A Framework for Fast and Stable Representations of Multiparameter Persistent Homology Decompositions

David Loiseaux,
Mathieu Carrière, Andrew J. Blumberg

Python package

Multipers :
https://github.com/DavidLapous/multipers

GUDHI Geometry Understanding in Higher Dimensions

References

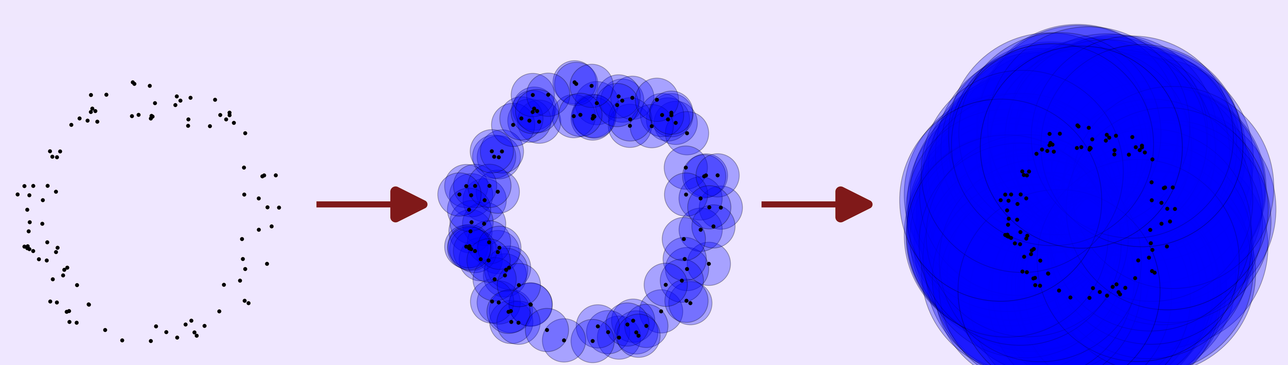
[1] F. Chazal, V. de Silva, M. Glisse, S. Oudot. The Structure and Stability of Persistence Modules
[2] M. B. Botnan, M. Lesnick. An introduction to multiparameter persistence
[3] D. Loiseaux, M. Carrière, A. Blumberg. Fast, Stable and Efficient Approximation of Multi-parameter Persistence Modules with MMA
[4] M. Carrière, A. Blumberg. Multiparameter Persistence Images for Topological Machine Learning
[5] A. Blumberg, M. Lesnick. Stability of 2-parameter persistent homology.

Abstract

The huge variety of dataset in the wild has brought many difficulties from a statistical point of view, such as the so-called « curse of the dimensionality ». Fortunately, data sets usually lie close to some hidden structure; which, if taken into account in the learning pipeline, can help mitigate this effect.
Topological Data Analysis (TDA) is a strategy that aims for a solution to this challenge, by providing compact descriptors inferring the topological features of this hidden structure, such as connectivity, loops, cavities; with nice guarantees. However, these main descriptors, the persistent modules, still suffers from some technical limitations; for instance, in computational biology, there are, in some cases no canonical way to compute them, as the input contains too much information.
This motivates their generalization : Multiparameter Persistence Modules.
The price to pay for this generalization is their computational cost. By leveraging on recent approximation technics, we propose a general framework, that take into account the majority of already known multiparameter persistent representations as well as a new powerful family of representations, for multiparameter topological machine learning pipelines.

Idea

A standard goal is to recover « topological features » from a dataset.
E.g., from points sampled on a circle, we want to retrieve this circle.
On this example, we look at the offsets of the dataset.



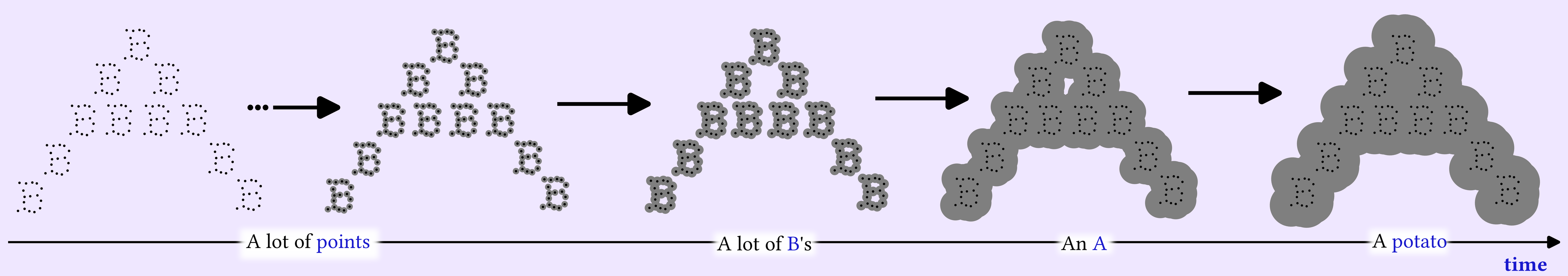
The circle «lives » on this time interval !

birth time — death time

This construction can be generalized to « filter functions »;
we look at the topology of the sublevelsets
$(\{x \in \mathcal{X} \mid f(x) \leq t\})_{t\in\mathbb{R}}$ for a function $f : \mathcal{X} \to \mathbb{R}$

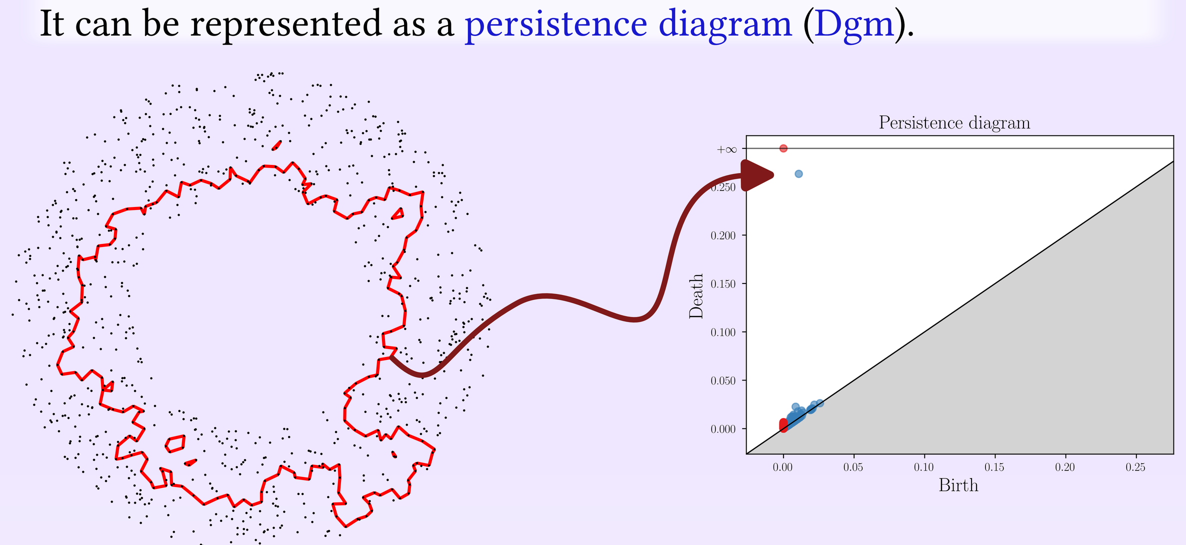
In that context, « growing balls » ≈ take the sublevelsets of the distance function to the dataset.

All scales at once !



Byproducts of TDA

The mathematical structure behind is the persistent homology (PH), which encodes birth and death time of each topological feature.
It can be represented as a persistence diagram (Dgm).



Nice properties

Universality

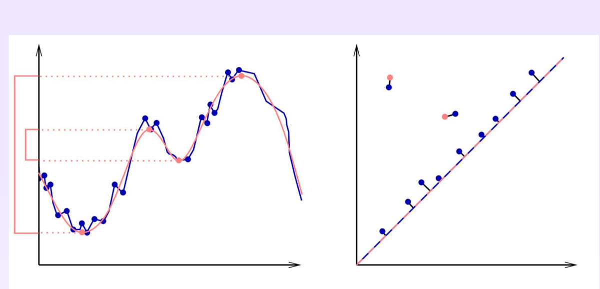
Any dataset having topological or geometrical signal can be used in this pipeline, and the output has always the same diagram structure.

Convergence

If $X^{(n)} = (X_1, \ldots, X_n) \sim \mu^{\otimes n}$ is a nice sampling of a space X_μ,
Then $\mathrm{Dgm}\left(X^{(n)}\right) \xrightarrow[n\to\infty]{d_b} \mathrm{Dgm}\left(X_\mu\right)$

Stability

For two functions $f, g : \mathcal{X} \to \mathbb{R}$,
$d_b\left(\mathrm{Dgm}(f), \mathrm{Dgm}(g)\right) \leq \|f - g\|_\infty$.

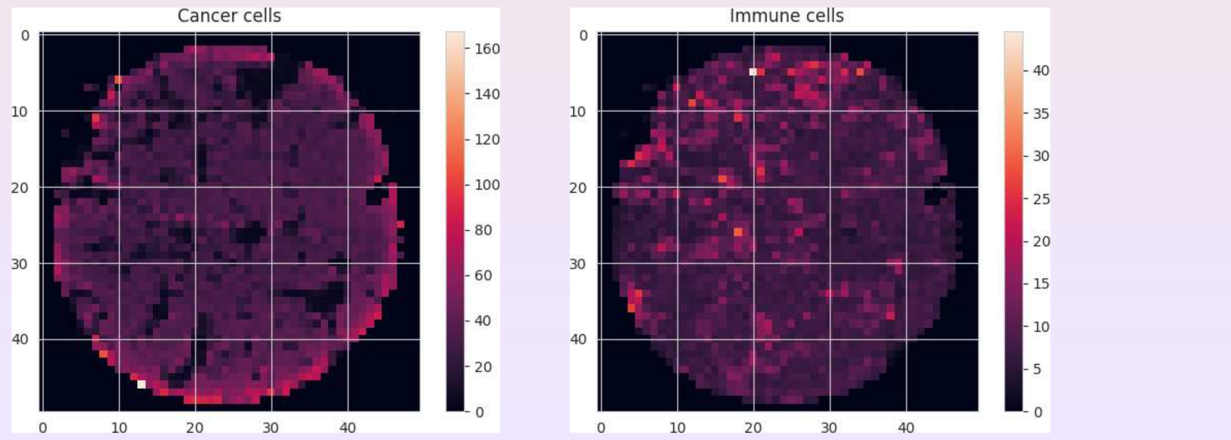


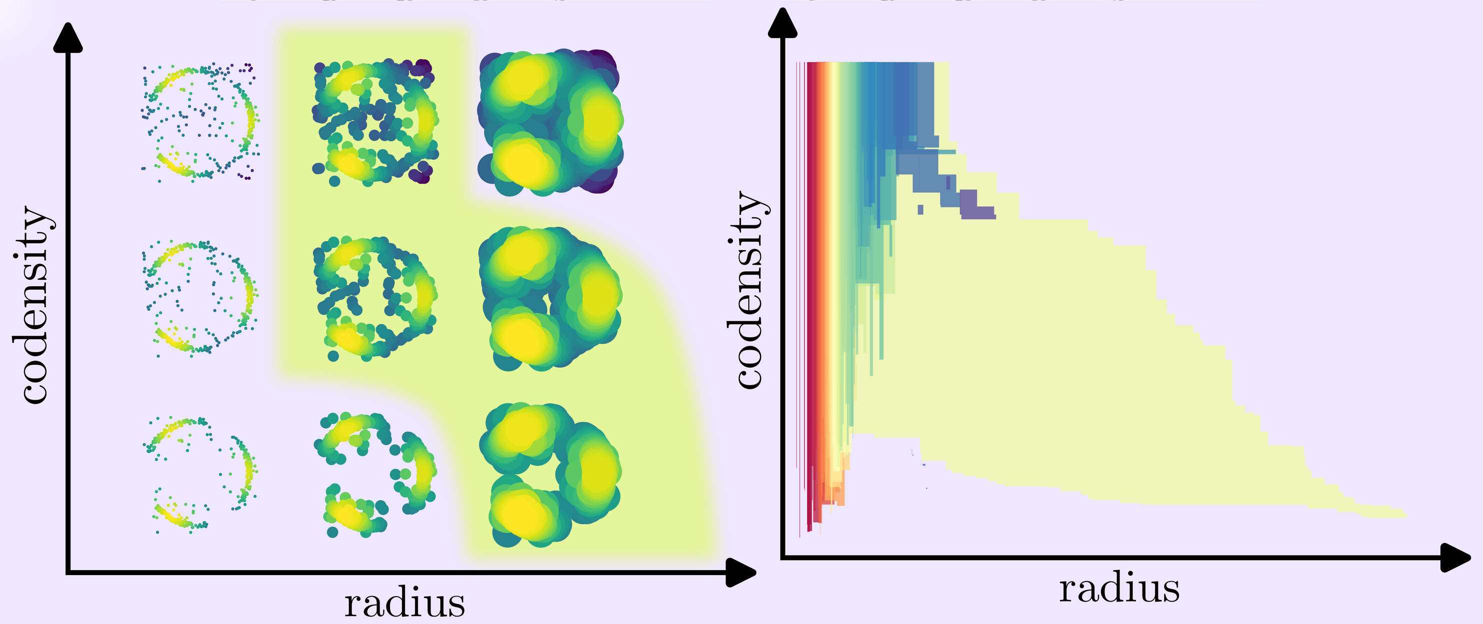
Multiparameter Persistence

Some datasets contain more than geometric information, or have an interesting sampling measure, which will not be taken into account in with PH. This motivates the construction of Multiparameter Persistent Homology (MPH) which looks at the topological persistence of a multi-filtered function $f : \mathcal{X} \to \mathbb{R}^n$.

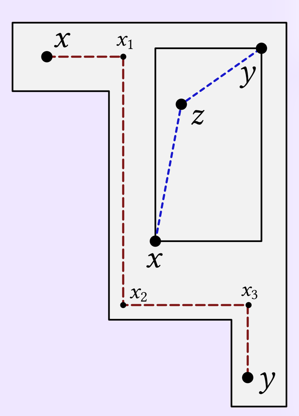
A Multiparameter Persistent Module M or an n-parameter persistent module is a family of vector spaces $(M_x)_{x\in\mathbb{R}^n}$ with some linear maps $M(x \leq y) : M_x \to M_y$ for $x \leq y \in \mathbb{R}^n$, satisfying

