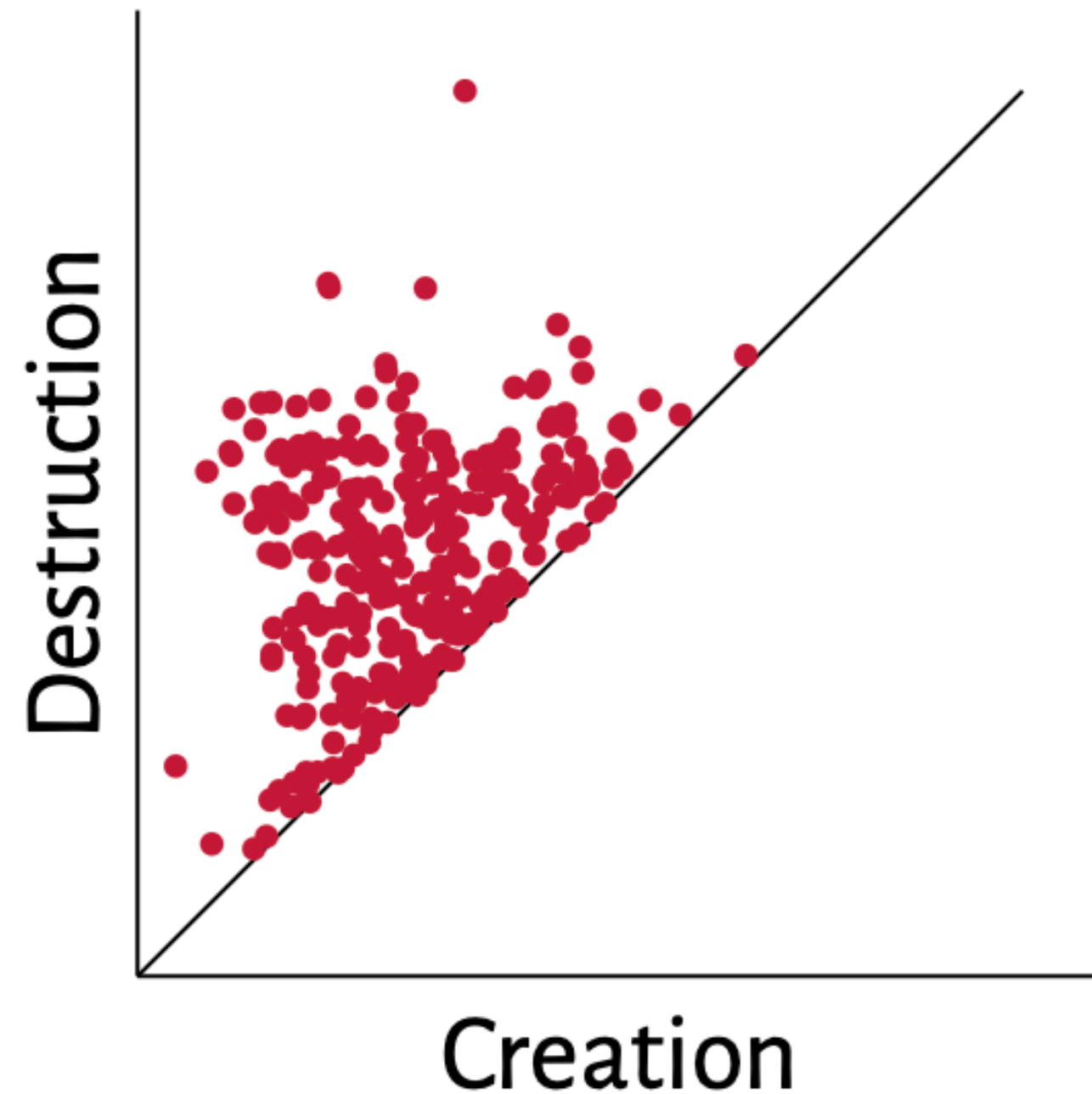


Топологические дескрипторы и их применения в анализе данных

по материалам лекций Bastian Rieck, “Topological Data Analysis for Machine Learning”
– <https://bastian.rieck.me/>

Максим Бекетов, аспирант МЛ АТиП ФКН ВШЭ,
школа “Торическая Топология, Комбинаторика и Анализ Данных” – СПб, октябрь 2022

Persistence Diagrams (PDs)



- точки в $\mathbb{R} \times \mathbb{R} \cup \{\infty\}$
- устойчивость (persistence) = расстояние до диагонали
- точки могут перекрываться
- место под диагональю не используется :)
- главный **вопрос**: как использовать в анализе?
как “подавать на вход” алгоритму классификации, например?

