A short note on distance measures and outlier robustness

January 6, 2023

Abstract

This note is an attempt to summarise some of the articles and constructions we have been looking at in previous meetings. I will also try to summarise what kind of outlier robustness (leave-one-out stability) these methods enjoy.

Some fixed notation: Given a (finite) point cloud X with $|X| = n$, let μ_X denote the empirical probability measure with support on X. That is, $\mu_X = \frac{1}{n}\sum_{x\in X}\delta_x$ where δ_x denotes the Dirac measure centred at x. Similarly, let μ_{X_0} denote the empirical measure on $X_0 = X \setminus \{x_0\}$ for some arbitrary $x_0 \in X$.

1 Uniform stability

Let Y be a metric space, and consider a function V taking a finite subset $X \subseteq Y$ to a persistence module V_X. We introduce the notion of uniform stability for such functions with inspiration from learning theory (cf. [1]). If $X \subseteq Y$ is a point cloud and $x_0 \in X$, we fix the notation $X_0 = X \setminus \{x_0\}$. The intuition behind uniform stability is that, given enough samples, the difference between V_X and V_{X_0} should be very small.

Definition 1.1. We say that V is *uniformly stable with respect to d* if for every finite subset $X \subseteq Y$ with $|X| = n$, there exists a positive real number $\beta(n) \in \mathcal{O}(n^{-1})$ such that

$$d\left(V_X, V_{X_0}\right) \leqslant \beta(n) \quad \text{holds for all } x_0 \in X. \tag{1}$$

For 1-parameter persistent homology, we can take d to be the usual interleaving distance d_I between persistence modules. By the isometry theorem of [4], this distance agrees with the bottleneck distance between barcodes/persistence diagrams. The interleaving distance extends to 2-parameter persistence modules which allows us to talk about uniform stability in the 2-parameter setting as well. In practice, one does not necessarily prove stability directly on the algebraic level of persistence modules, but rather on the level of (bi-)filtrations of topological spaces or simplicial complexes.

1.1 The multicover bifiltration and the Prohorov distance

The *offset filtration* $O(X)$ of a point cloud X is a filtered topological space given in filtration degree $r > 0$ as the union of all balls of radius r centred at the points in X. Taking the nerve of this collection of balls of radius r, we get the usual Čech filtration. The multicover bifiltration $\mathcal{M}(X)$ is a 2-parameter extension of the offset filtration which is sensitive to density. Given a metric measure space (Y, η_Y), the *measure bifiltration* $\mathcal{B}(Y)$ is given in filtration degree (k, r) by

$$\mathcal{B}(Y)_{(k,r)} = \{y \in Y \mid \eta_Y(B(y, r)) \geq k\}.$$

For a point cloud $X \subseteq \mathbb{R}^d$, the *unormalized multicover bifiltration* $\mathcal{M}^u(X)$ of X is then defined as $\mathcal{M}^u(X) = \mathcal{B}(\hat{\nu}_X)$ where $\hat{\nu}_X$ is the counting measure on X defined as $\hat{\nu}_X(A) = |A \cap X|$. In other words, $\mathcal{M}^u(X)_{(k,r)}$ consists of all $y \in \mathbb{R}^d$ such that the ball centred at y of radius r contains at least k distinct points x from X. The *(normalized) multicover filtration* $\mathcal{M}(X)$ is defined similarly by replacing $\hat{\nu}_X$ with the empirical probability measure $\nu_X(A) = |A \cap X|/|X|$. This normalizes the first parameter k by the number of points in X.

The *Prohorov distance* d_{Pr} between two (probability) measures μ and η on a metric space (Y, d_Y) is given by

$$d_{Pr}(\mu, \eta) = \sup_A \inf\{\delta \geq 0 \mid \mu(A) \leq \eta(A^\delta) + \delta \text{ and } \eta(A) \leq \mu(A) + \delta\}$$

where A ranges over all closed sets in Y. The notation A^δ means the δ-thickening of A in Y with respect to the metric d_Y. It follows from remark 2.16 in [6] that the Prohorov distance between the two empirical probability measures μ_X and μ_{X_0} is bounded above by $\frac{1}{n-1}$. The following special case of a theorem of [6] then implies uniform stability for the multicover bifiltration with respect to the interleaving distance for 2-parameter persistence modules.

