Large-Scale Topological Data Analysis

Björn H. Wehlin, KTH Royal Institute of Technology
Deptartment of Mathematics
On behalf of TDA group at KTH
(https://www.kth.se/math/act/tda)

Motivation & Research Goals

Persistent homology (PH) gives computers a sense of geometry. It is one of the cornerstones of Topological Data Analysis (TDA). If we want to compare different geometric datasets, we can use PH-derived functional invariants such as stable rank [2, 3, 4]. Traditionally, a lot of effort has been spent on making computations fast for a single invariant. Our aim is to leverage modern heterogeneous compute environments to accelerate TDA computations for massive datasets and experiments. In addition to providing the TDA community with new computational tools, we hope to gain insight for developing new mathematics.

Denoising through homology

Let f, g be $\mathcal{Y}$-valued *measurements* of $\mathcal{X}$:

$$\mathcal{X} \underset{g}{\overset{f}{\rightrightarrows}} \mathcal{Y}, \quad \text{such that} \quad d(f(X), g(X)) \leq \varepsilon \quad \text{for all } X \in \mathcal{X}, \text{ some } \varepsilon > 0.$$

Homology-based pipeline:

$$\text{Distance Spaces} \longrightarrow \begin{matrix}\text{Time Series of}\\ \text{Vector Spaces}\end{matrix} \longrightarrow \begin{matrix}\text{Lebesgue-}\\ \text{Measurable Functions}\\ [0,\infty) \to [0,\infty)\end{matrix}$$

$$Y \longrightarrow H_\bullet(\mathrm{VR}_t(Y); \mathbb{F}) \longrightarrow \widehat{\mathrm{rank}}[Y](t)$$

Theorem: (Chachólski, Riihimäki, 2020)

Let $\mathcal{Y}$ be a pseudometric space with distance d. Let $Y_1, Y_2 \in \mathcal{Y}$, and $p \geq 1$. Then,

$$L_p(\widehat{\mathrm{rank}}[Y_1], \widehat{\mathrm{rank}}[Y_1]) \leq c\, d(Y_1, Y_2)^{1/p}$$

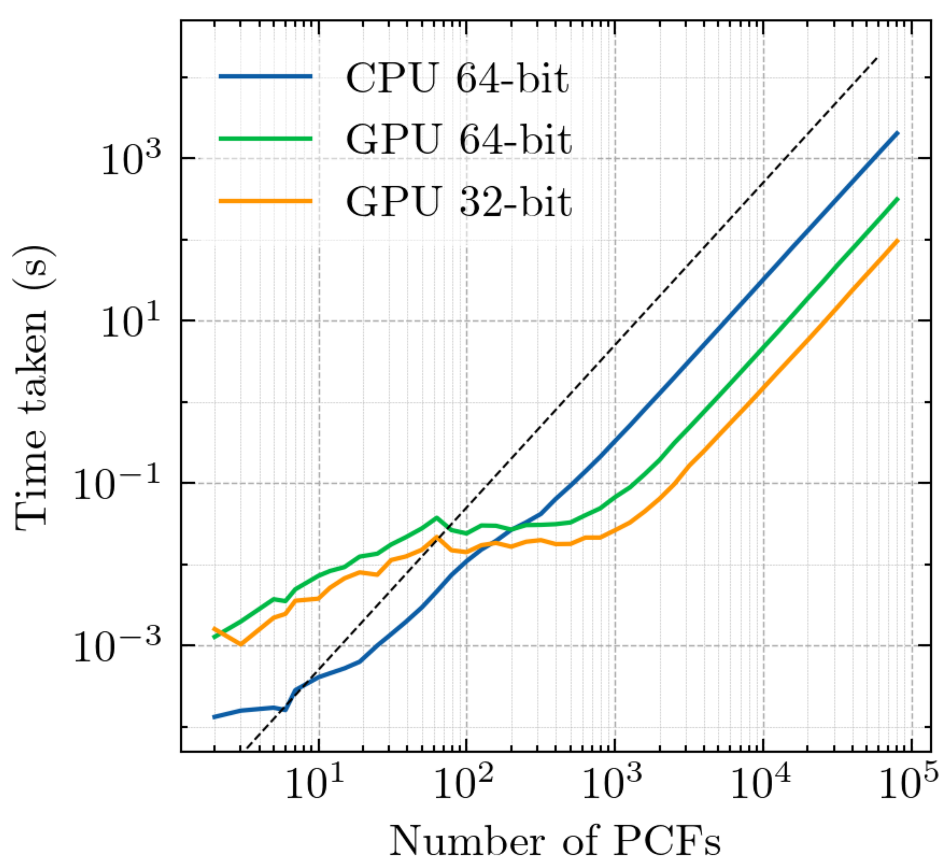
where $c = \max\{\widehat{\mathrm{rank}}[Y_1](0), \widehat{\mathrm{rank}}[Y_2](0)\}$.