Persistence Diagrams (PDs)

Bottleneck-расстояние между двумя диаграммами \mathcal{D} и \mathcal{D}' :

$$W_{\infty}(\mathcal{D}, \mathcal{D}') := \inf_{\eta: \mathcal{D} \rightarrow \mathcal{D}'} \sup_{x \in \mathcal{D}} \|x - \eta(x)\|_{\infty}$$

где $\eta: \mathcal{D} \rightarrow \mathcal{D}'$ – биекция между точками \mathcal{D} и \mathcal{D}' (если неравномощны – расстояние до диагонали)

“смягчение” – **расстояние Вассерштейна**:

$$W_p(\mathcal{D}_1, \mathcal{D}_2) := \left(\inf_{\eta: \mathcal{D}_1 \rightarrow \mathcal{D}_2} \sum_{x \in \mathcal{D}_1} \|x - \eta(x)\|_{\infty}^p \right)^{\frac{1}{p}}$$

Persistence Diagrams (PDs)

Диаграммы устойчивости **устойчивы** к шуму:

- как в bottleneck-метрике¹: $W_{\infty}(\mathcal{D}_f, \mathcal{D}_g) \leq \|f - g\|_{\infty}$
- так и в Вассерштейне²

¹ D. Cohen-Steiner, H. Edelsbrunner and J. Harer, 'Stability of persistence diagrams', *Discrete & Computational Geometry* 37.1, 2007, pp. 103–120

² Skraba, Primoz, and Katharine Turner. "Wasserstein stability for persistence diagrams." *arXiv preprint arXiv:2006.16824* (2020).

Persistence Diagrams (PDs)

- Однако, для вычисления расстояния – нужно решать задачу **оптимального транспорта**³
- Впрочем, есть эффективные алгоритмы⁴, втч приближения/эвристики⁵
- И для многих задач “сравнения PDs” этого может быть достаточно

3) G. Peyré, M. Cuturi et al., ‘Computational Optimal Transport’, *Foundations and Trends in Machine Learning* 11.5–6, 2019, pp. 355–607

4) M. Cuturi, ‘Sinkhorn Distances: Lightspeed Computation of Optimal Transport’, *Advances in Neural Information Processing Systems*, ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, vol. 26, Curran Associates, Inc., 2013

5) M. Kerber, D. Morozov and A. Nigmatov, ‘Geometry helps to compare persistence diagrams’, *Proceedings of the 18th Workshop on Algorithm Engineering and Experiments (ALENEX)*, ed. by M. Goodrich and M. Mitzenmacher, Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2016, pp. 103–112

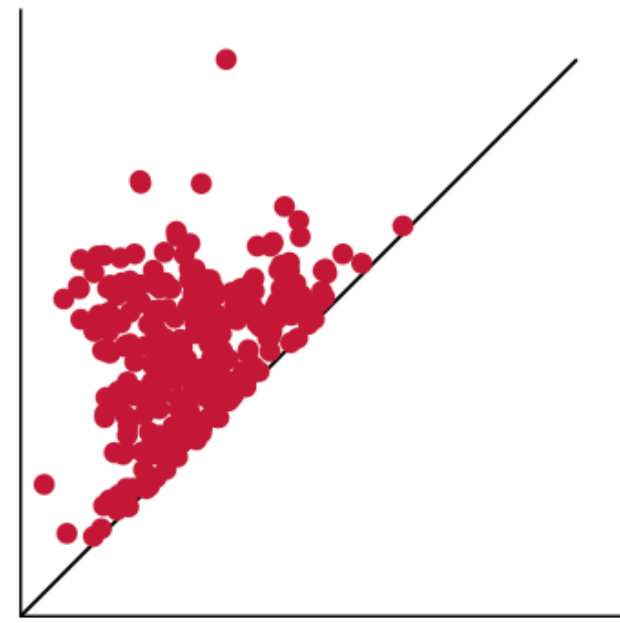
Интерлюдия: ядерные методы

- На мн-ве \mathcal{X} функция $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ называется **ядром**, если суц.
 - 1) гильбертово пр.-во \mathcal{H} (полное метрич. пр.-во со скал. произв.) и
 - 2) отображение $\Phi : \mathcal{X} \rightarrow \mathcal{H}$

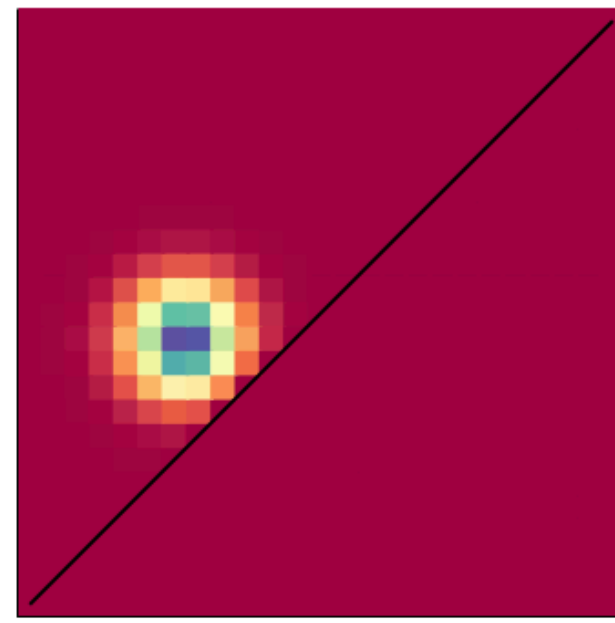
такие, что: $k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} \quad \forall x, y \in \mathcal{X}$

- В анализе данных / машинном обучении это нужно затем, чтобы вычислять “близость/похожесть” объектов, **не вычисляя явно расстояния** между ними!

