geq 0$, see Figure 2. However, a key issue of this approach is that the level $t \geq 0$ must be selected "by hand" based on heuristics, and the quality of geometric inference depends heavily on this choice. This problem motivates our use of persistent homology, which considers all sublevel sets simultaneously.

1.1 Contributions and outline

We propose a new scheme for topological data analysis of a finite point cloud $\mathcal{X} \subseteq [-1, 1]^n$, based on Christoffel-Darboux kernels. Our scheme unites recent applications of CD kernels in (statistical) data analysis with ideas from persistent homology. It consists of three steps:

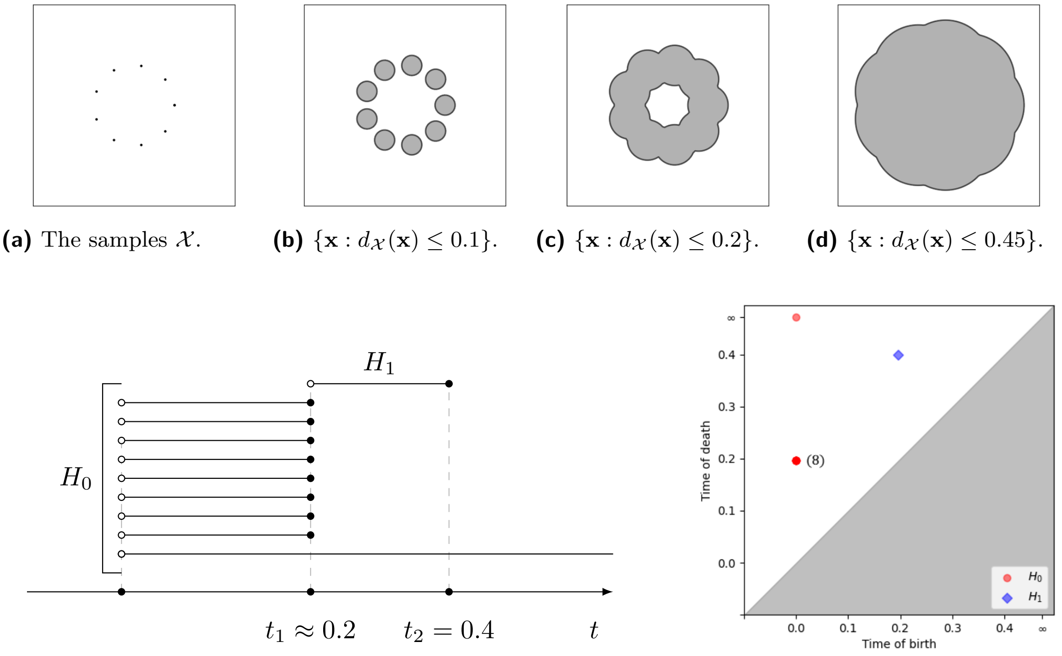
1. **Moment matrix**. Fix $d \in \mathbb{N}$. Choose a basis $\mathbf{b} = (b_\alpha)$ for the space $\mathbb{R}[\mathbf{x}]_d$ of n-variate polynomials of degree at most d. Compute the *moment matrix* $M_d(\mathbf{b})$ of size $\binom{n+d}{d}$, whose entries can be computed from $\mathcal{X}$ in linear time via:

$$M_d(\mathbf{b})_{\alpha,\beta} := \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X}} b_\alpha(\mathbf{x}) b_\beta(\mathbf{x}) \quad (\alpha, \beta \in \mathbb{N}^n,\ |\alpha|, |\beta| \leq d).$$

2. **Christoffel polynomial**. Invert the moment matrix to obtain the *Christoffel polynomial*:

$$P_d(\mathbf{x}) := \mathbf{b}(\mathbf{x})^\top \big(M_d(\mathbf{b})\big)^{-1} \mathbf{b}(\mathbf{x}),$$

whose sublevel sets are known to approximate the set $\mathcal{X}$, see Figure 2.



(a) The samples $\mathcal{X}$. **(b)** $\{\mathbf{x} : d_{\mathcal{X}}(\mathbf{x}) \leq 0.1\}$. **(c)** $\{\mathbf{x} : d_{\mathcal{X}}(\mathbf{x}) \leq 0.2\}$. **(d)** $\{\mathbf{x} : d_{\mathcal{X}}(\mathbf{x}) \leq 0.45\}$.

(e) The interval modules. **(f)** The persistence diagram.

Figure 1 A filtration of $[-1, 1]^2$ by the distance function $d_{\mathcal{X}} : \mathbf{x} \mapsto \text{dist}(\mathbf{x}, \mathcal{X})$ to a set of equidistant points $\mathcal{X}$ on a circle of radius 0.4, and the corresponding persistence diagram. Note that there are 8 intervals that are born at $t = 0$ and die at $t \approx 0.2$, which show up as a single dot in the diagram. Throughout, we indicate the number of such overlapping dots in the diagram if necessary for clarity.

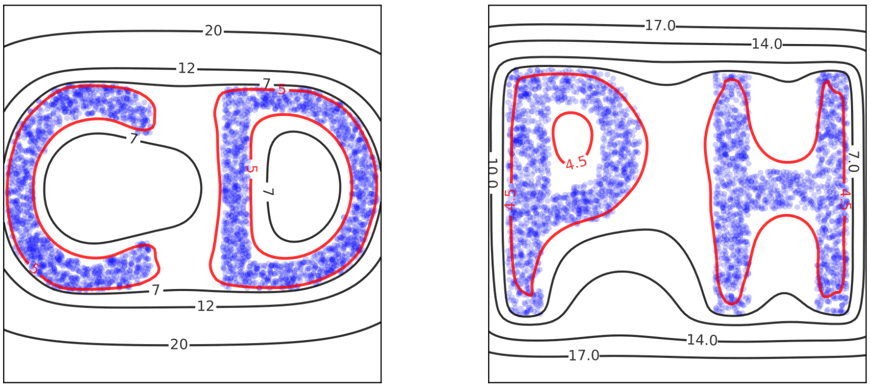


Figure 2 The level sets of the Christoffel polynomial $\mathbf{x} \mapsto P_{10}^{\mu_{\mathcal{X}}}(\mathbf{x})$ associated to the empirical measure $\mu_{\mathcal{X}}$ of two sample sets $\mathcal{X} \subseteq [-1, 1]^2$ (in blue). The level sets indicated in red capture the support of the underlying measure μ quite well.

3. **Persistence module.** Define the *sublevel set filtration*:

$$\mathbf{X}_t := \{\mathbf{x} \in [-1,1]^n : \log P_d(\mathbf{x}) \leq t\} \quad (t \geq 0),$$

and compute its associated *persistence module*:

$$\mathbb{CD}(\mathcal{X}, d) := \mathrm{PH}_*([-1,1]^n,\ \log P_d).$$

Robustness to statistical outliers. We show that the module $\mathbb{CD}(\mathcal{X}, d)$ is stable and robust under perturbations of the input data $\mathcal{X}$. To be precise, we show *local* Lipschitz continuity of the function $\mathcal{X} \mapsto \mathbb{CD}(\mathcal{X}, d)$, in the *Bottleneck* and *Wasserstein* distance. We also give an estimate of the Lipschitz constant in terms of a concrete measure of algebraic degeneracy of the set $\mathcal{X}$. This is our main technical result, see Section 3.1.

Exact algorithm with linear dependence on the number of samples. We give an exact algorithm for computing the persistence module $\mathbb{CD}(\mathcal{X}, d)$ in Section 3.2, whose runtime is linear in the number of data points, but depends exponentially on the dimension n. This algorithm is a combination of 1) a known procedure to compute CD kernels and 2) the recent work [1], in which the authors propose an algorithm for computing the persistent homology of *semialgebraic* filtrations.

Numerical examples. We provide several numerical examples in Section 4 that illustrate the geometric properties of our scheme, and its potential benefits and downsides compared to existing methods. Unfortunately, there is no practical implementation available of the algorithm proposed in [1]. In order to perform numerical experiments, we therefore propose a simple scheme for approximating $\mathbb{CD}(\mathcal{X}, d)$ in Section 3.3, based on a triangulation of the sample space $[-1,1]^n$. These experiments show that our novel persistence module is able to accurately capture underlying homological features of point clouds, even in the presence of outliers.

2 Background

Notations and conventions. Throughout, $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$ are n-dimensional variables. We denote by $\mathbb{R}[\mathbf{x}]$ the n-variate polynomial ring. We write $\mathbb{R}[\mathbf{x}]_d \subseteq \mathbb{R}[\mathbf{x}]$ for the subspace of polynomials of (total) degree at most d, which has (real) dimension $s(n,d) = \binom{n+d}{d}$. For ease of exposition, we assume throughout that sets of samples $\mathcal{X}, \mathcal{Y}$ are contained in the box $[-1,1]^n$, which can always be achieved by a rescaling.

2.1 Persistent homology

Persistent homology is a central tool in topological data analysis, and has received a lot of attention over recent years [8]. It serves to track homology classes through a diagram of spaces, typically arising from a filtration: Let $\mathbf{X}$ be a filtered topological space, that is, for each $t \in \mathbb{R}$, there is a subspace $\mathbf{X}_t \subseteq \mathbf{X}$, such that if $s \leq t$, $\mathbf{X}_s \subseteq \mathbf{X}_t$. For convenience, we assume that the filtration is exhaustive, i.e. $\bigcup_t \mathbf{X}_t = \mathbf{X}$. Applying homology with coefficients in $\mathbb{F}$ to each $\mathbf{X}_t$ then yields a diagram of spaces $\mathrm{H}_*(\mathbf{X}_t; \mathbb{F})$: For every $s \leq t$, the inclusion map $\iota_s^t : \mathbf{X}_s \to \mathbf{X}_t$ induces a map $h_s^t = (\iota_s^t)_* : \mathrm{H}_*(\mathbf{X}_s; \mathbb{F}) \to \mathrm{H}_*(\mathbf{X}_t; \mathbb{F})$, and the collection of maps $\{h_s^t\}$ satisfy:

1. For all $r \leq s \leq t$, $h_s^t \circ h_r^s = h_r^t$;
2. For all $t \in \mathbb{R}$, $h_t^t = id_{\mathrm{H}_*(\mathbf{X}_t;\mathbb{F})}$.