$$\forall x \leq y \leq z \in \mathbb{R}^n, \quad M(y \leq z) \circ M(x \leq y) = M(x \leq z) \quad \text{and} \quad M(x \leq x) = \mathrm{id}$$





Interval decomposition



Multiparameter Persistent Modules have, in general, a very complex structure. In order to simplificate our problem, we use previous work that approximate modules with interval decomposable modules.

An interval is a module that is convex and connected, i.e.,

- $\forall x \in \mathbb{R}^n, I_x \cong \mathbb{k}$ or $I_x \cong \{0\}$,
- $\forall x \leq y \in \mathbb{R}^n, \quad I_x \cong I_y \cong \mathbb{k} \implies I_x \to I_y = \mathrm{id}_{\mathbb{k}}$
- $\forall x \leq y \in \mathbb{R}^n, \quad I_x \cong I_y \cong \mathbb{k} \implies \forall x \leq z \leq y, I_z \cong \mathbb{k}$
- $\forall x, y \in \mathbb{R}^n, \quad I_x \cong I_y \cong \mathbb{k} \implies \exists x = x_0 \leq x_1 \geq \cdots \leq x_m = y$, satisfying $I_{x_1} \cong I_{x_2} \cong \cdots \cong I_{x_m} \cong \mathbb{k}$