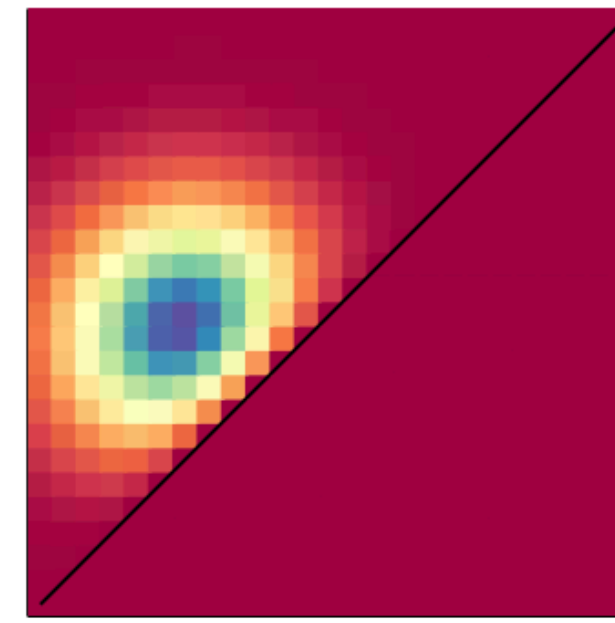
Ядра для PDs:



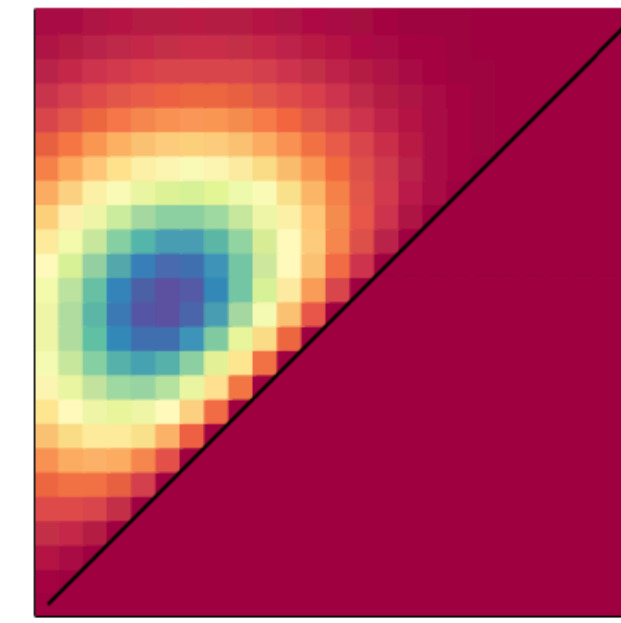
Persistence diagram



$\sigma = 0.1$



$\sigma = 0.5$



$\sigma = 1.0$

- Первое предложенное⁶ ядро для диаграмм устойчивости:

$$k(\mathcal{D}, \mathcal{D}') := \frac{1}{8\pi\sigma} \sum_{p \in \mathcal{D}, q \in \mathcal{D}'} \exp(-8^{-1}\sigma^{-1}\|p - q\|^2) - \exp(-8^{-1}\sigma^{-1}\|p - \bar{q}\|^2)$$

$$\Phi(x) := \frac{1}{4\pi\sigma} \sum_{p \in \mathcal{D}} \exp(-4^{-1}\sigma^{-1}\|x - p\|^2) - \exp(-4^{-1}\sigma^{-1}\|x - \bar{p}\|^2)$$

- просто сумма, не нужно решать задачу ОТ (!)

6) J. Reininghaus, S. Huber, U. Bauer and R. Kwitt, 'A stable multi-scale kernel for topological machine learning', *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Red Hook, NY, USA: Curran Associates, Inc., 2015, pp. 4741–4748

Ядра для PDs:

- Есть и другие предложения, основанные на: sliced Wasserstein distance⁷, ядерных вложениях⁸, идеях из Римановой геометрии⁹

Приложения:

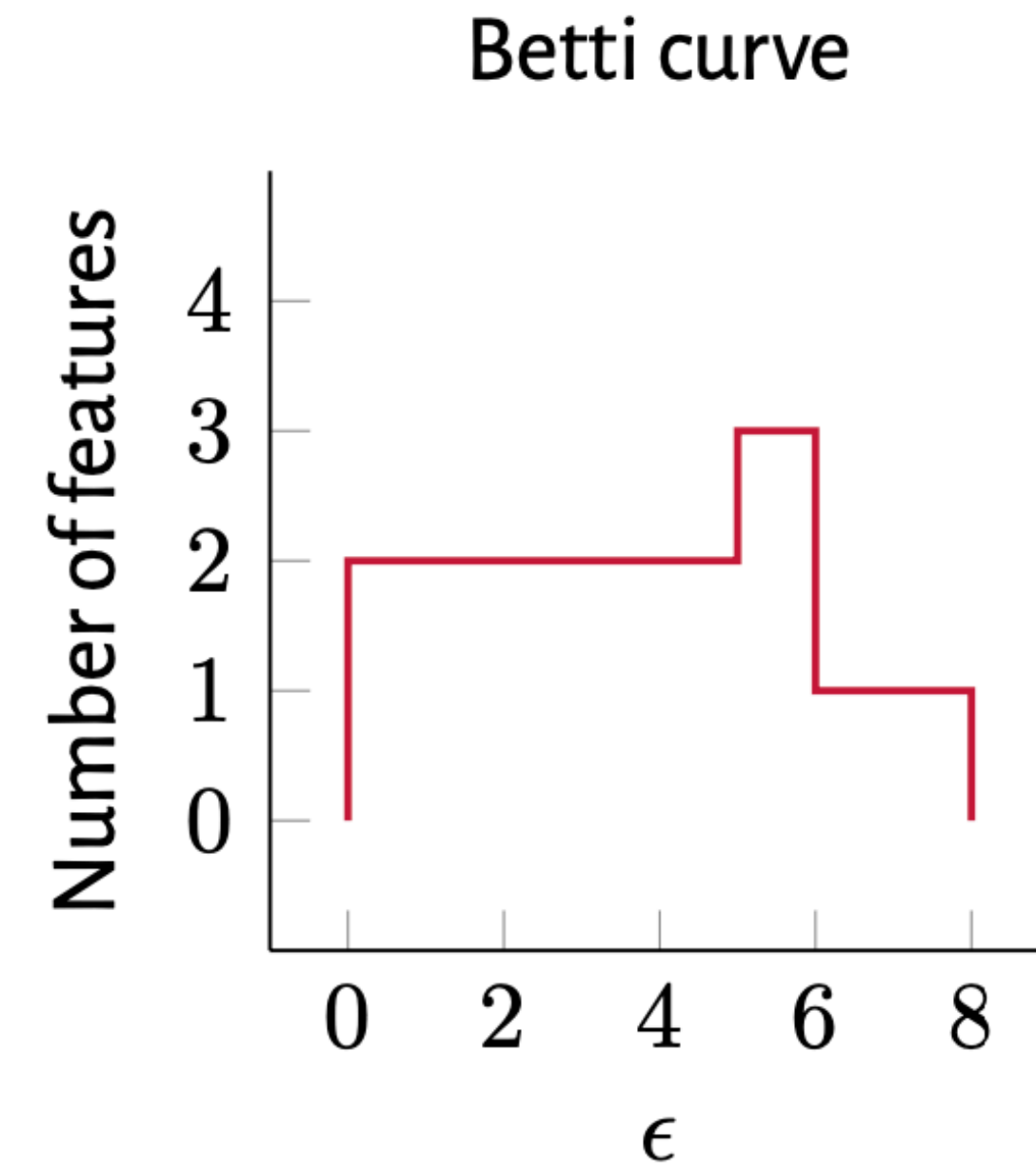
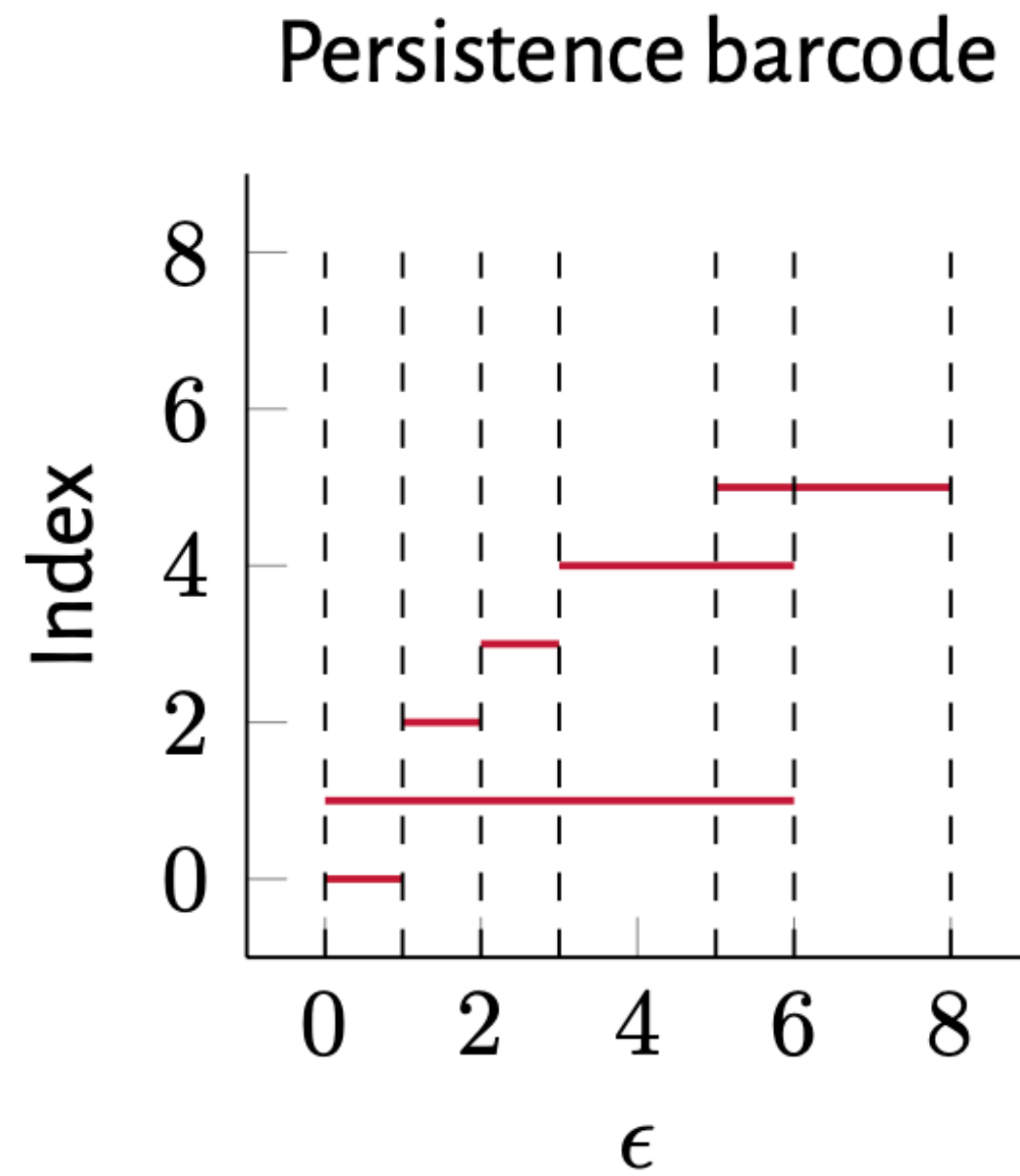
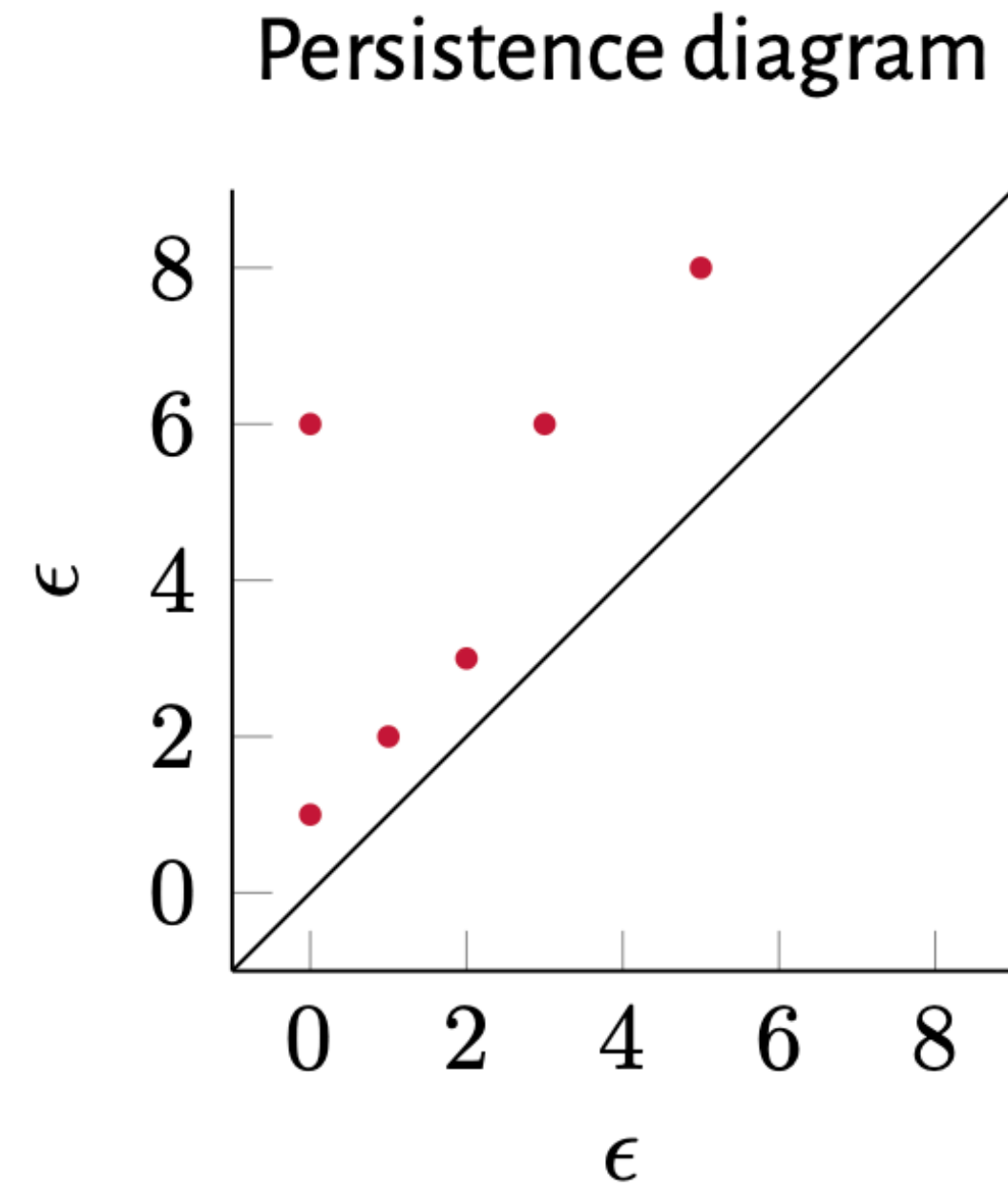
- kernel PCA – для визуализации / снижения размерности / генерации признаков
- kernel SVM/SVR – для классификации / регрессии

