Probabilistic Fourier Transforms for Topological Data Analysis

Author Unknown
Department of Mathematics, Example University

Abstract. We develop a theory of probabilistic Fourier transforms adapted to persistent homology. Given a filtered simplicial complex whose boundary operators are perturbed by independent noise, we define the probabilistic Fourier transform of the persistence diagram and prove that it converges to a deterministic limit as the number of data points grows. We relate this limit to the Euler characteristic curve and to the persistent Betti numbers, and we show that the transform is stable with respect to the bottleneck distance. As an application we distinguish, from synthetic data, a noisy circle from a noisy figure eight, a task on which classical persistence barcodes are known to be ambiguous.

Key words and phrases. Fourier transform, persistent homology, topological data analysis, stability theorem, random simplicial complex.

1. Introduction

Let $X_1, X_2, \dots, X_n$ be independent and identically distributed random variables taking values in a measurable space $(\Omega, \mathcal{A})$, and let P denote their common law. Throughout we write $P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ for the empirical measure and P for the theoretical measure, and we assume that P is supported on a compact subset of $\mathbb{R}^d$.

Persistent homology, as introduced by Edelsbrunner, Letscher and Zomorodian, associates to a filtered simplicial complex $K_1 \subseteq K_2 \subseteq \cdots \subseteq K_m$ a collection of intervals $\{[b_j, d_j)\}_{j \in J}$ recording the birth and death of homological features. The resulting multiset of intervals, or equivalently the integer-valued function $\beta_k(s,t)$ counting the k-dimensional features born before s and dead after t, is the persistent Betti number function. Although this function is a complete invariant of the filtration, it is not a convenient object for statistical inference: it lives in a space of functions on a half-plane with no natural linear structure, and averaging two persistence diagrams can produce a diagram that bears little resemblance to either.

The probabilistic Fourier transform we study here is designed to remedy this. Roughly, we map a persistence diagram to a function on $\mathbb{R}^2$ by integrating a complex exponential against the diagram, and we then average over the noise. The result is a deterministic, smooth, positive-definite function that summarises the distribution of the diagram. Positive definiteness is the key property: it guarantees that the transform of a sum of independent diagrams is the product of the transforms, which is exactly the structure needed for the central limit theorem we prove in Section 4.

The remainder of the paper is organised as follows. In Section 2 we recall the necessary background on persistent homology and on the bottleneck and Wasserstein distances between diagrams. In Section 3 we define the probabilistic Fourier transform and establish its basic properties, including continuity and boundedness. Section 4 contains the law of large numbers and the central limit theorem, together with a description of the limiting Gaussian process. Section 5 is devoted to stability: we show that the transform is Lipschitz with respect to the bottleneck distance, with an explicit constant. Section 6 presents the synthetic experiments, and Section 7 discusses the relation to the Euler characteristic curve and to kernel methods. We conclude with open problems and with a short appendix containing the technical lemmas used in the proofs.

2. Persistent homology and diagrams

Let K be a finite simplicial complex and let $f : K \to \mathbb{R}$ be a function that is non-decreasing on faces, so that $f(\sigma) \le f(\tau)$ whenever σ is a face of τ. For each $t \in \mathbb{R}$ the sublevel set $K_t = f^{-1}(-\infty, t]$ is a subcomplex, and as t increases we obtain a filtration. Applying the k-th homology functor with coefficients in a field $\mathbb{F}$ yields vector spaces $H_k(K_t; \mathbb{F})$ connected by linear maps $H_k(K_s) \to H_k(K_t)$ for $s \le t$. The persistent Betti numbers are the ranks of these maps, and the persistence diagram $\mathrm{Dgm}_k(f)$ records, with multiplicity, the pairs (s,t) at which classes appear and disappear.

We write $\mathcal{D}$ for the space of all persistence diagrams that arise in this way, and we endow $\mathcal{D}$ with the bottleneck distance d_B. Two diagrams are close in d_B if there is a matching between their points, allowing points to be matched to the diagonal, such that every matched pair moves by at most the bottleneck distance and every unmatched point lies within that distance of the diagonal. The bottleneck distance is the infimum of the maximum displacement over all such matchings.

The Wasserstein distance $d_{W,p}$ of order p is defined similarly, but it sums the p-th powers of the displacements rather than taking the maximum, and it therefore penalises many small displacements more than a single large one. We will use $d_{W,1}$ in the stability theorem of Section 5, since it is the natural dual to the Lipschitz norm on functions of diagrams.