Similarity computation

GPU-enabled computations on piecewise constant functions [1].

$$d(f,g) = \int_0^\infty |f(t) - g(t)|\, dt$$

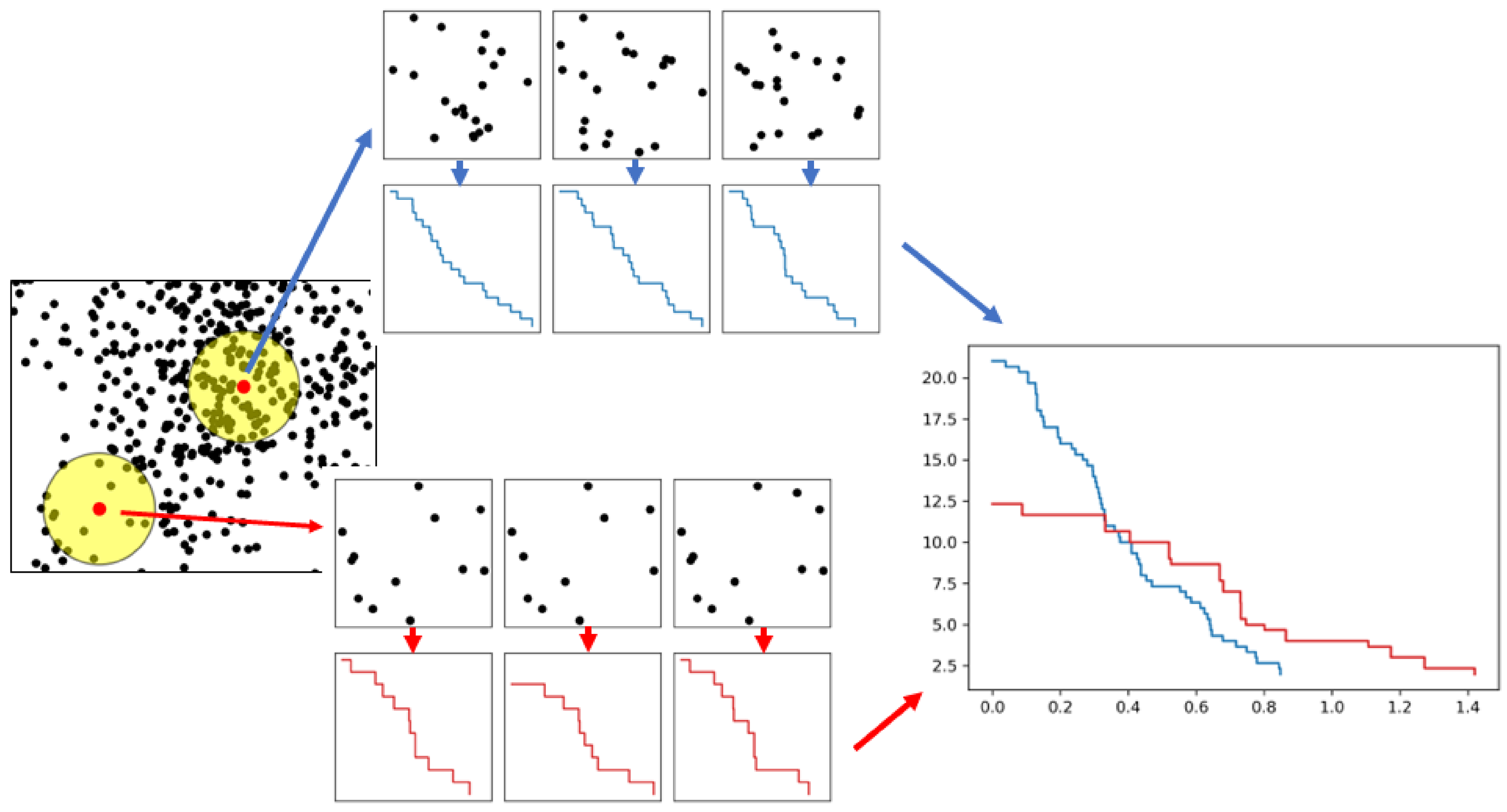
$$D = \begin{bmatrix} 0 & d(f_1,f_2) & d(f_1,f_3) & \cdots & d(f_1,f_{n-1}) \\ & 0 & d(f_2,f_3) & \cdots & d(f_1,f_{n-1}) \\ & & 0 & \ddots & \vdots \\ & & & 0 & d(f_{n-1},f_n) \\ & & & & 0 \end{bmatrix}$$



`pip install masspcf`

Geometric exploration

Explore geometry by subsampling from distributions within dataset or reference object [5]:



- How to sample efficiently when dataset is large?
  - Incremental lookup in kD tree (other metric trees?)
  - Fast pseudorandom number generation [6]
  - **Sub-ms sampling latency** for 30x25 points from 338k $\mathbb{R}^4$ point dataset
- To be developed: distributed $\widehat{\mathrm{rank}}$ computations from data
- Idea: run lots of experiments (datasets, hyperparameters, ...)
- Goal: new math (statistical tests, tools, ...)

Acknowledgments

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

[1] Wehlin, B. H. (2024). Massively Parallel Computation of Similarity Matrices from Piecewise Constant Invariants. arXiv:2404.07183.

[2] Chachólski, W., & Riihimäki, H. (2020). Metrics and stabilization in one parameter persistence. SIAM J. Appl. Algebra Geom., 4(1), 69-98.

[3] Gäfvert, O., & Chachólski, W. (2017). Stable invariants for multiparameter persistence. arXiv:1703.03632.

[4] Scolamiero, M., Chachólski, W., Lundman, A., Ramanujam, R., & Öberg, S. (2017). Multidimensional persistence and noise. Found. Comput. Math., 17, 1367-1406.

[5] Agerberg, J., Chacholski, W., & Ramanujam, R. (2023). Global and Relative Topological Features from Homological Invariants of Subsampled Datasets. TAG-ML, PMLR 221:302-312.

[6] Blackman, D. & Vigna, S. (2021). Scrambled Linear Pseudorandom Number Generators. ACM Trans. Math. Softw. 47, 4, Article 36

WALLENBERG AI,
AUTONOMOUS SYSTEMS
AND SOFTWARE PROGRAM