Theorem 1.2 (Theorem 1.6 [6])**.** For X and X' non-empty, finite subsets of $\mathbb{R}^d$,

$$d_I(\mathcal{M}(X), \mathcal{M}(X')) \leq d_{Pr}(\mu_X, \mu_{X'}).$$

Look at the subdivision and degree bifiltrations and understand how these are related to the multicover bifiltration.

1.2 The Wasserstein distance

The Wasserstein distance, also known as the Earth Mover's Distance, is a distance measure between probability measures. It can be used to compare the similarity between two distributions, with smaller values indicating greater similarity. Some methods proposed in TDA for dealing with noise and outliers gives guarantees in terms of the Wasserstein distance. For two probability measures μ and η on $Y = \mathbb{R}^d$, we define the *p-Wasserstein distance* between μ and ν to be

$$W_p(\mu, \eta) = \inf_{\nu} \left(\int_{Y \times Y} \|x - y\|^p \, d\nu \right)^{\frac{1}{p}}.$$

Here, ν ranges over all couplings of μ and η. A coupling is a joint probability distribution on $Y \times Y$ with marginals equal to μ and η, respectively. Such couplings are also often referred to as transport plans, and one minimizing the above expression is called an optimal transport plan. For two measures μ and η with finite support, a transport plan can be represented as a matrix $A = (a_{ij})$ where a_{ij} is the mass moved from x_j to x_i. In this case, the p-Wasserstein distance between μ and η is given as

$$W_p(\mu, \eta) = \inf_{A} \left(\sum_{i,j} \|x_j - y_i\|^p \cdot a_{ij} \right)^{\frac{1}{p}}.$$

For example, let μ_X and μ_{X_0} be the empirical measures with support X and X_0, respectively. One way to think about transport plans between μ_X and μ_{X_0} is as follows: If we add a point x_0 to go from X_0 to X, each of the other $n-1$ points have to give away some mass so that all points, including x_0, ends up with a mass of $\frac{1}{n}$. One possible transport plan is given by the $(n-1)\times n$-matrix

$$\begin{bmatrix} \frac{1}{n(n-1)} & \frac{1}{n} & 0 & 0 & \dots & 0 & 0 \\ \frac{1}{n(n-1)} & 0 & \frac{1}{n} & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{1}{n(n-1)} & 0 & 0 & 0 & \dots & \frac{1}{n} & 0 \\ \frac{1}{n(n-1)} & 0 & 0 & 0 & \dots & 0 & \frac{1}{n} \end{bmatrix}$$

Since the Wasserstein distance is the infimum over all such transport plans, we have the following upper bound

$$\begin{aligned} W_p(\mu_X, \mu_{X_0}) &\leq \left(\frac{1}{n(n-1)} \sum_{x \in X_0} \|x_0 - x\|^p \right)^{\frac{1}{p}} \\ &\leq \left(\frac{n-1}{n(n-1)} \operatorname{diam}(X)^p \right)^{\frac{1}{p}} = n^{-\frac{1}{p}} \operatorname{diam}(X). \end{aligned}$$

Idea:

Can we find a tighter bound for $d_W(\mu_X, \mu_{X_0})$?

Assuming that the diameter of X does not depend on $|X|$, the above inequality implies that the 1-Wasserstein distance between μ_X and μ_{X_0} is $O(n^{-1})$.

1.2.1 The distance to a measure ([3])

For a probability measure μ on a metric space Y and a mass parameter $m \in (0,1]$, define the function $\delta_{m,\mu} \colon \mathbb{R}^d \to \mathbb{R}_{>0}$ by sending a point y in $\mathbb{R}^d$ to the real number $\inf\{r > 0 \mid \mu(\bar{B}(y,r)) > m\}$. Intuitively, $\delta_{m,\mu}$ describes how large a disk around y we have to consider to capture at least m mass as defined by μ. Using δ, we define *the distance $d_{\mu,m}$ to measure μ* as

$$\begin{aligned} d_{\mu,m} \colon \mathbb{R}^d &\to \mathbb{R}_{>0} \\ y &\mapsto \sqrt{\frac{1}{m} \int_0^m \delta_{mu,l}(x)^2 dl}. \end{aligned}$$