This diagram of spaces is the *persistent homology* of $\mathbf{X}$, denoted by $\mathrm{PH}_*(\mathbf{X}_t;\mathbb{F})$. Any $\mathbb{R}$-indexed collection of vector spaces with maps satisfying **1.** and **2.** above is called a *persistence module.* More succinctly put, a persistence module is a functor from the poset $(\mathbb{R},\leq)$ to the category of vector spaces over some field. The vector spaces can be taken over any field $\mathbb{F}$, but we always work over a finite field.

▶ **Example 1.** Let $\mathbf{X}$ be a topological space, and let $f:\mathbf{X}\to\mathbb{R}$ be a continuous function. Then the sublevel set filtration of $\mathbf{X}$ with respect to f is given by $\mathbf{X}_t=\{\mathbf{x}\in\mathbf{X}\,|\,f(\mathbf{x})\leq t\}$. Applying homology to this filtration yields a persistence module, which we denote by $\mathrm{PH}_*(\mathbf{X},f)$.

If $\mathrm{H}_*(\mathbf{X}_t)$ is finite dimensional for each $t\in\mathbb{R}$ (which is a mild requirement, and will always be satisfied in our setting), then $\mathrm{PH}_*(\mathbf{X})$ is completely described by a set of intervals, denoted by $\mathrm{Dgm}(\mathrm{PH}_*(\mathbf{X}))$. The presence of an interval $[t_b,t_d)\in\mathrm{Dgm}(\mathrm{PH}_p(\mathbf{X}))$ tells us that a particular p-dimensional homology class is born at time t_b, and lives until time t_d, where it then dies. If $t_d=\infty$, this means that this homology class lives forever, and corresponds to a global homology class in $\mathrm{H}_p(\mathbf{X})$. The diagram $\mathrm{Dgm}(\mathrm{PH}_*(\mathbf{X}))$ can be conveniently visualized, see Figure 1.

2.1.1 Stability

An important property of a persistence module one needs to verify before using it in an application, is that it is stable with respect to the input data. Intuitively, this means that small perturbations of the input data should result only in small perturbations in the obtained persistence diagrams. We make this precise below.

▶ **Definition 2.** *A matching between two multi-sets A and B is a bijection χ between two subsets $A'\subset A$ and $B'\subset B$. We denote this by $\chi:A\twoheadrightarrow B$. If χ matches $a\in A$ to $b\in B$, we write $(a,b)\in\chi$. If $c\in A\cup B$ is unmatched by χ, we abuse notation and write $c\notin\chi$.*

▶ **Definition 3.** *Let $I=\langle t_{b_1},t_{d_1}\rangle$ and $J=\langle t_{b_2},t_{d_2}\rangle$ be two intervals in $\overline{\mathbb{R}}=\mathbb{R}\cup\{-\infty,\infty\}$. Then the cost of I is given by*

$$c(I):=(t_{d_1}-t_{b_1})/2$$

and the cost of the pair (I,J) is given by:

$$c(I,J):=\max\{|t_{b_1}-t_{b_2}|,\ |t_{d_1}-t_{d_2}|\}$$

Now let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two multi-sets of intervals in $\overline{\mathbb{R}}=\mathbb{R}\cup\{-\infty,\infty\}$. The cost of a matching $\chi:\mathcal{D}_1\twoheadrightarrow\mathcal{D}_2$ is defined as

$$\mathrm{cost}(\chi):=\max\Big\{\sup_{(I,J)\in\chi}c(I,J),\ \sup_{I\notin\chi}c(I)\Big\}$$

Finally, the Bottleneck distance *between $\mathcal{D}_1$ and $\mathcal{D}_2$ is given by:*

$$d_B(\mathcal{D}_1,\mathcal{D}_2):=\inf_{\chi:\mathcal{D}_1\twoheadrightarrow\mathcal{D}_2}\mathrm{cost}(\chi)$$

The Bottleneck distance is the most widely used distance on the space of persistence diagrams, and it satisfies the following:

▶ **Theorem 4** ([6]). *Suppose $\mathbf{X}$ is a CW-complex, and $f,g:\mathbf{X}\to\mathbb{R}$ are two continuous functions on $\mathbf{X}$. Then*

$$d_B(\mathrm{Dgm}(\mathrm{PH}_p(\mathbf{X},f)),\mathrm{Dgm}(\mathrm{PH}_p(\mathbf{X},g)))\leq\|f-g\|_\infty:=\max_{\mathbf{x}\in\mathbf{X}}|f(\mathbf{x})-g(\mathbf{x})|.$$

▶ **Definition 5.** *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be two subsets of a metric space* (M,d)*. Write* $U_\varepsilon(\mathbf{X}) := \{m \in M \,|\, \mathrm{dist}(m,\mathbf{X}) \leq \varepsilon\}$*. Then the Hausdorff distance between* $\mathbf{X}$ *and* $\mathbf{Y}$ *is given by:*

$$\begin{aligned} d_H(\mathbf{X},\mathbf{Y}) : &= \inf\{\varepsilon \geq 0 \,|\, \mathbf{Y} \subseteq U_\varepsilon(\mathbf{X}) \text{ and } \mathbf{X} \subseteq U_\varepsilon(Y)\} \\ &= \max\Big\{\sup_{\mathbf{x}\in\mathbf{X}} \mathrm{dist}(\mathbf{x},\mathbf{Y}),\ \sup_{\mathbf{y}\in\mathbf{Y}} \mathrm{dist}(\mathbf{y},\mathbf{X})\Big\}. \end{aligned}$$

▶ **Example 6.** Let $\mathbf{X} \subseteq M$, be a subset of a metric space (M,d). Then the distance function

$$d_{\mathbf{X}} : M \to \mathbb{R}, \quad \mathbf{x} \mapsto \mathrm{dist}(\mathbf{x},\mathbf{X})$$

is a continuous function on M, and defines a sublevel set filtration and a persistence module, which we suggestively denote by $\mathrm{PH}_p(\check{C}(\mathbf{X}))$. Note that if $\mathbf{Y}$ is another subset of M, then $\|d_{\mathbf{X}} - d_{\mathbf{Y}}\|_\infty = d_H(\mathbf{X},\mathbf{Y})$, so it follows from Theorem 4 that

$$d_B\left(\mathrm{Dgm}(\mathrm{PH}_p(\check{C}(\mathbf{X}))),\ \mathrm{Dgm}(\mathrm{PH}_p(\check{C}(\mathbf{Y})))\right) \leq d_H(\mathbf{X},\mathbf{Y}).$$

In the above example, $\mathbf{X}$ is typically a finite subset of $\mathbb{R}^n$, and this is often used as one of the motivating examples for persistent homology. When $\mathbf{X}$ is sampled from some unknown shape $\mathfrak{X}$ inside of $\mathbb{R}^n$, its persistent homology can be used to estimate the homology of $\mathfrak{X}$. It can be computed using the Čech filtration of $\mathbf{X}$, which is a filtered simplicial complex whose homotopy type at each stage agrees with that of the sublevel set of $d_{\mathbf{X}}$ at the same scale. It is true, but not entirely straight-forward, that the two persistence modules arising from these constructions are isomorphic [2, 5].

2.1.2 Wasserstein distance

In Section 3.1, we will show stability results for our novel persistence module in terms of the *Wasserstein distance.* The Wasserstein distance is a metric on the space of probability measures supported on $\mathbb{R}^n$. It is commonly used in the context of optimal transport and (statistical) data analysis, see, e.g., [16]. The primary advantage of the Wasserstein distance over the Hausdorff distance is that it is much less sensitive to outliers, and therefore more suited to applications in statistics. For our purposes, it is enough to consider probability measures with *finite support.*

▶ **Definition 7.** *Let* $\mu_{\mathcal{X}}, \mu_{\mathcal{Y}}$ *be two probability measures with finite supports* $\mathcal{X},\mathcal{Y} \subseteq \mathbb{R}^n$*, respectively. The Wasserstein distance* $d_W(\mu_{\mathcal{X}},\mu_{\mathcal{Y}})$ *is then given by the optimum solution to the linear program:*

$$d_W(\mu_{\mathcal{X}},\mu_{\mathcal{Y}}) := \min_{\gamma} \sum_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}} \gamma(\mathbf{x},\mathbf{y}) \cdot \|\mathbf{x}-\mathbf{y}\|_2 \tag{1}$$

$$\text{s.t.} \quad \sum_{\mathbf{y}\in\mathcal{Y}} \gamma(\mathbf{x},\mathbf{y}) = \mu_{\mathcal{X}}(\mathbf{x}) \tag{2}$$

$$\sum_{\mathbf{x}\in\mathcal{X}} \gamma(\mathbf{x},\mathbf{y}) = \mu_{\mathcal{Y}}(\mathbf{y}) \tag{3}$$

$$\gamma : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}. \tag{4}$$

One can think of $d_W(\mu_{\mathcal{X}},\mu_{\mathcal{Y}})$ as the amount of "work" required to transform the measure $\mu_{\mathcal{X}}$ into $\mu_{\mathcal{Y}}$. For instance, if $\mathcal{Y} = \{\mathbf{y}\}$ is a singleton, then $d_W(\mu_{\mathcal{X}},\mu_{\{\mathbf{y}\}})$ is simply given by:

$$d_W(\mu_{\mathcal{X}},\mu_{\{\mathbf{y}\}}) = \sum_{\mathbf{x}\in\mathcal{X}} \mu_{\mathcal{X}}(\mathbf{x}) \cdot \|\mathbf{x}-\mathbf{y}\|_2.$$

2.2 The Christoffel-Darboux kernel

In this section, we introduce some basic facts on Christoffel-Darboux kernels, with emphasis on the statistical setting. We refer to the book of Lasserre, Pauwels and Putinar [13] for a comprehensive treatment. Let μ be a finite, positive Borel measure on $\mathbb{R}^n$ with compact, full-dimensional support. (In our setting, it is helpful to think of μ as the restriction of the Lebesgue measure to a sufficiently nice compact subset of $\mathbb{R}^n$). Then μ induces an inner product on the space $\mathbb{R}[\mathbf{x}]$ of n-variate, real polynomials via:

$$\langle p, q\rangle_\mu := \int p(\mathbf{x})q(\mathbf{x})d\mu(\mathbf{x}). \tag{5}$$

We can choose an ortho*normal* basis $\mathbf{b} = \{b_\alpha : \alpha \in \mathbb{N}^n\}$ for $\mathbb{R}[\mathbf{x}]$ with respect to $\langle\cdot,\cdot\rangle_\mu$, which we order so that $b_\alpha \in \mathbb{R}[\mathbf{x}]$ is of total degree $|\alpha| = \sum_{i=1}^n \alpha_i$ for each $\alpha \in \mathbb{N}^n$. That is, we have the orthogonality relations:

$$\langle b_\alpha, b_\beta\rangle_\mu = \int b_\alpha(\mathbf{x})b_\beta(\mathbf{x})d\mu(\mathbf{x}) = \delta_{\alpha\beta} \quad (\alpha, \beta \in \mathbb{N}^n). \tag{6}$$

Using this orthonormal basis, we can define the Christoffel-Darboux kernel.

▶ **Definition 8.** *For $d \in \mathbb{N}$, the* Christoffel-Darboux kernel $K_d^\mu : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ *of degree d for the measure μ is defined as:*

$$K_d^\mu(\mathbf{x},\mathbf{y}) := \sum_{|\alpha|\le d} b_\alpha(\mathbf{x})b_\alpha(\mathbf{y}). \tag{7}$$

The Christoffel-Darboux kernel K_d^μ is also called the *reproducing kernel* for the Hilbert space $(\mathbb{R}[\mathbf{x}]_d, \langle\cdot,\cdot\rangle_\mu)$, as we have the reproducing property:

$$\int K_d^\mu(\mathbf{x},\mathbf{y})p(\mathbf{y})d\mu(\mathbf{y}) = \langle K_d^\mu(\mathbf{x},\cdot), p(\cdot)\rangle_\mu = p(\mathbf{x}) \quad (p \in \mathbb{R}[\mathbf{x}]_d). \tag{8}$$

We note that the kernel K_d^μ is independent of our choice of basis $\{b_\alpha\}$, and it can be computed via the Gram-Schmidt procedure even if we do not have access to an explicit orthonormal basis.

▶ **Proposition 9** (Gram-Schmidt, see Prop. 4.1.2 in [13]). *Let $d \in \mathbb{N}$ and let $\mathbf{b} = \{b_\alpha : |\alpha| \le d\}$ be* any *basis for $\mathbb{R}[\mathbf{x}]_d$. For $\mathbf{x} \in \mathbb{R}^n$, write $\mathbf{b}_d(\mathbf{x}) = (b_\alpha(\mathbf{x}))_{|\alpha|\le d} \in \mathbb{R}^{s(n,d)}$ and consider the matrix $M_d^\mu(\mathbf{b}) \in \mathbb{R}^{s(n,d)\times s(n,d)}$ given by the entrywise integral:*

$$M_d^\mu(\mathbf{b}) := \int \mathbf{b}_d(\mathbf{x})\mathbf{b}_d(\mathbf{x})^\top d\mu(\mathbf{x}),$$

$$\textit{i.e.,} \quad (M_d^\mu(\mathbf{b}))_{\alpha,\beta} = \int b_\alpha(\mathbf{x})b_\beta(\mathbf{x})d\mu(\mathbf{x}) \quad (\alpha,\beta \in \mathbb{N}_d^n). \tag{9}$$

The matrix $M_d^\mu(\mathbf{b})$ is strictly positive semidefinite, i.e., its eigenvalues are all stricly larger than 0. Moreover, we have:

$$K_d^\mu(\mathbf{x},\mathbf{y}) = \mathbf{b}_d(\mathbf{x})^\top \left(M_d^\mu(\mathbf{b})\right)^{-1}\mathbf{b}_d(\mathbf{y}).$$

The Christoffel polynomial. For our purposes, we are mostly interested in the *Christoffel polynomial* $P_d^\mu : \mathbb{R}^n \to \mathbb{R}$, defined in terms of an orthonormal basis $\mathbf{b}$ for $(\mathbb{R}[\mathbf{x}]_d, \langle\cdot,\cdot\rangle_\mu)$ as:

$$P_d^\mu(\mathbf{x}) := K_d^\mu(\mathbf{x},\mathbf{x}) = \sum_{|\alpha|\le d} b_\alpha(\mathbf{x})^2. \tag{10}$$

The Christoffel polynomial is a *sum of squares* of polynomials, implying immediately that $P_d^\mu(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$. In fact, by definiteness of the inner product (5), it is strictly positive on $\mathbb{R}^n$. It has a remarkable alternative definition in terms of a *variational problem*:

$$\frac{1}{P_d^\mu(\mathbf{z})} = \min_{p \in \mathbb{R}[\mathbf{x}]_d} \left\{ \int p^2(\mathbf{x}) d\mu(\mathbf{x}) : p(\mathbf{z}) = 1 \right\} \quad (\mathbf{z} \in \mathbb{R}^n).$$

The Christoffel polynomial encodes information on the support of μ. Roughly speaking, $P_d^\mu(\mathbf{x})$ is rather *large* when $\mathbf{x} \notin \mathrm{supp}(\mu)$, and rather *small* when $\mathbf{x} \in \mathrm{supp}(\mu)$ (see Figure 2). This can be made precise in the regime $d \to \infty$.

▶ **Proposition 10** (see Sec. 4.3 of [13]). *Under certain assumptions on the measure μ, we have:*

$$\lim_{d\to\infty} P_d^\mu(\mathbf{x}) = \begin{cases} O(d^n) & \mathbf{x} \in \mathrm{int}\big(\mathrm{supp}(\mu)\big), \\ \Omega(\exp(\alpha d)) & \mathbf{x} \notin \mathrm{supp}(\mu), \end{cases}$$

for any fixed $\mathbf{x} \in \mathbb{R}^n$. Here, the constant α is proportional to $\mathrm{dist}(\mathbf{x}, \mathrm{supp}(\mu))$.

When μ is the restriction of the Lebesgue measure to a sufficiently nice compact subset $\mathfrak{X} \subseteq \mathbb{R}^n$, a stronger result is shown by Lasserre and Pauwels [12].

▶ **Theorem 11** (reformulation of Thm. 7.3.2 in [13]). *Let $\mathfrak{X} \subseteq \mathbb{R}^n$ be a compact set satisfying the conditions of Assumption 7.3.1 in [13], and let μ be the restriction of the Lebesgue measure to $\mathfrak{X}$. Then there exist sequences $(t_k)_{k\in\mathbb{N}}$ and $(d_k)_{k\in\mathbb{N}}$ such that the sublevel sets $\mathbf{X}_k := \{\mathbf{x} \in \mathbb{R}^n : P_{d_k}^\mu(\mathbf{x}) \leq t_k\}$ satisfy:*

$$\lim_{k\to\infty} d_H(\mathbf{X}_k, \mathfrak{X}) = 0, \quad \textit{and} \quad \lim_{k\to\infty} d_H(\partial\mathbf{X}_k, \partial\mathfrak{X}) = 0.$$

Here, $\partial\mathfrak{X}$ and $\partial\mathbf{X}_k$ denote the boundary of $\mathfrak{X}$ and $\mathbf{X}_k$, respectively.

2.2.1 The empirical setting

Assume now that we do not have explicit knowledge of the measure μ, but are instead given a sequence $\mathcal{X} \subseteq \mathbb{R}^n$ of N samples $X_1, X_2, \ldots, X_N \in \mathbb{R}^n$, drawn independently according to μ. These samples induce a probability measure $\mu_\mathcal{X}$ given by $\mu_\mathcal{X} = \frac{1}{N}\sum_{i=1}^N \delta_{X_i}$, which we call the emperical measure associated to $\mathcal{X}$. Under a non-degeneracy assumption (Assumption 13 below), the measure $\mu_\mathcal{X}$ induces an inner product of the form (5) on $\mathbb{R}[\mathbf{x}]_d$ by:

$$\langle p, q \rangle_{\mu_\mathcal{X}} := \int p(\mathbf{x})q(\mathbf{x})d\mu_\mathcal{X}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^N p(X_i)q(X_i). \tag{11}$$

In light of (11) and Proposition 9, it is straightforward to compute the Christoffel-Darboux kernel $K_d^{\mu_\mathcal{X}}$ of degree d for the measure $\mu_\mathcal{X}$ (and thus to compute $P_d^{\mu_\mathcal{X}}$). Indeed, the entries of the matrix $M_d^{\mu_\mathcal{X}}(\mathbf{b})$ in (9) may each be computed in time $O(N)$, after which $M_d^{\mu_\mathcal{X}}(\mathbf{b})$ can be inverted in time $O(s(n,d)^3)$.

▶ **Proposition 12.** *The empirical Christoffel-Darboux kernel $K_d^{\mu_\mathcal{X}}$ of degree d for N samples in $\mathbb{R}^n$ may be computed in time $O(Ns(n,d)^2 + s(n,d)^3)$.*

The above procedure only works when the matrix $M_d^{\mu_\mathcal{X}}(\mathbf{b})$ is invertible, or equivalently, when the “inner product” $\langle\cdot,\cdot\rangle_{\mu_\mathcal{X}}$ on $\mathbb{R}[\mathbf{x}]_d$ is definite. We shall make this assumption throughout.

▶ **Assumption 13.** *We say a sample set $\mathcal{X}$ is non-degenerate (up to degree d) if the inner product $\langle\cdot,\cdot\rangle_{\mu_\mathcal{X}}$ associated to the empirical measure $\mu_\mathcal{X}$ induced by $\mathcal{X}$ via* (11) *is definite for polynomials up to degree d.*

Assumption 13 is satisfied if and only if the samples $\mathcal{X}$ are not contained in an algebraic hypersurface of degree d (i.e., the zero set of a polynomial of degree at most d). This implies in particular that $N \geq s(n,d)+1$.

Under certain assumptions on μ, Lasserre and Pauwels [12] show that the (emperical) Christoffel polynomial $P_d^{\mu_\mathcal{X}}$ converges to the (population) Christoffel polynomial P_d^{μ} as the number of samples $N \to \infty$. The rate of this convergence can be quantified, see [22].

