On barcodes and approximation of distributions

Serguei Barannikov

On barcodes and approximation of distributions.

Serguei Barannikov [1 2]

Abstract

We describe a novel tool, Barcode(P,Q), that, given a pair of distributions in a high-dimensional space, tracks multiscale topology spacial discrepancies between manifolds on which the distributions are concentrated.

1. Introduction

Reconstruction of the data distribution from observing only a subset of its points has made a significant step forward since the invention of Generative Adversarial Networks (Goodfellow et al., 2014). Deep generative networks try to model a distribution which is as similar as possible to the true data distribution. Despite the exceptional success that was achieved by the deep generative models, there exists a longstanding challenge of good assessment of the generated samples quality and diversity. We describe a framework, based on Cross-Barcode(P,Q), a novel tool aiming at attacking this problem. For a pair of point clouds the Cross-Barcode(P,Q) describes the differences in multiscale manifold topology between the manifolds approximated by the two clouds.

2. Cross-Barcode and Manifold Topology Divergence.

2.1. Multiscale simplicial approximation of manifolds.

According to the well-known Manifold Hypothesis (Goodfellow et al., 2016) the support of the data distribution $\mathcal{P}_{\text{data}}$ is often concentrated on a low-dimensional manifold M_{data}. We construct a framework for comparing numerically such distribution $\mathcal{P}_{\text{data}}$ with a similar distribution $\mathcal{Q}_{\text{model}}$ concentrated on a manifold M_{model} produced by a generative neural network. The immediate difficulty here is that the manifold M_{data} is unknown and is described only through discrete sets of samples from the distribution $\mathcal{P}_{\text{data}}$. One standard approach to resolve this difficulty is to approximate the manifold M_{data} by simplices with vertices given by the sampled points. The simplices approximating the manifold are picked based on proximity information given by the pairwise distances between sampled points. The standard approach is to fix a threshold $r > 0$ and to take the simplices with vertices that are sufficently close to each other in comparison with the threshold r. The choice of threshold is essential here since if it is too small, then only the initial points, i.e., separated from each other 0-dimensional simplices, are allowed. And if the threshold is too large, then all possible simplices with sampled points as vertices are included and their union is simply the big blob representing the convex hull of the sampled points. Instead of trying to guess the right value of the threshold, the standard recent approach is to study all thresholds at once. This can be achieved thanks to the mathematical tool, called barcode, that quantifies the evolution of topological features over multiple scales. For each value of r the barcode describes the topology, namely the numbers of holes or voids of different dimensions, of the union of all simplicies included up to the threshold r.

2.2. Measuring the differences in simplicial approximation of two manifolds

However, to estimate numerically the degree of similarity between the manifolds $M_{\text{model}}, M_{\text{data}} \subset \mathbb{R}^D$, it is important not just to know the numbers of topological features across different scales for simplicial complexes approximating $M_{\text{model}}, M_{\text{data}}$, but to be able to verify that the similar topological features are located at similar places and appear at similar scales.

Our method measures the differences in the simplicial approximation of the two manifolds, represented by samples P and Q, by constructing sets of simplices, describing discrepancies between the two manifolds. To construct these sets of simplices we take the edges connecting $P-$points with $Q-$points, and also $P-$points between them, ordered by their length, and start adding these edges one by one, beginning from the smallest edge and gradually increasing the threshold. We add also the triangles and $k-$simplices at the threshold when all their edges have been added. It is assumed that all edges between $Q-$points were already in the initial set. We track in this process the birth and death of topological features, where the topological features are

[1]Skolkovo Institute of Science and Technology, Moscow, Russia. [2]IMJ, Paris University, France. Correspondence to: <Serguei.Barannikov@imj-prg.fr>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

allowed here to have boundaries on any simplices formed by $Q-$points. The longer the lifespan of the topological feature across the change of threshold the bigger the described by this feature discrepancy between the two manifolds.

Homology is a tool that permits to single out topological features that are similar, and to decompose any topological feature into a sum of basic topological features. More specifically, we use the homology of a pair. In our case, a $k-$cycle is a collection of $k-$simplices formed by $P-$ and $Q-$ points, such that their boundaries cancel each other, perhaps except for the part of boundaries consisting of simplices formed only by $Q-$points. For example, a cycle of dimension $k = 1$ corresponds to a path connecting a pair of $Q-$points and consisting of edges passing through a set of $P-$points. A cycle which is a boundary of a set of $(k+1)-$simplices is considered trivial. Two cycles are topologically equivalent if they differ by a boundary, and by collection of simplices formed only by $Q-$points. A union of cycles is again a cycle. Each cycle can be represented by a vector in the vector space where each simplex corresponds to a generator. In practice, the vector space over $\{0, 1\}$ is used most often. The union of cycles corresponds to the sum of vectors. The homology vector space H_k is defined as the factor of the vector space of all $k-$cycles modulo the vector space of boundaries and cycles consisting of simplices formed only by $Q-$points. A set of vectors forming a basis in this factor-space corresponds to a set of basic topological features, so that any other topological feature is equivalent to some partial sum of the basic features.

The homology are also defined for manifolds and for arbitrary topological spaces. This definition is technical and we have to omit it due to limited space, and to refer to e.g. (Hatcher, 2005; Moraleda et al., 2019) for details. The relevant properties for us are the following. For each topological space X the vector spaces $H_k(X)$, $k = 0, 1 \ldots$, are defined. The dimension of the vector space H_k equals to the number of independent $k-$dimensional topological features (holes,voids etc). An inclusion $Y \subset X$ induces a natural map $H_k(Y) \to H_k(X)$

In terms of homology, we would like to verify that not just the dimensions of homology groups $H_*(M_{\rm model})$ and $H_*(M_{\rm data})$ are the same but that more importantly the natural maps:

$$\varphi_r : H_*(M_{\rm model} \cap M_{\rm data}) \to H_*(M_{\rm model}) \quad (1)$$
$$\varphi_p : H_*(M_{\rm model} \cap M_{\rm data}) \to H_*(M_{\rm data}) \quad (2)$$

induced by the embeddings are as close as possible to isomorphisms. The homology of a pair is precisely the tool that measures how far such maps are from isomorphisms. Given a pair of topological spaces $Y \subset X$, the homology of a pair $H_*(X, Y)$ counts the number of independent topological features in X that cannot be deformed to a topological feauture in Y plus independent topological features in Y that, after the embedding to X, become deformable to a point. An equivalent description, the homology of a pair $H_*(X, Y)$ counts the number of independent topological features in the factor-space X/Y, where all points of Y are contracted to a single point. The important fact for us is that the map, induced by the embedding, $H_*(Y) \to H_*(X)$ is an isomorphism if and only if the homology of the pair $H_*(X, Y)$ are trivial. Moreover the embedding of simple simplicial complexes $Y \subset X$ is an equivalence in homotopy category, if and only if $H_*(X/Y)$ are trivial (Whitehead, 1968).

To define the counterpart of this construction for a pair of manifolds represented by point clouds, we employ the following strategy. Firstly, we replace the pair $(M_{\rm model} \cap M_{\rm data}) \subset M_{\rm model}$ by the equivalent pair $M_{\rm model} \subset (M_{\rm data} \cup M_{\rm model})$ with the same factor-space. Then, we represent $(M_{\rm data} \cup M_{\rm model})$ by the union of point clouds $P \cup Q$, where the point clouds P, Q are sampled from the distributions $\mathcal{P}_{\rm data}$, $\mathcal{Q}_{\rm model}$. Our principal claim here is that taking topologically the quotient of $(M_{\rm data} \cup M_{\rm model})$ by $M_{\rm model}$ is equivalent in the framework of multiscale analysis of topological features to the following operation on the matrix $m_{P\cup Q}$ of pairwise distances of the cloud $P \cup Q$: **we set to zero all pairwise distances within the subcloud $\mathbf{Q} \subset (P \cup Q)$.**

2.3. Cross-Barcode(P,Q)

Let $P = \{p_i\}$, $Q = \{q_j\}$, $p_i, q_j \in \mathbb{R}^D$ are two point clouds sampled from two distributions $\mathcal{P}$, $\mathcal{Q}$. To define Cross-Barcode(P, Q) we construct first the following filtered simplicial complex. Let $(\Gamma_{P\cup Q}, m_{(P\cup Q)/Q})$ be the metric space defined as the metrized complete graph on the union of point clouds $P \cup Q$ with the distance matrix given by the pairwise distance in $\mathbb{R}^D$ for the pairs of points (p_i, p_j) or (p_i, q_j) and with all pairwise distances within the cloud Q that we set to zero. Our filtered simplicial complex is the Vietoris-Rips complex of $(\Gamma_{P\cup Q}, m_{(P\cup Q)/Q})$.

Recall that given such a graph Γ with matrix m of pairwise distances between vertices and a parameter $\alpha > 0$, the Vietoris-Rips complex $R_\alpha(\Gamma, m)$ is the abstract simplicial complex with simplices that correspond to the non-empty subsets of vertices of Γ whose pairwise distances are less than α as measured by m. Increasing parameter α adds more simplices and this gives a nested family of collections of simplices know as filtered simplicial complex. Recall that a **simplicial complex** is described by a set of vertices $V = \{v_1, \ldots, v_N\}$, and a collection of simplices S, where a $k-$simplex is defined as a $(k+1)-$elements subset of the set of vertices V. The set of simplices S should satisfy the condition that for each simplex $s \in S$ all the $(k-1)-$simplices obtained by the deletion of a vertex from the sub-

set of vertices of s belong also to S. The **filtered simplicial complexes** is the family of simplicial complexes S_α with nested collections of simplices: for $\alpha_1 < \alpha_2$ all simplices of S_{α_1} are also in S_{α_2}.