An interval decomposable module is a module that can be written as a direct sum of interval modules, i.e.,
M is interval decomposable if, for some family of interval modules $\mathcal{I}$,

$$M \cong \bigoplus_{I\in\mathcal{I}} I$$

General Framework for Decomposition Representation

Given an interval decomposable module

$$M = \bigoplus_{1\leq i\leq m} M_i$$

One can consider

$$V_{\mathrm{op},w,\phi}(M) = \mathrm{op}(\{w(M_i)\cdot\phi(M_i)\}_{i=1}^m),$$

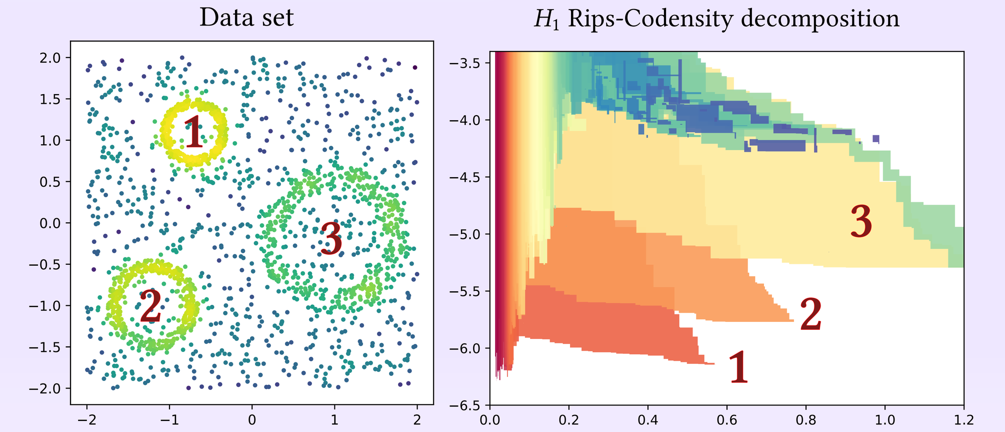
- op is a permutation invariant operation, e.g., sum, mean, max, min
- $w : \mathcal{M} \to \mathbb{R}$ is a weight function, and
- $\phi : \mathcal{M} \to \mathcal{H}$ is a kernel embedding.

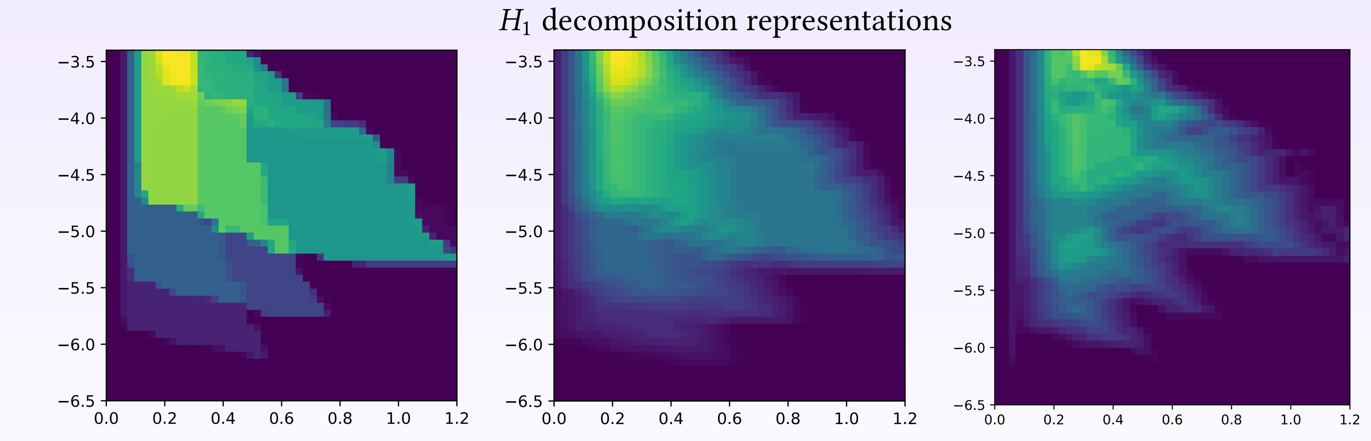
Now, considering stable functions, e.g.,