▶ **Theorem 14** ([12], see also [13]). *Let $\mathcal{X} = (X_1, X_2, \ldots, X_N)$ be sampled from μ as in the above. Then for each $\mathbf{x} \in [-1,1]^n$, we have $\lim_{N\to\infty} |P_d^{\mu}(\mathbf{x}) - P_d^{\mu_\mathcal{X}}(\mathbf{x})| = 0$ almost surely.*

3 A persistence module based on the Christoffel polynomial

Theorems 11 and 14 motivate the use of the Christoffel polynomial in (statistical) data analysis. They show that certain sublevel sets of the empirical Christoffel polynomial approximate the support of the underlying population measure μ well (in Hausdorff distance) as the number of samples grows. However, Theorem 11 gives very little explicit information on *which* (sub)level set to consider. This is the primary motivation for considering a persistent scheme instead, which we introduce now.

▶ **Definition 15.** *Fix $d \in \mathbb{N}$, and let $\mathcal{X} \subseteq [-1,1]^n$ be a set of samples whose associated empirical measure $\mu_\mathcal{X}$ satisfies Assumption 13. Let $P_d^{\mu_\mathcal{X}} : \mathbb{R}^n \to \mathbb{R}$ be the corresponding Christoffel polynomial* (10)*. For $t \geq 0$, we consider the compact sublevel set*

$$\mathbf{X}_t := \{\mathbf{x} \in [-1,1]^n : \log P_d^{\mu_\mathcal{X}}(\mathbf{x}) \leq t\} = \{\mathbf{x} \in [-1,1]^n : P_d^{\mu_\mathcal{X}}(\mathbf{x}) \leq 10^t\}, \tag{12}$$

which is well-defined as $P_d^{\mu_\mathcal{X}}(\mathbf{x}) \geq 1$ for all $\mathbf{x} \in \mathbb{R}^n$. By definition $(\mathbf{X}_t)_{t\geq 0}$ is a filtration, i.e, $\mathbf{X}_t \subseteq \mathbf{X}_{t'}$ for any $t \leq t'$. In light of Example 1, we can therefore define the persistence module

$$\mathbb{CD}(\mathcal{X}, d) := \mathrm{PH}_*([-1,1]^n,\ \log P_d^{\mu_\mathcal{X}}). \tag{13}$$

Notably, we do not consider the level sets of $P_d^{\mu_\mathcal{X}}$, but rather those of $\log P_d^{\mu_\mathcal{X}}$. Before we motivate this choice, let us first observe that from a computational perspective, this logarithmic rescaling makes no difference. Indeed, one can obtain the persistence module of the filtration $([-1,1]^n, \log P_d^{\mu_\mathcal{X}})$ by first computing the module associated to $([-1,1]^n, P_d^{\mu_\mathcal{X}})$ and then rescaling all interval modules. We choose to work with $\log P_d^{\mu_\mathcal{X}}$ for two reasons. First, as we will see below, this choice allows us to prove a stronger and more elegant stability result for the module $\mathbb{CD}(\mathcal{X}, d)$. Second, the logarithmic scaling produces persistence diagrams that better fit the underlying topology in practice. Proposition 10 provides a rather convincing theoretical argument for this observation. Indeed, if $\mathbf{x} \in [-1,1]^n$ is a point outside of the support of the underlying measure μ, it tells us that $P_d^{\mu}(\mathbf{x}) \approx \exp\big(\mathrm{dist}(\mathbf{x}, \mathrm{supp}(\mu)) \cdot d\big)$, which is to say that $\log P_d^{\mu}(\mathbf{x})$ scales *linearly* in the distance $\mathrm{dist}(\mathbf{x}, \mathrm{supp}(\mu))$, an intuitively desirable property.

3.1 Stability and robustness

In this section, we show that the module $\mathbb{CD}(\mathcal{X}, d)$ is *locally* stable under small perturbations of the sample set $\mathcal{X}$, measured in the Wasserstein distance (1). Namely, we show in Proposition 19 that:

$$d_B\big(\mathrm{Dgm}(\mathbb{CD}(\mathcal{X}, d)),\ \mathrm{Dgm}(\mathbb{CD}(\mathcal{Y}, d))\big) \leq \log\big(C_\mathcal{X} \cdot d_W(\mu_\mathcal{X}, \mu_\mathcal{Y}) + 1\big)$$

for fixed $\mathcal{X} \subseteq [-1,1]^n$ and any $\mathcal{Y} \subseteq [-1,1]^n$ for which $d_W(\mu_\mathcal{X}, \mu_\mathcal{Y})$ is sufficiently small. Here, the constant $C_\mathcal{X}$ depends on n, d, and the supremum norm

$$\| \log P_d^{\mu_\mathcal{X}} \|_\infty := \max_{\mathbf{x} \in [-1,1]^n} | \log P_d^{\mu_\mathcal{X}}(\mathbf{x})|,$$

which we interpret as a "measure of algebraic degeneracy" of the set $\mathcal{X}$ (see Assumption 13). For *arbitrary* sets $\mathcal{X}$ and $\mathcal{Y}$, we show in Proposition 18 that

$$d_B\big(\mathrm{Dgm}(\mathbb{CD}(\mathcal{X}, d)),\ \mathrm{Dgm}(\mathbb{CD}(\mathcal{Y}, d))\big) \leq \log\big(C_{\mathcal{X},\mathcal{Y}} \cdot d_W(\mu_\mathcal{X}, \mu_\mathcal{Y}) + 1\big), \tag{14}$$

where the constant $C_{\mathcal{X},\mathcal{Y}}$ now additionally depends on $\| \log P_d^{\mu_\mathcal{Y}} \|_\infty$. If one restricts to "sufficiently non-degenerate" sample sets (i.e., those having $\| \log P_d^{\mu_\mathcal{X}} \|_\infty$ bounded from above), relation (14) may be read as a *global* stability result.

For the proof of these statements, note first that in light of Corollary 4, we have:

$$d_B\big(\mathrm{Dgm}(\mathbb{CD}(\mathcal{X}, d)),\ \mathrm{Dgm}(\mathbb{CD}(\mathcal{Y}, d))\big) \leq \| \log P_d^{\mu_\mathcal{X}} - \log P_d^{\mu_\mathcal{Y}} \|_\infty \tag{15}$$

and so it suffices to consider the quantity $\| \log P_d^{\mu_\mathcal{X}} - \log P_d^{\mu_\mathcal{Y}} \|_\infty$. We start by showing the following.

▶ **Theorem 16.** *Let $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^n$ be as in the above. Write $C_{n,d} := 4 \cdot s(n,d) \cdot d^2$, where $s(n,d) := \dim \mathbb{R}[\mathbf{x}]_d = \binom{n+d}{d}$. Then for all $\mathbf{x} \in \mathbb{R}^n$, we have that:*

$$|P_d^{\mu_\mathcal{X}}(\mathbf{x}) - P_d^{\mu_\mathcal{Y}}(\mathbf{x})| \leq C_{n,d} \cdot \|P_d^{\mu_\mathcal{X}}\|_\infty \cdot d_W(\mu_\mathcal{X}, \mu_\mathcal{Y}) \cdot P_d^{\mu_\mathcal{Y}}(\mathbf{x}).$$

We prove Theorem 16 in [17, Appendix A] by adapting the techniques of Section 6.2 in [13] to our setting. This requires some small technical statements on Wasserstein distance and supremum norms of polynomials on compact domains, which we also give there.

We have the following immediate consequence:

▶ **Corollary 17.** *Let $\mathcal{X}, \mathcal{Y}$ as in the above, and let $C_{n,d} > 0$ be the constant of Theorem 16. Then for all $\mathbf{x} \in [-1,1]^n$, we have:*

$$|P_d^{\mu_\mathcal{X}}(\mathbf{x}) / P_d^{\mu_\mathcal{Y}}(\mathbf{x}) - 1| \leq C_{n,d} \cdot \|P_d^{\mu_\mathcal{X}}\|_\infty \cdot d_W(\mu_\mathcal{X}, \mu_\mathcal{Y}).$$

After taking logarithms (see [17, Appendix A] for details), we then obtain:

▶ **Proposition 18.** *Let $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{R}^n$ be as above, and let $C_{n,d} > 0$ be the constant of Theorem 16. Then we have:*

$$\| \log P_d^{\mu_\mathcal{X}} - \log P_d^{\mu_\mathcal{Y}} \|_\infty \leq \log\big(C_{n,d} \cdot \max\{\|P_d^{\mu_\mathcal{X}}\|_\infty, \|P_d^{\mu_\mathcal{Y}}\|_\infty\} \cdot d_W(\mu_\mathcal{X}, \mu_\mathcal{Y}) + 1\big). \tag{16}$$

Note that when the quantity $C := C_{n,d} \cdot \max\{\|P_d^{\mu_\mathcal{X}}\|_\infty, \|P_d^{\mu_\mathcal{Y}}\|_\infty\}$ is close to 0, the bound (16) tells us that $\| \log P_d^{\mu_\mathcal{X}} - \log P_d^{\mu_\mathcal{Y}} \|_\infty \leq C \cdot d_W(\mu_\mathcal{X}, \mu_\mathcal{Y})$. On the other hand, if $C \gg 1$, the bound on $\| \log(P_d^{\mu_\mathcal{X}}) - \log(P_d^{\mu_\mathcal{Y}}) \|_\infty$ is exponentially smaller than in Theorem 16. Finally, we may get rid of the dependence on $\mathcal{Y}$ in Corollary 17 after making an additional assumption.

