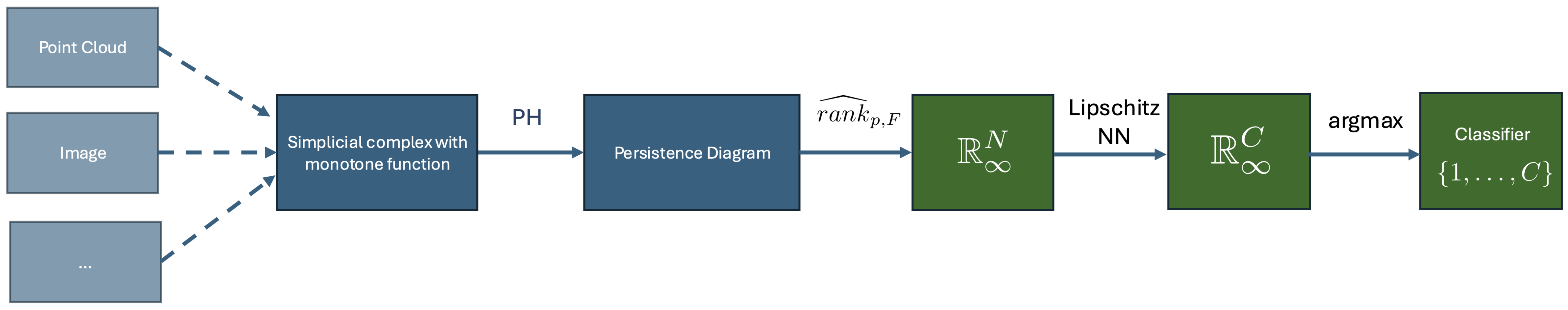
Certifying Robustness via Topological Representations

Jens Agerberg, Andrea Guidolin, Andrea Martinelli, Pepijn Roos Hoefgeest, David Eklund, Martina Scolamiero
KTH Royal Institute of Technology

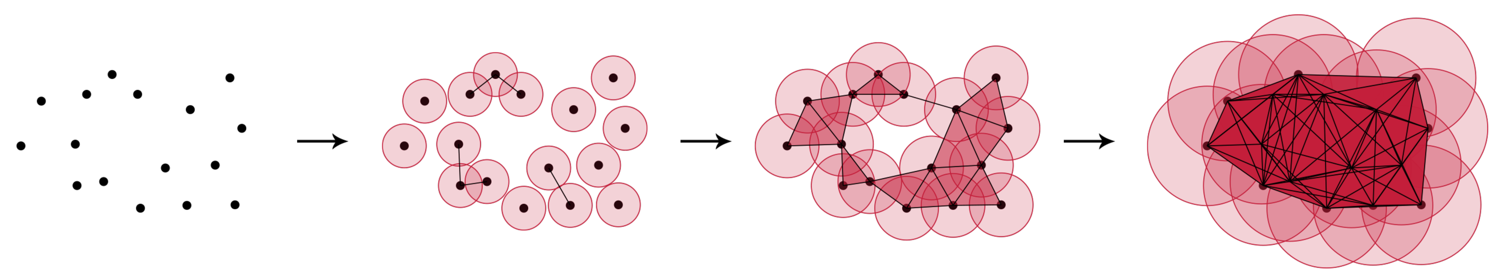
Architecture

We propose an end-to-end robust Persistent Homology/ML pipeline, to learn representations with stability guarantees w.r.t. metrics in the data space. We analyze the robustness in the framework of Adversarial Machine Learning.



Persistent Homology

From a point cloud we can construct a Vietoris-Rips complex, a combinatorial object encoding its geometry, parametrized by $t \in [0,\infty)$.



By taking homology we get (for each homological degree) a vector space for each t and a linear map for each $\tau \leq t \in [0,\infty)$.

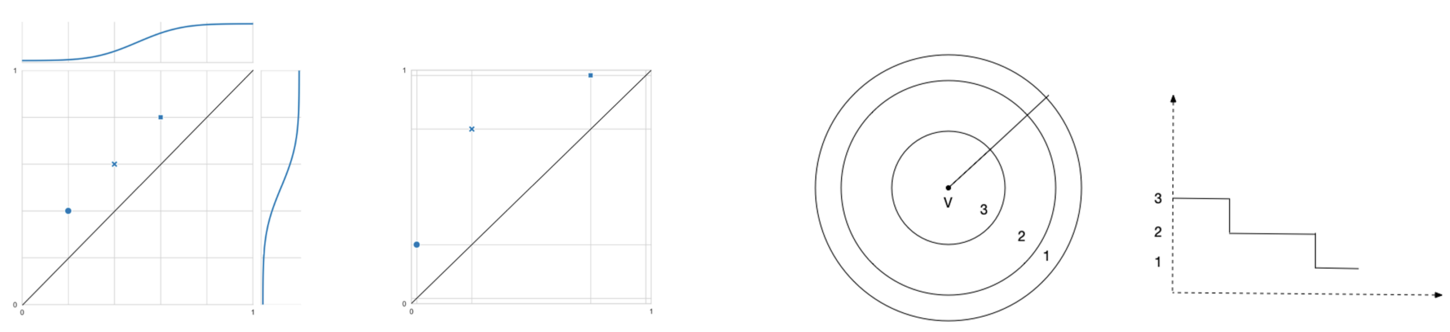
This object can be decomposed into a **persistence diagram** $D = \{(a_i,b_i)\}_{i=1}^N$ and endowed with a Wasserstein-type distance W_p $(p \in [1,\infty])$.