7) M. Carrière, M. Cuturi and S. Oudot, ‘Sliced Wasserstein Kernel for Persistence Diagrams’, ed. by D. Precup and Y. W. Teh, vol. 70, Proceedings of Machine Learning Research, PMLR, 2017, pp. 664–673

8) G. Kusano, K. Fukumizu and Y. Hiraoka, ‘Kernel Method for Persistence Diagrams via Kernel Embedding and Weight Factor’, Journal of Machine Learning Research 18.189, 2018, pp. 1–41

9) T. Le and M. Yamada, ‘Persistence Fisher Kernel: A Riemannian Manifold Kernel for Persistence Diagrams’, Advances in Neural Information Processing Systems, ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, vol. 31, Curran Associates, Inc., 2018, pp. 10007–10018

Betti curves (sequences):



- просто кусочно-постоянная функция $\mathcal{B} : \mathbb{R} \rightarrow \mathbb{N}$

10) Yuhei Umeda. Time series classification via topological data analysis. *Information and Media Technologies*, 12:228–239, 2017.

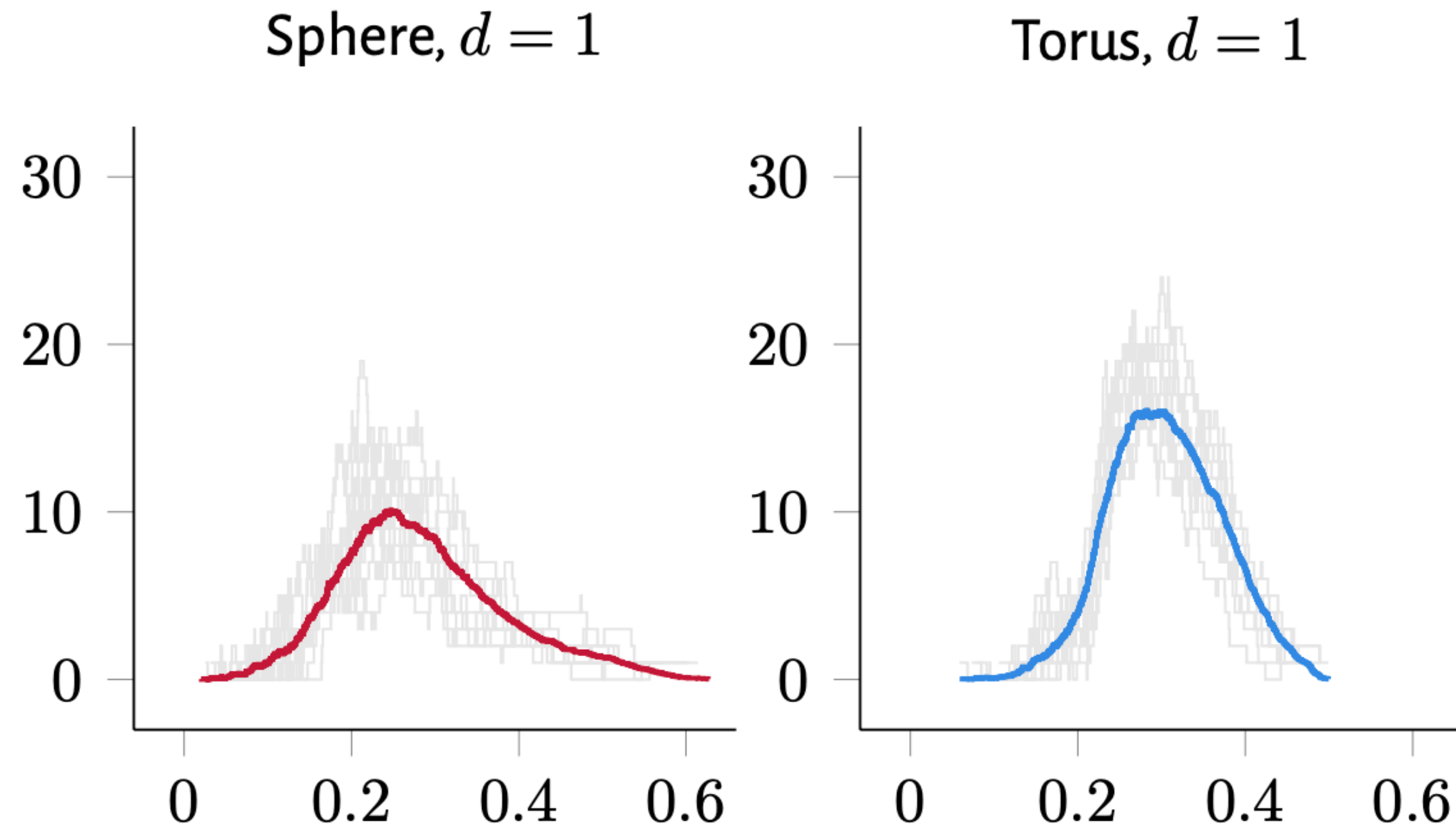
Betti curves (sequences):