▶ **Proposition 19.** *Let $\mathcal{X} \subseteq [-1,1]^n$ be a (fixed) sample set satisfying Assumption 13. Let $C_{n,d} > 0$ be the constant of Theorem 16, and assume that $\mathcal{Y} \subseteq [-1,1]^n$ is such that $C_{n,d} \cdot \|P_d^{\mu_\mathcal{X}}\|_\infty \cdot d_W(\mu_\mathcal{X}, \mu_\mathcal{Y}) \leq 1/2$. Then we have $\|P_d^{\mu_\mathcal{Y}}\|_\infty \leq 2\|P_d^{\mu_\mathcal{X}}\|_\infty$. In particular, the bound* (16) *then reads:*

$$\| \log P_d^{\mu_\mathcal{X}} - \log P_d^{\mu_\mathcal{Y}} \|_\infty \leq \log\big(C_{n,d} \cdot 2\|P_d^{\mu_\mathcal{X}}\|_\infty \cdot d_W(\mu_\mathcal{X}, \mu_\mathcal{Y}) + 1\big).$$

3.2 An exact algorithm for the persistence module

As we explain in Section 2.2 (see Proposition 12), the Christoffel polynomial $P_d^{\mu_\mathcal{X}}$ of degree d may be computed in time $O(s(n,d)^3 + Ns(n,d)^2)$, where $N = |\mathcal{X}|$ is the number of samples and $s(n,d) = \binom{n+d}{d}$ is the size of the moment matrix (9). Once we have access to $P_d^{\mu_\mathcal{X}}$, it remains to compute the persistent homology of the filtration $\{\mathbf{x} \in [-1,1]^n : P_d^{\mu_\mathcal{X}}(\mathbf{x}) \leq t\}_{t\geq 0}$. The set $[-1,1]^n$ is a particularly simple example of a basic (closed) *semialgebraic set*. That is, a subset of $\mathbb{R}^n$ defined by a finite number of polynomial (in)equalities; namely:

$$[-1,1]^n = \{\mathbf{x} \in \mathbb{R}^n : g_i(\mathbf{x}) := 1 - \mathbf{x}_i^2 \geq 0 \quad \text{for } i = 1, 2, \dots, n\}.$$

We may therefore use the following recent result of Basu and Karisani [1].

▶ **Theorem 20** (Basu, Karisani (2022)). *For fixed $p \in \mathbb{N}$, there is an algorithm that takes as input a description of a closed and bounded semialgebraic set $\mathbf{X} = \{\mathbf{x} \in \mathbb{R}^n : g_i(\mathbf{x}) \geq 0,\ i = 1, 2, \dots, m\}$, and a polynomial $P \in \mathbb{R}[\mathbf{x}]$, and outputs the persistence diagram associated to the filtration $\{\mathbf{x} \in \mathbf{X} : P(\mathbf{x}) \leq t\}_{t\geq 0}$ up to dimension p. The complexity of this algorithm is bounded by $(md)^{O(n)}$, where $d = \max\{\deg(g_i), \deg(P)\}$ is the largest degree amongst P and the polynomial inequalities defining $\mathbf{X}$.*

▶ **Corollary 21** (Exact algorithm). *For fixed $p \in \mathbb{N}$, there is an exact algorithm that computes the persistence diagram associated to $\mathbb{CD}(\mathcal{X}, d)$ up to dimension p. Its runtime is bounded by $O(s(n,d)^3 + Ns(n,d)^2) + (nd)^{O(n)}$.*

3.3 An effective approximation scheme

To the authors' knowledge, no implementation exists of the algorithm mentioned in Theorem 20 at the time of writing. In order to perform numerical experiments, we use a simple approximation scheme for $\mathbb{CD}(\mathcal{X}, d)$. This method works in the more general case of approximating $\mathrm{PH}_*([-1,1]^n, f)$ for any Lipschitz continuous function $f : [-1,1]^n \to \mathbb{R}$. Succinctly put, we first fix $m \in \mathbb{N}$, and construct the Freudenthal triangulation [10, 7] of $[-1,1]^n$, with vertices equal to the lattice points of $\frac{2}{m} \cdot \mathbb{Z}^n$ contained in $[-1,1]^n$. See Figure 3. We denote this triangulation by $\mathcal{K}_m$. Note that the diameter of any simplex in this triangulation is equal to $2\sqrt{n}/m$. We then evaluate f on each of the vertices, and compute the persistent homology of the lower-star filtration on $\mathcal{K}_m$ induced by these function values. This persistence module, denoted by by $\mathrm{PH}_*(\mathcal{K}_m, f)$, approximates $\mathrm{PH}_*([-1,1]^n, f)$. The diagram of this module can be computed in polynomial time in the number of lattice points, which in our case is $(m+1)^n$. The following proposition gives a guarantee on the quality of this approximation:

▶ **Proposition 22.** *Let $f : [-1,1]^n \to \mathbb{R}$ be a Lipschitz continuous function with Lipschitz constant L_f, choose $m \in \mathbb{N}$, and let $\mathcal{K}_m$ be as above. Then*

$$d_B(\mathrm{Dgm}(\mathrm{PH}_*([-1,1]^n, f)),\ \mathrm{Dgm}(\mathrm{PH}_*(\mathcal{K}_m, f))) \leq L_f \cdot 2\sqrt{n}/m.$$

Since the function $\log P_d^{\mu_\mathcal{X}} : [-1,1]^n \to \mathbb{R}$ is differentiable and $[-1,1]^n$ is compact, $\log P_d^{\mu_\mathcal{X}}$ is Lipschitz continuous, and so the above proposition applies to our setting. For a proof and a more detailed discussion on this approximation scheme, we refer to [17, Appendix B].

▶ Remark 23. The stability results (14), (15) apply directly to $\mathrm{PH}_*(\mathcal{K}_m, P_d^{\mu_\mathcal{X}})$. That is, the approximated modules are also stable in the Wasserstein distance, see [17, Appendix B].

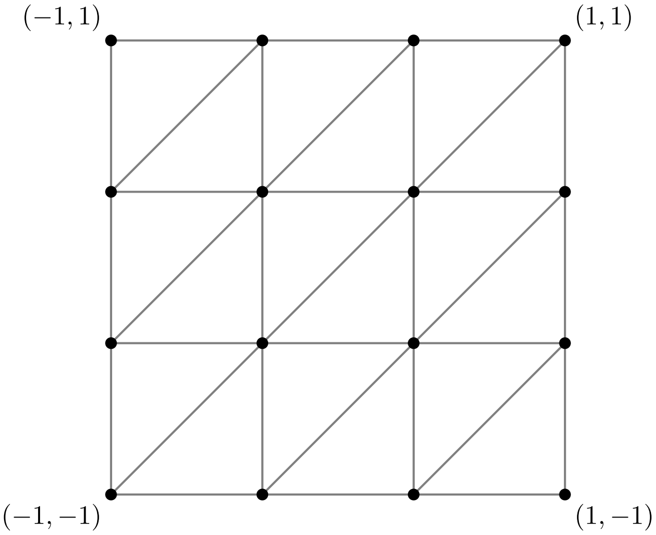


Figure 3 The Freudenthal triangulation of $[-1,1]^2$ on the lattice $\frac{2}{m}\cdot\mathbb{Z}^2$ (with $m=3$).

4 Numerical examples

In this section we illustrate our new scheme by computing the persistence diagram of $\mathbb{CD}(\mathcal{X},d)$ for various toy examples $\mathcal{X}\subseteq[-1,1]^n$ for $n=1,2,3$. In each case, the sets $\mathcal{X}$ are drawn from a measure supported $[-1,1]^n$, after which some additional noise may be added. We compute the Christoffel polynomials using the method outlined in Section 2.2.1 and the linear algebra packages of NumPy [11] and Scipy [21]. To *approximate* the persistence of the resulting filtrations, we employ the method of Section 3.3, where we set the resolution $m=250$ for $n=1,2$ and $m=50$ for $n=3$. See [17, Appendix B]. The persistence of the resulting lower-star filtration is computed using Gudhi [20]. We also use Gudhi to compute the (exact) persistent homology of the Čech filtration.

We add noise to our data sets $\mathcal{X}$ in two ways. We say we add *uniform noise* when the noisy data $\tilde{\mathcal{X}}$ is obtained from $\mathcal{X}$ by adding M points chosen uniformly at random from $[-1,1]^n$. We say we add *Gaussian noise* (with standard deviation $\sigma\geq 0$), when $\tilde{\mathcal{X}}$ is obtained from $\mathcal{X}$ by adding to each coordinate of each sample $\mathbf{x}\in\mathcal{X}$ an independently drawn perturbation $t\sim N(0,\sigma)$.

A univariate example. It is rather instructive to consider first a simple univariate example. Let $\mathcal{X}\subseteq[-1,1]$ be drawn from a uniform measure μ supported on five disjoint intervals $I_1,I_2,\ldots,I_5\subseteq[-1,1]$. The corresponding Christoffel polynomials ($d=4,8,12$) and persistence diagrams for this situation are plotted in Figure 4. We would expect the persistence diagram of $\mathbb{CD}(\mathcal{X},d)$ to reflect the simple topology of $\mathrm{supp}(\mu)$; namely we expect $\mathbb{CD}(\mathcal{X},d)$ to consist of five interval modules, each corresponding to one of the connected components of $\mathrm{supp}(\mu)$. For $d=8,12$, we observe that this is indeed the case. For $d=4$, however, we see there are only four interval modules. In the one-dimensional setting, this can be explained rather nicely. Indeed, "birth-events" correspond to (local) *minima* of $P_d^{\mu_{\mathcal{X}}}$, and "death-events" correspond to (local) *maxima*. As $P_4^{\mu_{\mathcal{X}}}$ is of degree 8, it can have at most 7 critical points, resulting in four interval modules (one of which is of infinite length).

The figure eight. In Figure 5, we plot $\log P_{12}^{\mu_{\mathcal{X}}}$ and $\mathbb{CD}(\mathcal{X}, 12)$ when $\mathcal{X}$ is drawn from a measure supported on two circles in three different configurations: disjoint, just intersecting, and overlapping. We observe that $\mathbb{CD}(\mathcal{X}, 12)$ correctly captures the underlying topology in all three cases (for clarity, the number of overlapping points in the diagram is indicated).

Feature sizing. We consider the influence of feature size in the data on our persistence module. We draw samples $\mathcal{X}$ from a configuration of two circles in $[-1, 1]^n$. In Figure 6, we depict the diagram of $\mathbb{CD}(\mathcal{X}, 10)$ as we decrease the *radius* of one of the circles. Similarly, in Figure 7, we depict the diagram of $\mathbb{CD}(\mathcal{X}, 10)$ as we decrease the *number of samples* drawn from one of the circles. In both cases, the decrease in feature size corresponds to a decrease in the length of the corresponding interval in $\mathbb{CD}(\mathcal{X}, 10)$. Interestingly, this is due to an earlier time of death in the former case, and due to a later time of birth in the latter.

Comparison to the Čech module. We compare our new module $\mathbb{CD}(\mathcal{X}, 12)$ to Čech persistence in Figure 8, where $\mathcal{X}$ is drawn from a measure supported on a ball, a triangle, and a square ($N = 10000$). We consider four cases: a pure sample $\mathcal{X}$, two samples with uniform noise ($M = 50$ and $M = 2500$) in $[-1, 1]^n$, and a sample with Gaussian noise ($\sigma = 0.03$). We observe that both the Christoffel-Darboux and Čech module are able to correctly capture the underlying topology for the pure sample and for the sample with Gaussian noise. However, only the Christoffel-Darboux module is able to do so for the samples with uniform noise, whereas the Čech module recovers no meaningful information there.

Stability under uniform noise. We consider a set $\mathcal{X}$ consisting of evenly spaced points on the 1-skeleton of a cube[1] in $\mathbb{R}^3$ (50 points per edge) with edge-length 1.5. The significant persistent features of $\mathcal{X}$ consist of a single interval (of infinite length) in H_0; five intervals in H_1 and one interval in H_2, see Figure 9. We investigate how well the features in H_1 can be recovered after adding an increasing amount of uniform noise to $\mathcal{X}$, comparing Čech persistence to $\mathbb{CD}(\mathcal{X}, d)$, $d = 6, 8, 10$. To measure this, we follow [3] and use the *signal-to-noise ratio*; meaning the ratio between the size of the smallest interval in H_1 inherent to $\mathcal{X}$ and the size of the largest interval (in H_1) induced by the noise. In Table 1, we report the median signal-to-noise ratios over 100 experiments. We reiterate that the Christoffel-Darboux persistence is computed *approximately*, which could affect these results. See [17, Appendix B] for a more detailed discussion. We observe that the ratios for the Christoffel-Darboux modules are much better than those for the Čech module. Furthermore, we note that the module of degree $d = 6$ outperforms the modules of degree $d = 8$ and $d = 10$. This is consistent with our stability results in Section 3.1, which are stronger for small values of d.

Table 1 Signal-to-noise ratios for persistent homology in dimension 1 for the cube skeleton in the presence of uniform noise (median values over 100 experiments). See also Figure 9.

uniform noise (M)	**baseline**	**25**	**50**	**100**	**250**	**500**	**1000**
Čech	$\gg 10$	3.6	2.4	2.0	1.6	1.4	1.3
Christoffel-Darboux ($d = 6$)	$\gg 10$	7.8	5.7	5.3	3.7	2.3	1.2
Christoffel-Darboux ($d = 8$)	$\gg 10$	4.9	2.8	2.6	2.5	2.1	1.2
Christoffel-Darboux ($d = 10$)	$\gg 10$	8.8	4.0	2.3	1.9	1.8	1.3

[1] Because the cube skeleton is degenerate, we add a small amount of Gaussian noise ($\sigma = 0.025$) to $\mathcal{X}$ to ensure the Christoffel polynomial for $\mu_{\mathcal{X}}$ is well-defined.

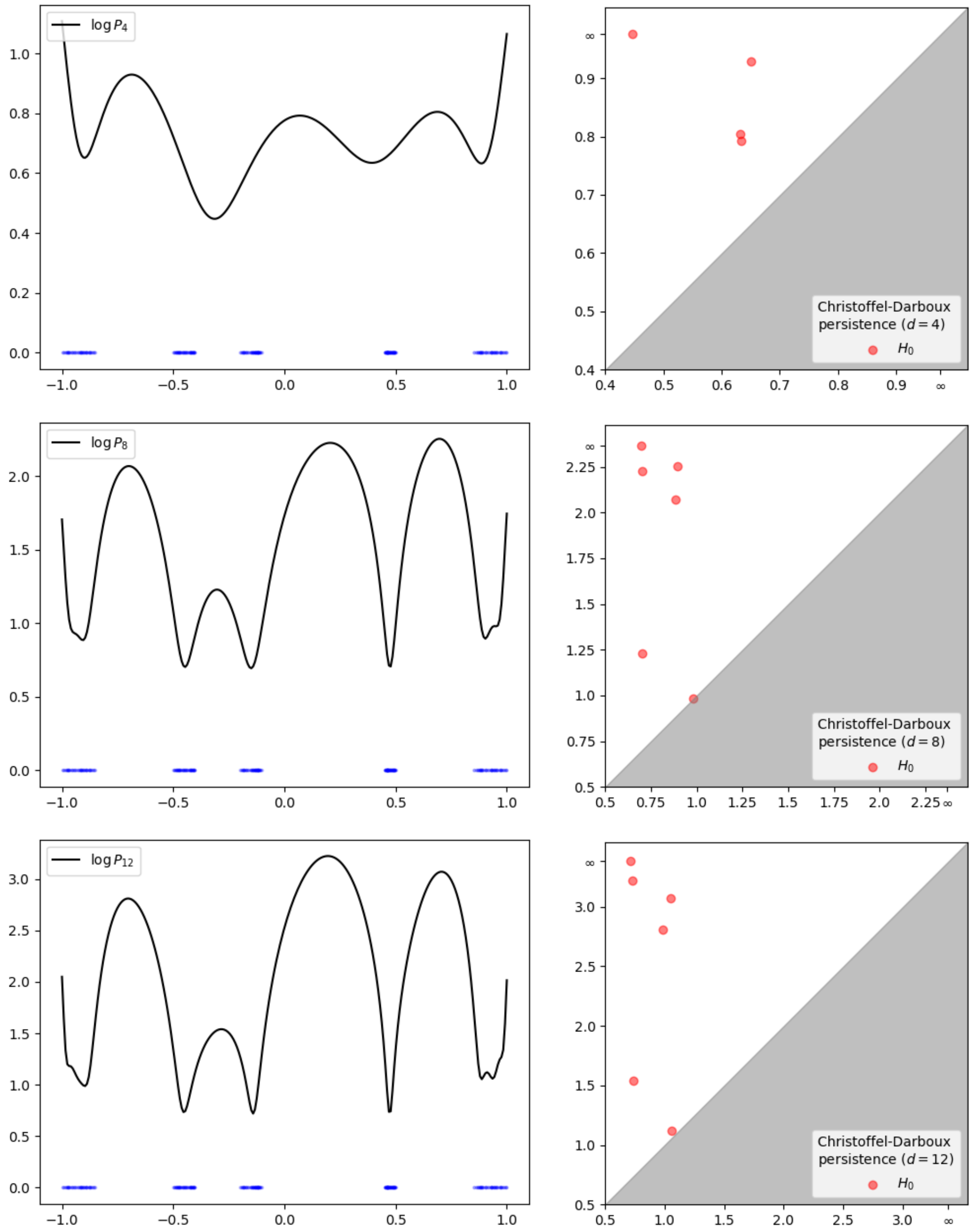


Figure 4 Left: Christoffel polynomials ($d = 4, 8, 16$) for samples $\mathcal{X} \subseteq [-1, 1]$ drawn from a measure supported on five intervals in $[-1, 1]$ (in blue, $N = 500$). Right: diagrams of $\mathbb{CD}(\mathcal{X}, d)$.

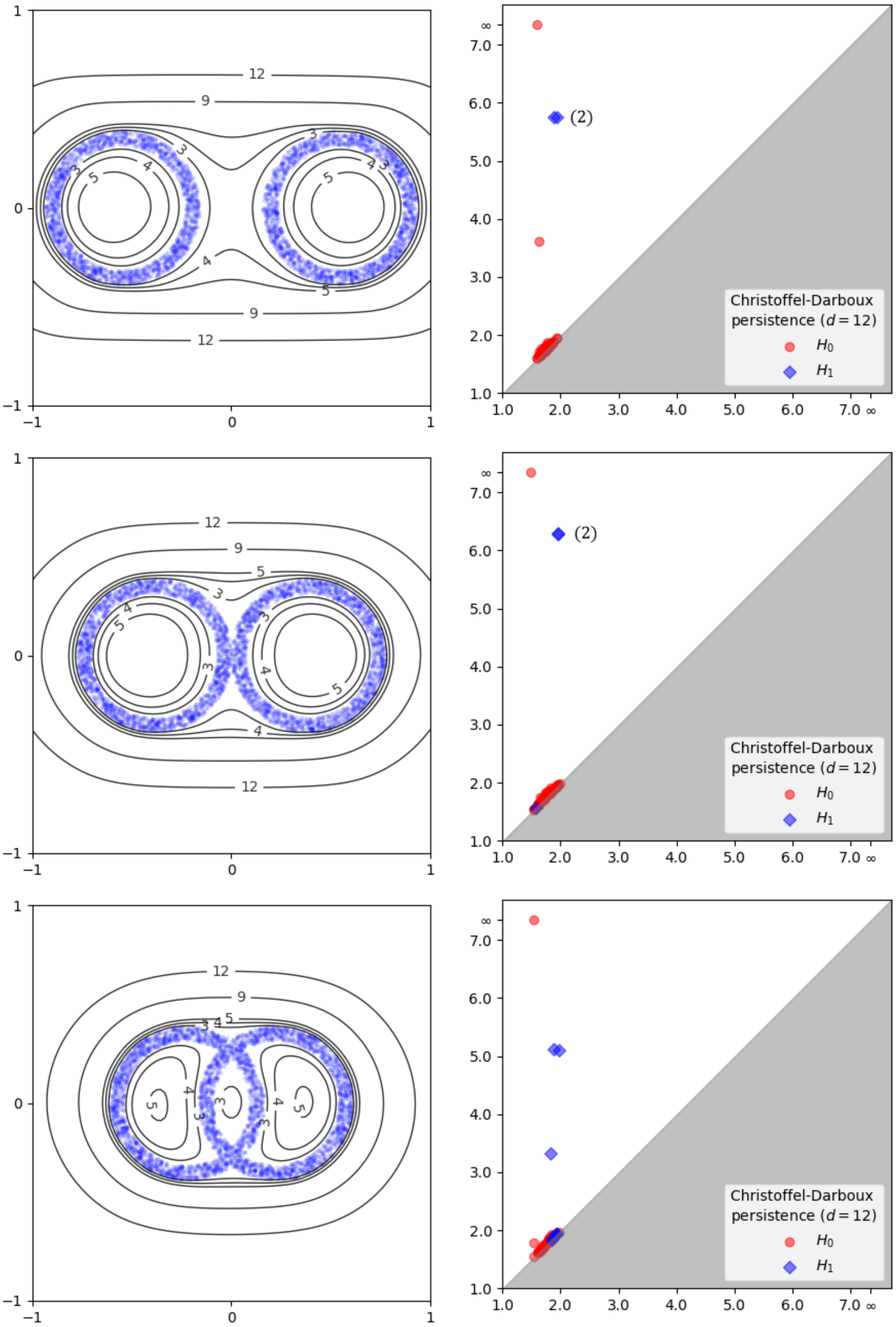


Figure 5 Left: level sets (in black) of the Christoffel polynomial $P_{12}^{\mu,\mathcal{X}}$ of degree 12 for three sets of samples $\mathcal{X} \subseteq [-1,1]^n$ (in blue, $N = 3000$). Right: the corresponding diagrams of $\mathbb{CD}(\mathcal{X}, 12)$.

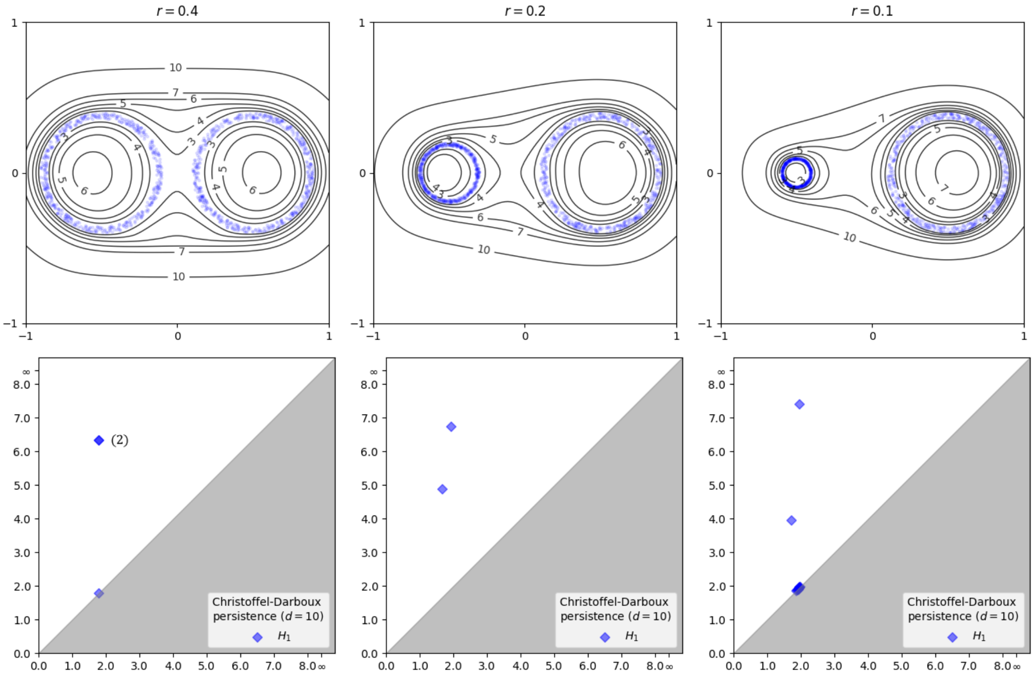


Figure 6 Level sets of $\log P_{10}^{\mu,\mathcal{X}}$ and diagrams of $\mathbb{CD}(\mathcal{X}, 10)$ for samples $\mathcal{X}$ drawn from a measure supported on two circles. The radius r of the left circle decreases. Only degree 1 homology is shown.

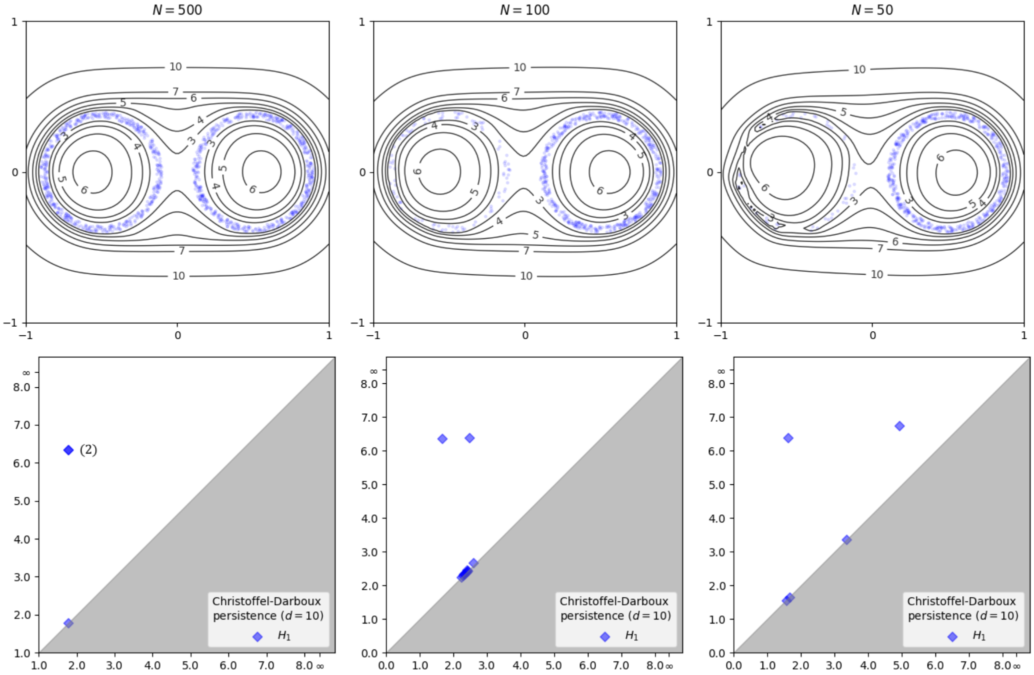


Figure 7 Level sets of $\log P_{10}^{\mu,\mathcal{X}}$ and diagrams of $\mathbb{CD}(\mathcal{X}, 10)$ for samples $\mathcal{X}$ drawn from a measure supported on two circles. The number of samples N drawn from the left circle decreases. Only degree 1 homology is shown.

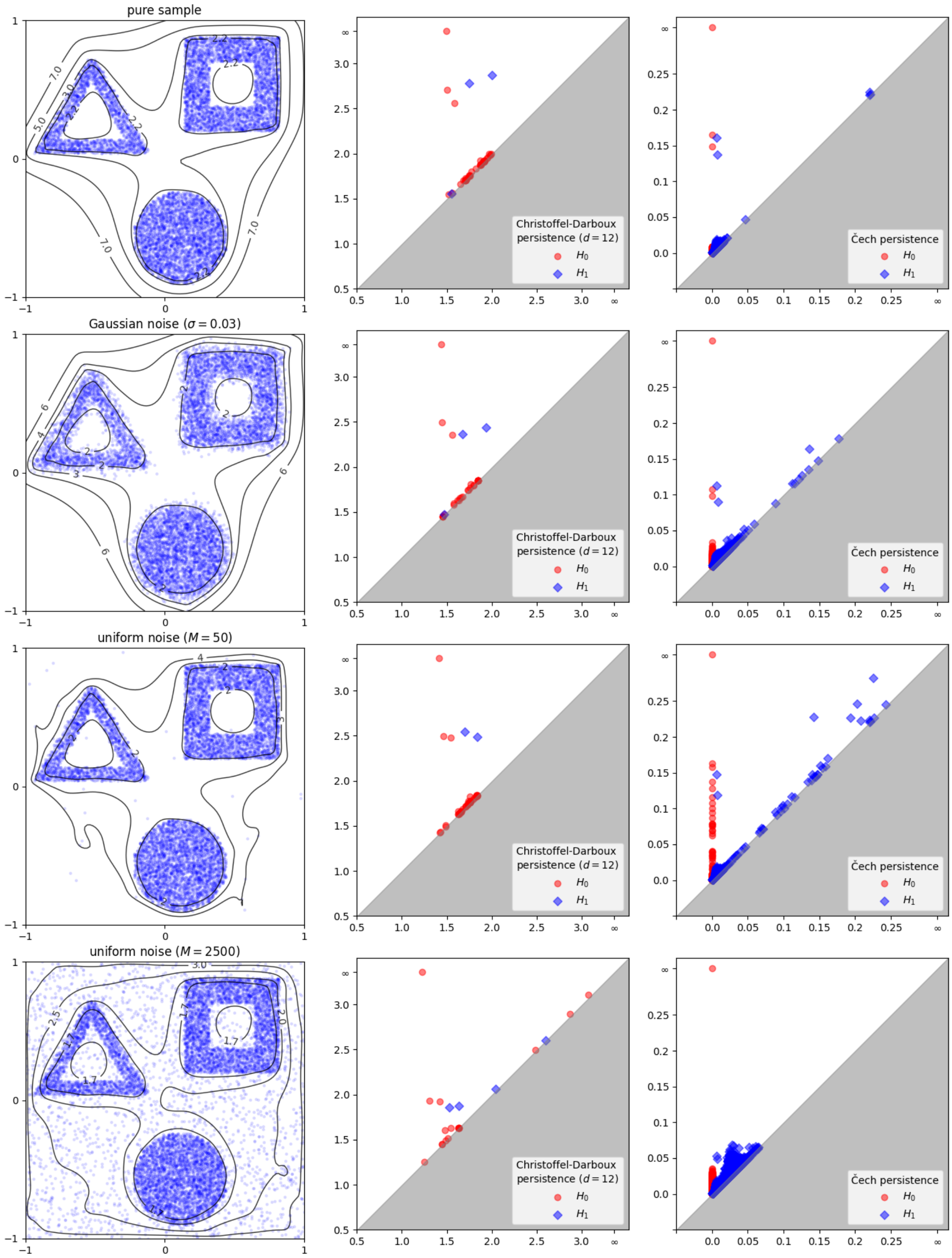


Figure 8 Left: level sets of $\log P_{12}^{\mu_{\mathcal{X}}}$ for sample sets $\mathcal{X}$ in $[-1,1]^2$ with different types of noise. Center: persistence diagrams of $\mathbb{CD}(\mathcal{X},12)$. Right: persistence diagrams for the Čech filtration.

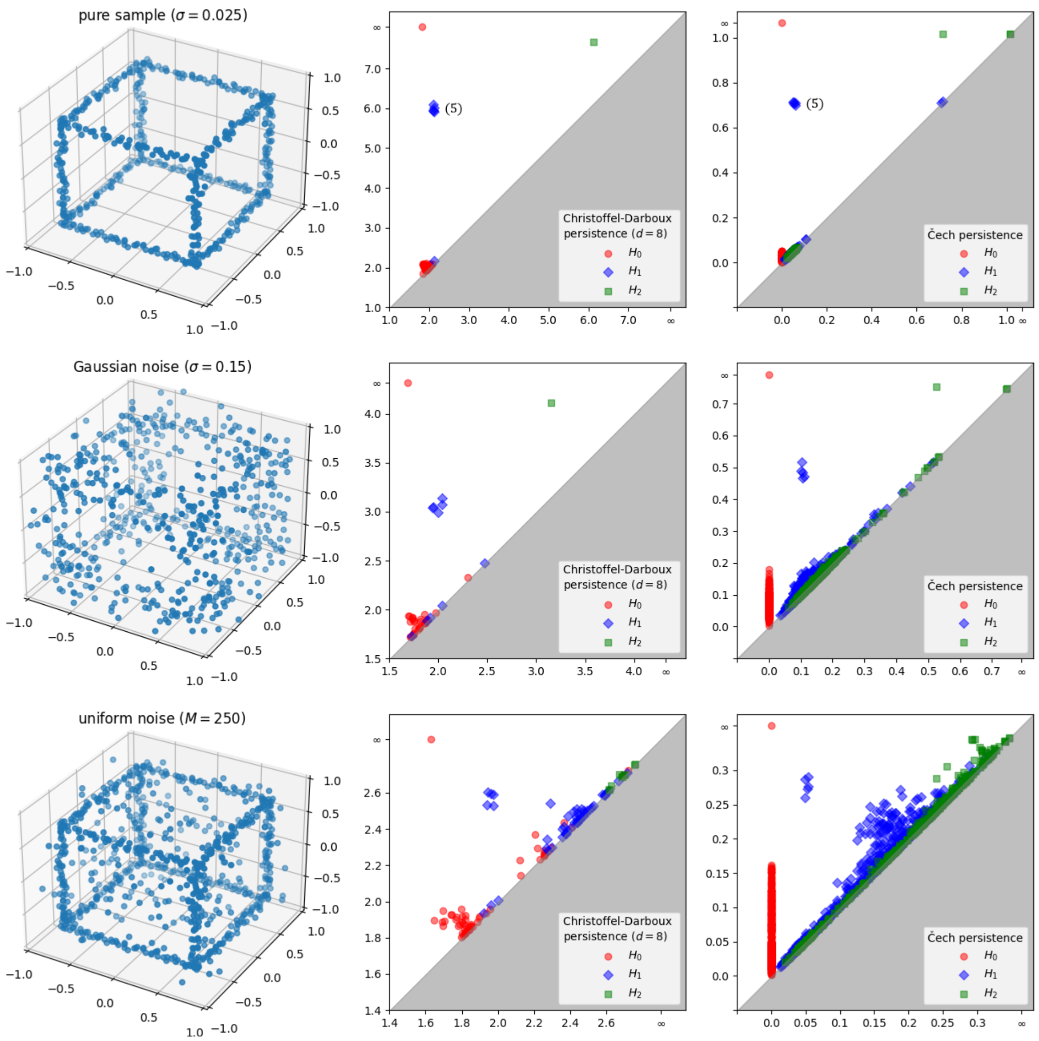


Figure 9 Left: sample sets obtained from the 1-skeleton of a cube in $[-1,1]^3$ by adding Gaussian and uniform noise, respectively. Right: the corresponding Čech and Christoffel-Darboux persistence.

5 Discussion

We have introduced a new scheme for computing persistent homology of a point cloud in $\mathbb{R}^n$, based on the theory of Christoffel-Darboux kernels. Our scheme is stable w.r.t. the Wasserstein distance. It admits an exact algorithm whose runtime is linear in the number of samples, but depends rather heavily on the ambient dimension n and the degree d of the kernel. In several examples ($n = 1, 2, 3$), it was able to capture key topological features of the point cloud, even in the presence of uniform noise.

Computing the persistent homology. The persistence module $\mathbb{CD}(\mathcal{X}, d)$ arises from a particularly simple filtration of a semialgebraic set by a polynomial. This is what allows us to invoke the result of Basu & Karisani in Section 3.2 to compute its persistence diagram. Their result in fact applies in a much more general setting, but no practical implementation of the resulting algorithm exists. Our present work thus motivates the search for *effective* exact algorithms in simple special cases. We instead rely in this work on the approximation scheme described in Section 3.3. We are only able to bound the error of this scheme in terms of the Lipschitz constant of $P_d^{\mu_\mathcal{X}}$ (which may be large, see also below). The practical performance of our scheme appears to be much better than this bound would suggest. It would therefore be desirable to prove further theoretical results that back this up.

Regularization. Our stability results of Section 3.1 depend on the "algebraic degeneracy" of the sample set $\mathcal{X}$. Such dependence is undesirable, and not present in stability results for most conventional persistence modules. This dependence can potentially be avoided by considering a *regularization* of the Christoffel polynomial, obtained by adding a small multiple of the identity to the moment matrix (9): $M \leftarrow M + \varepsilon \cdot \mathrm{Id}$. One can also think of this as adding a small multiple of the uniform measure on $[-1,1]^n$ to the empirical measure $\mu_\mathcal{X}$, which ensures that the corresponding inner product is definite. In [14], the authors already studied the impact of such modifications in the setting of functional approximation. There, it allows them to work over (near-)degenerate sets $\mathcal{X}$, while still accurately capturing the geometry of their problem. Preliminary experiments show this approach can be applied in our setting as well, and it would be very interesting to explore this further.

Selecting the degree *d*. Another important consideration is the selection of the degree d of the Christoffel polynomial $P_d^{\mu_\mathcal{X}}$. On the one hand, Proposition 10 and Theorem 11 suggest that the polynomial captures the support of the underlying measure μ better when d is large. On the other hand, Proposition 18 and Table 1 suggest that $\mathbb{CD}(\mathcal{X}, d)$ is more stable under perturbations of the data $\mathcal{X}$ for smaller d. Furthermore, computing $P_d^{\mu_\mathcal{X}}$ rapidly becomes more costly as d grows. It is a hard open question what the "optimal" choice of d is w.r.t. n.

References

1 Saugata Basu and Negin Karisani. Persistent homology of semi-algebraic sets, 2022. URL: https://arxiv.org/abs/2202.09591.

2 Ulrich Bauer, Michael Kerber, Fabian Roll, and Alexander Rolle. A unified view on the functorial nerve theorem and its variations, 2022. URL: https://arxiv.org/abs/2203.03571.

3 Mickaël Buchet, Frédéric Chazal, Steve Y. Oudot, and Donald R. Sheehy. Efficient and robust persistent homology for measures. *Computational Geometry*, 58:70–96, 2016.

4 Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. Geometric inference for probability measures. *Foundations of Computational Mathematics*, 11:733–751, 2011.

5 Frédéric Chazal and Steve Oudot. Towards persistence-based reconstruction in euclidean spaces. *Proceedings of the Annual Symposium on Computational Geometry*, 2008. doi:10.1145/1377676.1377719.

6 David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37:103–120, 2007.

7 B.C. Eaves. *A Course in Triangulations for Solving Equations with Deformations*. Lecture notes in economics and mathematical systems. Springer-Verlag, 1984.

8 Herbert Edelsbrunner and John Harer. Persistent homology—a survey. *Discrete & Computational Geometry - DCG*, 453, 2008. doi:10.1090/conm/453/08802.

9 Herbert Edelsbrunner and Ernst Mucke. Three-dimensional alpha shapes. *ACM Transactions on Graphics*, 13, 1994. doi:10.1145/147130.147153.

10 Hans Freudenthal. Simplizialzerlegungen von beschrankter flachheit. *Annals of Mathematics*, 43:580, 1942.

11 Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert

Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

12 Jean-Bernard Lasserre and Edouard Pauwels. The empirical Christoffel function with applications in data analysis. *Advances in Computational Mathematics*, 45(3):1439–1468, 2019.

13 Jean-Bernard Lasserre, Edouard Pauwels, and Mihai Putinar. *The Christoffel–Darboux Kernel for Data Analysis*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2022. doi:10.1017/9781108937078.

14 Swann Marx, Edouard Pauwels, Tillmann Weisser, Didier Henrion, and Jean-Bernard Lasserre. Semi-algebraic approximation using Christoffel–Darboux kernel. *Constructive Approximation*, 54:391–429, 2021.

15 Edouard Pauwels, Mihai Putinar, and Jean-Bernard Lasserre. Data analysis from empirical moments and the Christoffel function. *Foundations of Computational Mathematics*, 21(1):243–273, 2021. doi:10.1007/s10208-020-09451-2.

16 Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

17 Pepijn Roos Hoefgeest and Lucas Slot. The Christoffel-Darboux kernel for topological data analysis (full version). URL: https://arxiv.org/abs/2211.15489.

18 Vin Silva and Gunnar Carlsson. Topological estimation using witness complexes. *Proc. Sympos. Point-Based Graphics*, 2004. doi:10.2312/SPBG/SPBG04/157-166.

19 Bernadette J. Stolz. Outlier-robust subsampling techniques for persistent homology, 2021. URL: https://arxiv.org/abs/2103.14743.

20 The GUDHI Project. *GUDHI User and Reference Manual*. GUDHI Editorial Board, 2015. URL: http://gudhi.gforge.inria.fr/doc/latest/.

21 Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

22 Mai Trang Vu, François Bachoc, and Edouard Pauwels. Rate of convergence for geometric inference based on the empirical Christoffel function. *ESAIM: PS*, 26:171–207, 2022.

23 Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete and Computational Geometry*, 33:249–274, 2005.