Lipschitz continuity of W_p w.r.t. various metrics on input spaces has been shown. For instance for d_{GH} Gromov-Hausdorff distance between finite metric spaces X, Y:

$$W_\infty(D(X),D(Y)) \leq d_{\mathsf{GH}}(X,Y).$$

Stable Rank vectorization

Given a persistence diagram $D = \{(a_i,b_i)\}_{i=1}^N$ and an increasing bijection F we have: $F(D) = \{(F(a_i),F(b_i))\}_{i=1}^N$



The **stable rank** of D corresponding to F and $p \in [1,\infty]$ is the function:

$$\widehat{\mathrm{rank}}_{p,F}(D)(t) := \min\{\mathrm{rank}(D') \mid D' \in \mathcal{PD} \text{ and } W_p(F(D),F(D')) \leq t\}$$

A distance $d_{\bowtie}$ can be defined between stable ranks, equivalent to an L_∞ distance.

Proposition $d_{\bowtie}(\widehat{\mathrm{rank}}_{p,F}(D), \widehat{\mathrm{rank}}_{p,F}(D')) \leq K W_p(D,D')$

where K is the Lipschitz constant of F.

Lipschitz Neural Network

L_∞ Neural Networks (*Zhang et al.*) propose to replace the MLP layers with layers composed of neurons of the form:

$$u(\mathbf{x},w,b) = ||\mathbf{x} - w||_\infty + b.$$

Neural networks formed by such layers are by design 1-Lipschitz stable w.r.t. input in an L_∞ space.

Robustness in Adversarial Machine Learning

We have a PH pipeline $\phi : \mathcal{X} \to \{1,\dots,C\}$, which classifies samples in the data space $\mathcal{X}$ to one of C classes.

A sample $x \in \mathcal{X}$ with ground truth label c is ϵ-robust if:

$$g(x') = c, \forall x' \in \mathcal{X} \text{ s.t. } d_{\mathcal{X}}(x,x') \leq \epsilon.$$

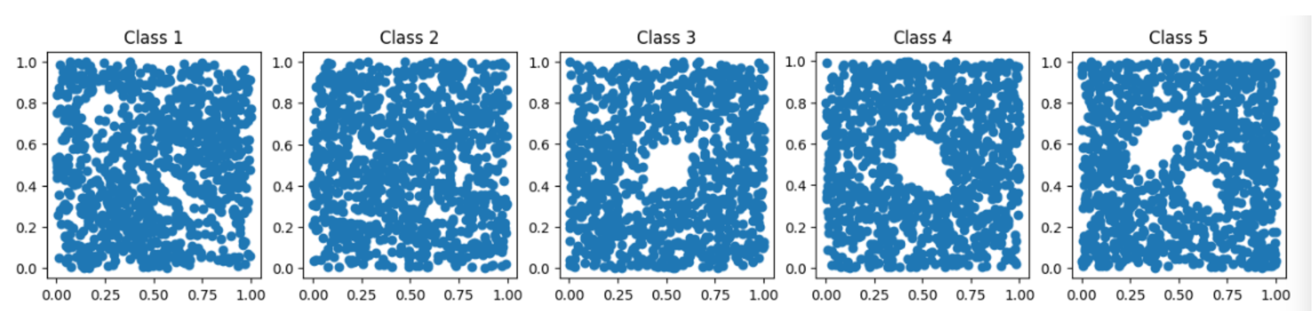
If the whole PH pipeline ϕ has known Lipschitz constant K, we can derive its robustness in Adversarial ML sense.

For a sample $x \in \mathcal{X}$, we define its margin $M_x = \phi(x)_{c_x} - \max_{i \neq c_x} \phi(x)_i$, where c_x is the ground truth label of x.

Then x is ϵ-robust for $\epsilon = \frac{M_x}{2K}$.

Results

We consider the dataset with realizations of point processes introduced in *Perslay (Carrière et al.)*:



We can compare a lower-bound of ϵ-robustness for our pipeline (SRN) to an upper-bound for Perslay (derived from methods to find adversarial examples), at the level of persistence diagrams.

Acc. at $\epsilon =$	0	10^{-5}	10^{-2}	10^{-1}	1
Perslay (H_1 only)	84.4	27.4	27.4	24.8	24.8
SRN (H_1 only)	79.6	79.6	78.8	74.6	51.3

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM