REVIEW OF ESTIMATION AND QUANTIZATION OF EXPECTED PERSISTENCE DIAGRAMS

YVES LECONTE

Abstract. The article Estimation and Quantization of Expected Persistence Diagrams [1] tackles the theoretical study and practical development of statistical estimation of Expected Persistence Diagrams (EPD). The EPD is used to summarise information from Persistence Diagrams (PDs), which are commonly used in Topological Data Analysis (TDA).

Contents

Introduction

A persistence diagram (PD) is a discrete measure $\sum_{i\in I}\delta_{x_i}$ supported on $\Omega = \{(t_1,t_2)\in\mathbb{R}^2, t_2 > t_1\}$, where the x-axis represents the birth moment of a topological feature and the y-axis the death of a topological feature. The metric of distance that is considered in this space is the optimal partial transport metric denoted OT_p with $1 \leq p \leq \infty$. More precisely, given two measures μ and ν,

$$OT_p(\mu,\nu) := \inf_{\pi\in \mathrm{Adm}(\mu,\nu)} \left(\int\int_{\bar{\Omega}\times\bar{\Omega}} ||x-y||^p \, d\pi(x,y) \right)^{1/p}$$

for $p < \infty$, and

$$OT_\infty(\mu,\nu) := \sup_{(x,y)\in supp(\pi)} ||x-y||$$

when $p = \infty$, which is then called the bottleneck distance. The space $(\mathcal{M}^p, OT_p)$ of persistence measures is the space of Radon measures $\mu \in \Omega$ such that $\int ||x-\partial\Omega||^p \, d\mu(x) < \infty$, and thus in this space $OT_p < \infty$.

The framework studied in the article is the following, we consider samples of points in $\mathbb{R}^d$, from which we can derive PDs. Resulting is a sample of PDs denoted as $\mu_1, ...\mu_n$ on Ω, following an unknown distribution of probability P. There are various well-known approaches to analyse our original dataset, while fewer to analyse the resulting sample of PDs. Moreover, the considered space is a metric space but not a vector space, which bottlenecks the development of common data analysis and statistical estimation techniques. Consider P a probability distribution on $(\mathcal{M}^p, OT_p)$, and $A \subset \Omega$ compact,

$$E(P)(A) := \mathbb{E}_p[\mu(A)]$$

where $\mu \sim P$ and $\mu(A)$ represents the number of observations of μ in A. We want to empirically study the expectation $E(P)$, it is thus natural to consider the estimator $\bar{\mu}_n = \frac{1}{n}\sum_{i=1}^n \mu_i$ of the EPD $E(P)$. This idea of approximating a continuous measure with a measure of discrete and fixed size is called quantization, here being a quantization of a sequence of PDs.

Date: February 2024.

Estimation of EPD

Using the law of large numbers, is proven that

$$OT_p^p(\bar{\mu}_n, E(P)) \longrightarrow 0$$

a.s. when $\mathbb{E}_P\left[\int ||x - \partial\Omega||^p \, d\mu(x)\right] < \infty$. It is then of natural interest to study the convergence rate of the estimator $\bar{\mu}_n$.

Let $\mathcal{P}_{L,M}^q$ the set of probability distributions supported on $\mathcal{M}_{L,M}^q$, and $P \in \mathcal{P}_{L,M}^p$, $1 \leq p < \infty$, $0 \leq q < p$, $\mu_1, \ldots, \mu_n$ and n-sample following the distribution P with $\bar{\mu}_n$ the associated estimator of the EPD, then

$$\mathbb{E}\left[OT_p^p(\bar{\mu}_n, E(P))\right] \leq c_{p,q} M L^{p-q} \left(\frac{1}{n^{1/2}} + \frac{a_p(n)}{n^{p-q}}\right)$$

with $a_p(n) = 1$ when $p > 1$ and $a_p(n) = log(n)$ when $p = 1$.
This result implies that the considered estimator converges at speed $\sqrt{n}$ when $p \geq q + 1/2$. From this first convergence rate, we can wonder whether this estimator is optimal or not. Following this idea, we can remind that the minimax rate of estimating $E(P)$ on $\mathcal{P}$ is given by

$$\mathcal{R}_n(\mathcal{P}) := \inf_{\hat{\mu}_n} \sup_{P \in \mathcal{P}} \mathbb{E}\left[OT_p^p(\hat{\mu}_n, E(P)\right]$$

and we can show that indeed when $p \geq q + 1/2$ then $\bar{\mu}_n$ is a minimax estimator on $\mathcal{P}_{L,M}^q$.
Let $1 \leq p < \infty$ and $q \geq 0$, L,$M > 0$, we have for a certain c,

$$\mathcal{R}_n\left(\mathcal{P}_{L,M}^q\right) \geq c_{p,q} M L^{p-q} n^{-1/2}.$$

We can notice that we didn't further assumed any regularity assumption on the EPD, and still the estimator $\bar{\mu}_n$ is optimal from a minimax perspective using the OT_p loss.

Quantization of EPD

The goal of this section is to minimize the quantity

$$((m_1, c_1), \ldots, (m_k, c_k)) \mapsto OT_p\left(\sum_j m_j \delta_{c_j}, \mu\right)$$

with $m_j \in \mathbb{R}_+$ and $c_j \in \Omega$. In other words, we try to approximate the continuous measure μ using a discrete and finite set of points. Additionally, we have that given a persistence measure μ and a codebook $c = (c_1, ..., c_k)$, it is optimal to choose m_j such that $m_j = \mu\left(V_j(c)\right)$. More precisely, we have that for $\hat{\mu}(c) := \sum_{j=1}^k \mu\left(V_j(c)\right) \delta_{c_j}$ and $\nu = \sum_{j=1}^k m_j \delta_{c_j}$ with $m_j \geq 0$,

$$OT_p\left(\hat{\mu}(c), \mu\right) \leq OT_p\left(\nu, \mu\right).$$

Following this result, we do have that taking $1 \leq p < \infty$, the above problem can be rewritten as the minimization of

$$\begin{aligned} R_{k,p}(c) :&= OT_p\left(\hat{\mu}(c), \mu\right) \\ &= \left(\sum_{j=1}^{k+1} \int_{V_j(c)} ||x - c_j||^p \, d\mu(x)\right)^{1/p}. \end{aligned}$$

The existence of such minimizer, meaning of an optimal codebook, is proven in the paper.
However, trying to solve this minimization problem using algorithms based on optimal transport literature isn't feasible as empirical EPD has large number of points, which impacts computational efficiency. Additionally, for our algorithm to be tractable with large sequences of large diagrams, we need to have an online type of algorithm leveraging the fact that we have a sequence of diagrams.

The alternative algorithm proposed in [1] is the following,

Algorithm 1 Online quantization of EPDs

A sequence $\mu_1, \dots, \mu_n$, integer k, parameter p.
Divide indices $\{1, \dots, n\}$ into batches $(B_1, \dots, B_T)$ of size $(n_1, \dots, n_T)$. Furthermore, divide $(B_t)_t$ into two halves $B_t^{(1)}$ and $B_t^{(2)}$.
Set $\bar{\mu}_t^{(\alpha)} = \frac{2}{n_t} \sum_{i \in B_t^{(\alpha)}} \mu_i$ for $1 \leq t \leq T$, $\alpha \in \{1, 2\}$.
Sample $c_1^{(0)} \dots c_k^{(0)}$ from the diagrams.
for $t = 0, \dots, T-1$ **do**
    $c^{(t+1)} = U_p(t, c^{(t)}, \bar{\mu}_{t+1}^{(1)}, \bar{\mu}_{t+1}^{(2)})$
end for
The final codebook $c^{(T)}$.

where for $p > 1$,

$$U_p(t, c, \mu, \mu') := c - \frac{\left(\frac{\mu(V_j(c))}{\mu'(V_j(c))}(c_j - v_p(c, \mu)_j)\right)_j}{t+1},$$

where $v_p(c, \mu)_j$ is the p-center of mass of μ over the cell $V_j(c)$:

$$v_p(c, \mu)_j := \arg\min_y \left(\int_{V_j(c)} ||y - x||^p d\mu(x) \right)^{\frac{1}{p}}.$$

The idea behind this algorithm is that we are pushing c_j toward the point that would decrease $R_{k,p}$ over the cell $V_j(c)$ the most with a learning rate of $\frac{1}{1+t}$. As previously mentioned, considering the case where $p = \infty$ is of main interest in TDA as we often want to use the bottleneck distance OT_∞. The speed complexity of this algorithm is linear w.r.t. the points of μ belonging to $V_j(c)$ when μ is a discrete measure.

Conclusion

Throughout this article, was presented a thorough study of the optimal estimation of expected persistence diagrams as well as the quantization of such problem. The online algorithm that was introduced is very convenient as we can even use it to handle the case where $p = \infty$, and additionally it doesn't require us to adjust any hyper parameter linked to the diagonal.

References

[1] Vincent Divol and Théo Lacombe. Estimation and quantization of expected persistence diagrams, 2021.