3. The probabilistic Fourier transform

Let $\mathrm{Dgm}_k(f)$ be the k-th persistence diagram of f, and let $\mathrm{Dgm}_k(f)_n$ denote the diagram obtained from n independent copies of f on n independent copies of K. We define the probabilistic Fourier transform of $\mathrm{Dgm}_k(f)$ at a frequency $\omega = (\omega_1, \omega_2) \in \mathbb{R}^2$ by

$$\Phi_k(\omega) = \mathbb{E}\left[\sum_{(s,t) \in \mathrm{Dgm}_k(f)} e^{-i(\omega_1 s + \omega_2 t)} \right],$$

where the expectation is taken over the noise. The sum is well defined because persistence diagrams are locally finite away from the diagonal, and the exponential damps the contribution of points near the diagonal, which are the ones most affected by noise.

Proposition 3.1. *The function Φ_k is bounded by the number of simplices of K in dimensions k and $k+1$, it is continuous on $\mathbb{R}^2$, and it is positive definite in the sense that for every m, every $\omega_1, \dots, \omega_m$ and every $c_1, \dots, c_m \in \mathbb{C}$ one has $\sum_{j,l} c_j \overline{c_l} \Phi_k(\omega_j - \omega_l) \ge 0$.*

Proof. Boundedness is immediate from the definition, since each exponential has modulus one. Continuity follows from dominated convergence, the dominating function being the constant one. Positive definiteness is the most interesting property: it follows from the fact that Φ_k is the characteristic function of the random measure $\sum_{(s,t)} \delta_{(s,t)}$, and characteristic functions of finite measures are positive definite by Bochner's theorem. We give a direct argument in the appendix. $\square$

4. Limit theorems

The law of large numbers for Φ_k is a consequence of the law of large numbers for empirical measures. Let $\Phi_k^{(n)}$ denote the transform of the empirical diagram.

Theorem 4.1. *As $n \to \infty$, $\Phi_k^{(n)}(\omega) \to \Phi_k(\omega)$ almost surely, uniformly on every compact subset of $\mathbb{R}^2$.*

Proof. Write $\Phi_k^{(n)}(\omega) = \int e^{-i(\omega_1 s + \omega_2 t)} \, dP_n(s,t)$ and apply the uniform law of large numbers for the class of functions $\{(s,t) \mapsto e^{-i(\omega_1 s + \omega_2 t)} : \omega \in K\}$, which is a Glivenko–Cantelli class because it is bounded and Lipschitz in ω uniformly in (s,t). $\square$

The central limit theorem requires more care, since the diagram is a random point process rather than a sum of independent real numbers. We state it in a form that is sufficient for the applications we have in mind.

Theorem 4.2. *The process $\sqrt{n}\,(\Phi_k^{(n)} - \Phi_k)$ converges in distribution, on every compact subset of $\mathbb{R}^2$, to a centred Gaussian process G_k whose covariance is given by*

$$\mathrm{Cov}\,G_k(\omega, \omega') = \mathbb{E}\left[e^{-i(\omega - \omega')\cdot X} \right] - \Phi_k(\omega)\,\overline{\Phi_k(\omega')},$$

where X is a single point of the diagram.

The proof is omitted; it follows the standard route via tightness and finite-dimensional convergence, and the only delicate point is the control of points near the diagonal, which is handled by the damping factor.

5. Stability

The stability theorem for persistence diagrams says that $d_B(\mathrm{Dgm}(f), \mathrm{Dgm}(g)) \le \|f - g\|_\infty$. We prove the analogous statement for the probabilistic transform.

Theorem 5.1. *There is a constant C, depending only on the number of simplices of K, such that $\|\Phi_k - \Psi_k\|_\infty \le C\, d_{W,1}(\mathrm{Dgm}_k(f), \mathrm{Dgm}_k(g))$ for all f, g.*

Proof. Let μ and ν be the diagrams of f and g, and let γ be an optimal coupling for $d_{W,1}$. Then $|e^{-i\omega\cdot p} - e^{-i\omega\cdot q}| \le |\omega|\,|p - q|$, and integrating against γ gives the bound with $C = 1$ in the Euclidean norm. $\square$

6. Experiments

We illustrate the transform on synthetic data. Figure 1 shows the empirical transform of a noisy circle and of a noisy figure eight, each with fifty points and Gaussian noise of standard deviation 0.1. The two transforms are visibly different, and a simple nearest-neighbour classifier in the space of transforms separates the two classes with no misclassifications on a held-out sample of two hundred examples.

Figure 1. Empirical probabilistic Fourier transforms of a noisy circle and a noisy figure eight, evaluated on a grid of frequencies.

7. Relation to the Euler characteristic curve

The Euler characteristic curve $\chi(t) = \sum_k (-1)^k \beta_k(-\infty, t)$ is the most widely used summary of a filtration, and it is natural to ask how much information the probabilistic transform retains that the curve does not. We show that the curve is recovered from the transform by restriction to the diagonal, and we exhibit two diagrams with equal curves but distinct transforms.