- Легко построить, кусочно-постоянные функции
- Хотя есть вопросы по устойчивости
- Можно говорить о **линейных операциях**: сложение, усреднение
- Можно ввести¹¹ ядро:

$$k_p(\mathcal{D}, \mathcal{D}') := - \left(\int_{\mathbb{R}} |\mathcal{B}_{\mathcal{D}}(x) - \mathcal{B}_{\mathcal{D}'}(x)|^p dx \right)^{\frac{1}{p}}$$

11) B. Rieck, F. Sadlo and H. Leitte, ‘Topological Machine Learning with Persistence Indicator Functions’, *Topological Methods in Data Analysis and Visualization V*, ed. by H. Carr, I. Fujishiro, F. Sadlo and S. Takahashi, Cham, Switzerland: Springer, 2020, pp. 87–101, arXiv: 1907.13496 [math.AT]

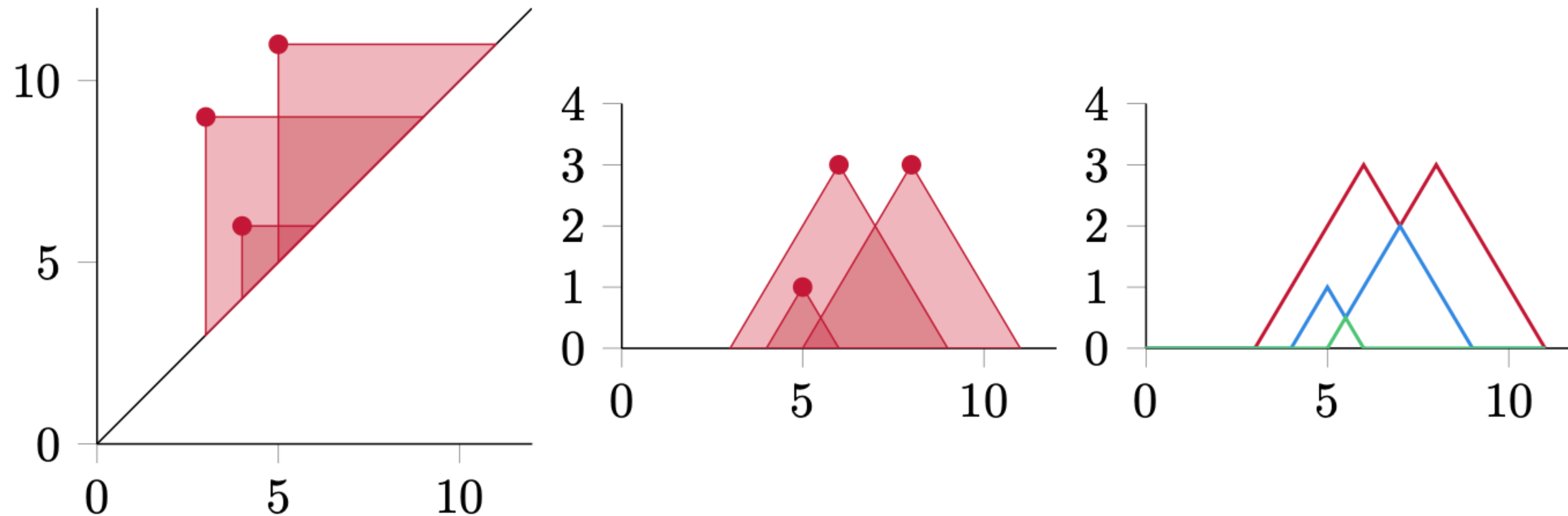
Betti curves (sequences):



- МОЖНО ЗАНИМАТЬСЯ СТАТИСТИКОЙ, ВТЧ. **ТЕСТИРОВАНИЕМ ГИПОТЕЗ**

11) B. Rieck, F. Sadlo and H. Leitte, 'Topological Machine Learning with Persistence Indicator Functions', *Topological Methods in Data Analysis and Visualization V*, ed. by H. Carr, I. Fujishiro, F. Sadlo and S. Takahashi, Cham, Switzerland: Springer, 2020, pp. 87–101, arXiv: 1907.13496 [math.AT]

Persistence landscapes:



- “иерархическое” представление¹²

12) P. Bubenik, ‘Statistical Topological Data Analysis Using Persistence Landscapes’, Journal of Machine Learning Research 16, 2015, pp. 77–102

Persistence landscapes:

- Взаимно однозначное кодирование (без потери информации)
- Устойчивость как для PDs, “иерархия”
- Можно “хранить” не сами функции, а “**семплировать**” их значения с каким-то шагом – и получать “вектор”-признак фикс. длины
- Потому можно подавать на вход нейросети (использовать “как слой”)¹³

13) K. Kim, J. Kim, M. Zaheer, J. Kim, F. Chazal and L. Wasserman, ‘PLLay: Efficient Topological Layer based on Persistent Landscapes’, *Advances in Neural Information Processing Systems*, ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin, vol. 33, Curran Associates, Inc., 2020, pp. 15965–15977

Persistence Images:



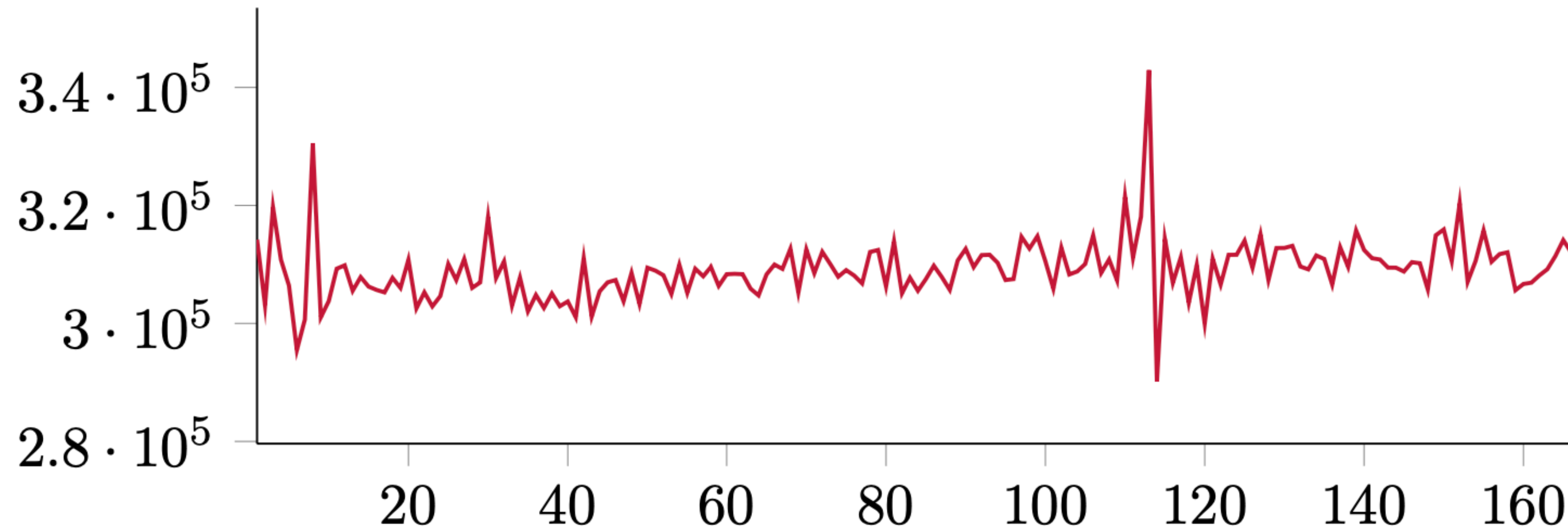
- Функция¹⁴ $\Psi : \mathbb{R}^2 \rightarrow \mathbb{R}$ диаграммы $\Psi(z) = \sum_{x,y \in \mathcal{D}} w(x,y) \Phi(x,y,z)$, где $w(\cdot)$ – PL-функция “весов”, а $\Phi(\cdot)$ – некая плотность вер-ти (часто изотроп. нормальное)
- Дискретизируется в картинку $r \times r$ (r for resolution)

14) H. Adams et al., ‘Persistence Images: A Stable Vector Representation of Persistent Homology’, Journal of Machine Learning Research 18.8, 2017, pp. 1–35

Persistence Images:

- устойчивость
- своего рода ядерная оценка плотности (KDE)
- r надо эмпирически выбирать
- не-разреженное (non-sparse) представление
- но можно использовать методы, хорошо работающие на картинках (и вполне успешно используют на тестовых датасетах)

Нормы PDs:



- для time-varying данных можно смотреть на **норму PD** от времени:

$$\|\mathcal{D}\|_{\infty} := \max_{x,y \in \mathcal{D}} \text{pers}^p(x,y) \quad \text{или} \quad \|\mathcal{D}\|_p := \left(\sum_{x,y \in \mathcal{D}} \text{pers}^p(x,y) \right)^{1/p}$$

Резюме: Какой метод использовать?