- $w : I \in \mathcal{I} \mapsto d_I(I, 0) \in \mathbb{R}$
- $\phi_\delta(M) : x \in \mathbb{R}^n \mapsto d_I\left(M|_{x+\delta K}, 0\right) \in \mathbb{R}$

where $K \subseteq B_{\|\cdot\|_\infty}(\mathbf{0}, 1) \subseteq \mathbb{R}^n$ is an interval containing $\mathbf{0}$,

$$V_{p,\delta}(M) := \sum_{i=1}^m \frac{w(M_i)^p}{\sum_{j=1}^m w(M_j)^p}\phi_\delta(M_i), \qquad V_{\infty,\delta}(M) := \sup_{1\leq i\leq m} \phi_\delta(M_i)$$





Stability result

Let $M = \oplus_{i=1}^m M_i$ and $M' = \oplus_{j=1}^{m'} M'_j$ be two interval decompositions.
Assume thatwe have $\frac{1}{m}\sum_i w(M_i), \frac{1}{m'}\sum_j w(M'_j) \geq C$, for some $C > 0$.
Then for any $\delta > 0$, one has

$$\left\|V_{0,\delta}(M) - V_{0,\delta}(M')\right\|_\infty \leq 2(d_b(M, M') \wedge \delta)/\delta,$$
$$\left\|V_{1,\delta}(M) - V_{1,\delta}(M')\right\|_\infty \leq \left[4 + \frac{2}{C}\right](d_b(M, M') \wedge \delta)/\delta,$$
$$\left\|V_{\infty,\delta}(M) - V_{\infty,\delta}(M')\right\|_\infty \leq (d_I(M, M') \wedge \delta)/\delta.$$

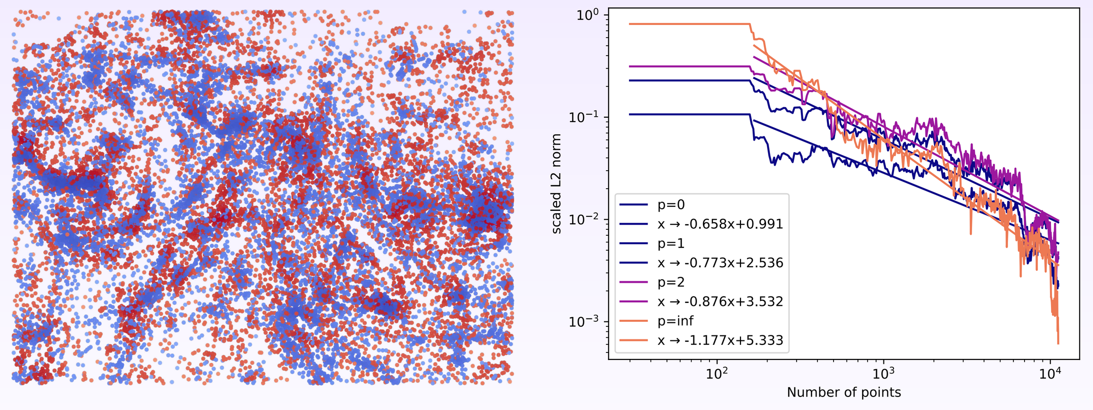
Application: Convergence Rates

This stablity along with known kernel density estimation, and persistence convergence rates,
allows us to have convergence rates to the ground truth, with respect to the number of sampling points.

In this example, the bifiltration is given by
- a density estimation of the red points, and
- a density estimation of the blue points.

The (pointwise) theorical convergence rate is

$$\left\|V_{p,\delta}(M) - V_{p,\delta}(\hat{M}_n)\right\|_\infty \lesssim \frac{1}{\delta}\sqrt{\frac{\ln n}{n}}$$



Inria
UNIVERSITÉ CÔTE D'AZUR — INTERDISCIPLINARY INSTITUTE FOR ARTIFICIAL INTELLIGENCE 3IA CÔTE D'AZUR — FRANCE 2030
COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

Poster link

Code link