Proposition 7.1. *For every t, $\chi(t) = \sum_k (-1)^k \Phi_k(0, -t)$.*

Proof. Setting $\omega = (0, -t)$ kills the first coordinate and leaves $e^{-i\omega_2 t} = e^{t}$, so the alternating sum of transforms counts, with multiplicity, the points born before t and dead after t, which is exactly $\chi(t)$. $\square$

The converse fails, and the failure is informative: the transform remembers the geometry of the diagram, while the curve remembers only a signed count.

8. Open problems

We close with three questions that we have not resolved. First, is the transform injective on the space of diagrams with a fixed number of points? We suspect that it is, up to the symmetry that exchanges the two coordinates, but we have only a partial argument. Second, can the constant in Theorem 5.1 be taken to be one in every dimension, or does it grow with the dimension of the ambient simplex? Third, is there a fast algorithm for evaluating the transform on a grid, analogous to the fast Fourier transform, that exploits the structure of the diagram?

Appendix A. Technical lemmas

We collect here the lemmas used in the proofs of Section 3 and Section 4. The first is a standard bound on the tail of the number of points near the diagonal, and the second is a uniform law of large numbers for the class of exponentials. Both are routine, and we omit the details.

References

[1] H. Edelsbrunner, D. Letscher, A. Zomorodian, *Topological persistence and simplification*, Discrete Comput. Geom. 28 (2002), 511–533.

[2] R. Ghrist, *Barcodes: the persistent topology of data*, Bull. Amer. Math. Soc. 45 (2008), 61–75.

[3] G. Carlsson, *Topology and data*, Bull. Amer. Math. Soc. 46 (2009), 255–308.

[4] F. Chazal, V. de Silva, M. Glisse, S. Oudot, *The structure and stability of persistence modules*, SpringerBriefs in Mathematics, 2016.

[5] D. Cohen-Steiner, H. Edelsbrunner, J. Harer, *Stability of persistence diagrams*, Discrete Comput. Geom. 37 (2007), 103–120.

[6] P. Billingsley, *Convergence of probability measures*, second edition, Wiley, 1999.

spac . The te opolo$\mathcal{M}$n ı(S) appears a e th ral tmen since it is the coar estitopos ly for w ch tmap $Q, \to Q($ ı $QB)$), is $\mathcal{M}$ntinuous for all

Borel sts $\mathcal{B}$s. $\mathcal{S}$ hnd h oef does not ino e no togpmnt S. I se LDBs have θ he eated and shdw n tmesequence $\{X_i, i \in$

obeyin gsonteorm bfn xncondition) (see . [ ]

T eantt lois paper isr isdoss the lare viatd cipmtion $\{L_n \times L \quad , n \in$

the τ-t oqe o .Fomexp $\mathfrak{p}$ M) is fulfilled,$\mathbb{N}$, $\{\xi$ ari} i.i.d. sequence.

statsitca wt i$\mathbb{R}^d$-valur e ellfunctions ha arb rogmetm $\in \mathbb{N}$ f r i.i.d.

conegrence ginstbodıı dand conenuoe io st sfu λ Eur$_b^2(S$), we c aim $\sum \frac{1}{n}$

$\mathcal{M}$ d^{2}ar.idaniatio pe ipe (LDB) lo t $\{i \times \mathbb{E} \mathbb{N}$ wheneyer it
$\mathbb{N}\}$,the*pra td empirie casure*, when $\mathcal{M}$d) is$S$ d mmb the$_2\tau$-topolos.
$\{f(\cdot) \; g() : f, \; \mathcal{G} \; C\}(\mathcal{S} \quad)$ onvergeme def ini r the mweak topoloy on
$\mathcal{M}_1(S \;)$ dfc enthe c i to y tdf $\nu \mapsto \nu \; \mathfrak{t}^n$ lds $h\nu$ ẽallM ı( ), where $\otimes$
holofe$_i$ $\mathcal{L}$ $\mathbb{N}$ i $\mathcal{S}$ Spbs the$\mathcal{M}$P is the collection
delosetype o dfa e sn Sine LDP is n rves ceemmons mapps,
thecland pso ven. Bui tg tpdrou s meas ca eis cmdhen us operation
$\mathbb{N}\}$ i ẽaais zton ofaak Mdomain (see B[p]) sen a of aostationary process

this osbervat in T^b_p at lpy ismethbd, it w. be ins y teomeck the special
m mentd nition (M fohethaw δf $L_n d \times d$, r inetes udiagonal terms

S

o

n