Let $\mathrm{Dgm}(d_{\mu,m})$ denote the persistence diagram corresponding to the sublevel set homology of $d_{\mu,m}$. In [3], the following Wasserstein stability result is established for the distance to a measure function:

Theorem 1.3 (Theorem 3.1 [3]). For two probability measures μ and ν and a triangulable metric space Y and a mass parameter m, we have that

$$d_B(\mathrm{Dgm}(d_{\mu,m}), \mathrm{Dgm}(d_{\nu,m})) \leq \frac{1}{\sqrt{m}} W_2(\mu, \nu).$$

In particular, for the empirical measures μ_X and μ_{X_0} this implies

$$d_B(\mathrm{Dgm}(d_{\mu_X,m}), \mathrm{Dgm}(d_{\mu_{X_0},m})) \leq \frac{1}{\sqrt{n}} \frac{1}{\sqrt{m}} \mathrm{diam}(X).$$

Approximation via a weighted Čech (or Rips) complex

Write something about stability in case of the complex used to approximate DTM. See [5]. There are also some stability results wrt. the Wasserstein distance in [5].

Computing the sublevel set homology of $d_{\mu,m}$ in practice is difficult. In [7] the authors propose a way to approximate the sublevel set homology by using weighted complexes.

1.2.2 Christoffel polynomials [7]

In the paper [7], the authors investigate use of Christoffel polynomials to achieve robustness to outliers in the input data. Given a point cloud X, one is interested in the sublevel set persistent homology module $\mathbb{CD}(X, k)$ of the logarithm[1] of the Christoffel polynomial $P_k^{\mu_X}$ of degree k induced by the empirical measure on X. Stability results are given with respect to the 1-Wasserstein measure. The following stability result is established in [7]:

Proposition 1.3.1. Let X and Y be point clouds in $[-1, 1]^d$. Then,

$$d_B\left(\mathrm{Dgm}(\mathbb{CD}(X, k)), \mathrm{Dgm}(\mathbb{CD}(Y, k))\right) \leq \log(C_X d_W(\mu_X, \mu_Y) + 1)$$

whenever $C_X \cdot d_W(\mu_X, \mu_Y) \leq 1$. Here $C_X = 2C_{d,k} \|P_k^{\mu_X}\|_\infty$ where $C_{d,k} = 4\binom{d+k}{k} k^2$.

1.2.3 Kernel density estimation

Kernel density estimation is a non-parametric statistical technique for estimating the underlying probability density. It involves placing a kernel function,

[1]The sublevel sets of $\log P_k^{\mu_X}$ rather than $P_k^{\mu_X}$ are considered. The rationale given in the original paper is that this allows for stronger and more elegant stability results, and also that the logarithmic scaling produces persistence diagrams that better fit the underlying topology. See page 8 of [7].

which is typically a smooth, continuous curve such as a normal distribution, at each data point and summing all of these kernel functions to estimate the true density function. The smoothness of the resulting curve is controlled by a parameter called the bandwidth. Sublevel set persistent homology can be applied to a kernel density estimator (KDE) to identify topological features in the data.

Definition 1.4. Let x be a point in Y. A *kernel g centred in x* is a function $g\colon Y \to \mathbb{R}_{>0}$ with $g(y) = h(\|x-y\|)$ for some continuous function $h\colon [0,\infty) \to \mathbb{R}_{>0}$. In addition, we require h to be differentiable on $(0,\infty)$ with $h'(t) < 0$. We call h the *height function* associated to g.

Given a kernel g centred in 0, we denote by g_x the shifted kernel given by $g_x(y) = g(x-y)$. If X is a finite point cloud, we define the *kernel density estimator (KDE) on X* (with kernel g) to be the function $f_X = \frac{1}{n}\sum_{x\in X} g_x\colon Y \to \mathbb{R}_{>0}$. Showing that sublevel set homology of a kernel density estimator is uniformly stable is straight-forward.

Proposition 1.4.1. The function taking a point cloud X to the sublevel set persistent homology of f_X is uniformly stable.

Proof. Let f_X and f_{X_0} be the kernel density estimators with kernel g on X and X_0, respectively.

$$f_X - f_{X_0} = \frac{1}{n} g_{x_0} - \frac{1}{n(n-1)} \sum_{x\in X_0} g_x = \frac{1}{n}\left(g_{x_0} - f_{X_0}\right).$$

Consequently, we have that $\|f_X - f_{X_0}\|_\infty \in O(n^{-1})$ and hence uniform stability follows from classical stability of persistence diagrams (cf. [2]). □

A grid based approach to computing the sublevel set homology

One practical challenge with a KDE based approach is to compute the persistence homology. For low dimensions such as $d = 1$ and 2, a piecewise linear approximation using a grid can be used. However, this approach depends on the grid resolution, and the number of vertices increases exponentially in the number of dimensions which makes it unfeasible for higher dimensional data.

The following approximation scheme is used by [7] to compute the sublevel set homology of a Lipschitz continuous function $f\colon [-1,1]^d \to \mathbb{R}$:

1. Fix a positive integer m and construct the Freudenthal triangulation $\mathcal{K}_m$ of $[-1,1]^d$.

2. This gives us a vertex set, with $(m+1)^d$ vertices, equal to the lattice points of $\frac{2}{m} \cdot \mathbb{Z}^n$.

3. Evaluate f on each vertex and compute the persistent homology module of the lower-star filtration on $\mathcal{K}_m$ induced by these function values. Denote the persistence diagram of this persistence module by $\mathrm{Dgm}(f|_{\mathcal{K}_m})$.

The following proposition gives an upper bound on the error of this approximation in terms of the Lipschitz constant for f:

Proposition 1.4.2 (Proposition 22 [7]). Let $f\colon [-1,1]^d \to \mathbb{R}$ be Lipschitz continuous with Lipschitz constant L_f. Then

$$d_B\left(\mathrm{Dgm}(f|_{\mathcal{K}_m}), \mathrm{Dgm}(f)\right) \leq \frac{2L_f\sqrt{d}}{m}$$

Idea:

We know that all interesting topology of the sublevel sets of f_X is contained in the convex hull of X. If we consider a triangulation of the convex hull (e.g., the Delaunay triangulation or some refinement of it), can we give guarantees for the approximation using the lower-star filtration on this triangulation? (See proof and discussion in Appendix B of [7]).

Idea:

Can we first use a grid, and then apply some sparsification method (with approximation guarantees) to reduce the number of vertices?

References

[1] Olivier Bousquet and André Elisseeff. “Stability and generalization”. In: *The Journal of Machine Learning Research* 2 (2002), pp. 499–526.

[2] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. “Stability of persistence diagrams”. In: *Proceedings of the twenty-first annual symposium on Computational geometry*. 2005, pp. 263–271.

[3] Mickael Buchet et al. *Efficient and Robust Persistent Homology for Measures*. 2013. DOI: `10.48550/ARXIV.1306.0039`. URL: `https://arxiv.org/abs/1306.0039`.

[4] Ulrich Bauer and Michael Lesnick. “Induced matchings and the algebraic stability of persistence barcodes”. en. In: *Journal of Computational Geometry* (2015), Vol. 6 No. 2 (2015): Special issue of Selected Papers from SoCG 2014. DOI: `10.20382/JOCG.V6I2A9`. URL: `https://jocg.org/index.php/jocg/article/view/2983`.

[5] Hirokazu Anai et al. *DTM-based Filtrations*. 2018. DOI: `10.48550/ARXIV.1811.04757`. URL: `https://arxiv.org/abs/1811.04757`.

[6] Andrew J. Blumberg and Michael Lesnick. *Stability of 2-Parameter Persistent Homology*. 2020. DOI: `10.48550/ARXIV.2010.09628`. URL: `https://arxiv.org/abs/2010.09628`.

[7] Pepijn Roos Hoefgeest and Lucas Slot. *The Christoffel-Darboux kernel for topological data analysis*. 2022. DOI: `10.48550/ARXIV.2211.15489`. URL: `https://arxiv.org/abs/2211.15489`.