At the initial moment, $\alpha = 0$, the simplicial complex $R_\alpha(\Gamma_{P\cup Q}, m_{(P\cup Q)/Q})$ has trivial homology H_k for all $k > 0$ since it contains all simplices formed by $Q-$points. The dimension of the $0-$th homology equals at $\alpha = 0$ to the number of $P-$points, since no edge between them or between a $P-$point and a $Q-$point is added at the beginning. As we increase α, some cycles, holes or voids appear in our complex R_α. Then, some combinations of these cycles disappear. The **persistent homology** principal theorem (Barannikov, 1994) states that it is possible to choose the set of generators in the homology of filtered complexes $H_k(R_\alpha)$ across all the scales α such that each generator appears at its specific "birth" time and disappears at its specific "death" time. These sets of "birth" and "deaths" times of topological features in R_α are registered in **Barcode** of the filtered complex. The Cross-Barcode$_i(P, Q)$ is thus a list of intervals consisting of the "birth" and "deaths" times of $i-$dimensional topological features in the filtered simplicial complex $R_\alpha(\Gamma_{P\cup Q}, m_{(P\cup Q)/Q})$. Topological features with longer "lifespan" are considered essential, and topological features with a short "lifespan" are considered less essential. The topological features with "birth"="death" are trivial by definition and do not appear in Cross-Barcode$_*(P, Q)$.

2.4. Cross-Barcode$_*(P, Q)$ as obstructions to assigning P points to distribution $\mathcal{Q}$

Geometrically, the lowest dimensional Cross-Barcode$_0(P, Q)$ is the record of relative hierarchical clustering of the following form. For a given threshold r, let us consider all points of the point cloud Q plus the points of the cloud P lying at a distance less than r from a point of Q as belonging to the single $Q-$cluster. It is natural to form simultaneously other clusters based on the threshold r, with the rule that if the distance between two points of P is less than threshold r then they belong to the same cluster. When the threshold r is increased, two or more clusters can collide. And the threshold, at which this happens, corresponds precisely to the "death" time of one or more of the colliding clusters. At the end, for very large r only the unique $Q-$cluster survives. Then Cross-Barcode$_0(P, Q)$ records precisely the survival times for this relative clustering.

Notice that in certain situations, like, for example, in Figure 1, it is difficult to attribute confidently certain points of P to the same distribution as the point cloud Q even when they belong to the "big" $Q-$cluster at a small threshold r, because of the nontrivial topology. Such "paths/membranes" of $P-$points in void space, are obstacles

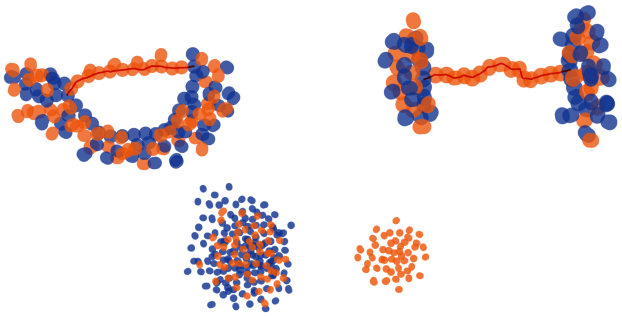

Figure 1. Paths/membranes (red) in the void that are formed by small intersecting disks around P points (orange), and are ending on Q (blue), are obstacles for identification of the distribution $\mathcal{P}$ with $\mathcal{Q}$. These obstacles are quantified by Cross-Barcode$_1(P, Q)$. Separate clusters are the obstacles quantified by Cross-Barcode$_0(P, Q)$.

for assigning points from P to distribution $\mathcal{Q}$. These obstacles are quantified by the segments from the higher barcodes $\text{Cross} - \text{Barcode}_{\geq 1}(P, Q)$. The bigger the length of the associated segment in the barcode, the further the membrane passes away from Q.

2.5. Basic properties of Cross-Barcode$_*(P, Q)$.

Here is the list of basic propertis of Cross-Barcode$_*(P, Q)$:

- If the two clouds coincide then Cross-Barcode$_*(P, P) = \varnothing$
- Factoring by the empty cloud changes nothing and Cross-Barcode$_*(P, \varnothing) =$ Barcode$_*(P)$ the usual barcode of the point cloud itself
- We have verified empirically the diminishing of Cross-Barcode$_*(Q_1, Q_2)$ when number of points in Q_1, Q_2 goes to $+\infty$ and Q_1, Q_2 are sampled from the same uniform distribution on the 2D disk of radius 1.

2.6. The Manifold Topology Divergence (MTop-Div)

Since Cross-Barcode$_*(P, P) = \varnothing$, the closeness of Cross-Barcode$_*(P, Q)$ to the empty set is a measure of discrepancy between $\mathcal{P}$ and $\mathcal{Q}$. Various numerical characteristics capture this discrepancy.

Each Cross-Barcode$_i(P, Q)$ is a multiset of intervals describing the persistent homology H_i. To measure the closeness to the empty set, one can use: number of segments, sum of lengths of segments, sum of squared lengths of segments, the longest segment, specific quantile of segments lengths or histogram of relative living time and its distance to the histogram of the empty barcode. We assume that various characteristics of different H_i could be useful in

various cases, but the cross-barcodes for H_0 and H_1 can be calculated relatively fast.

Our **MTop-Divergency**$(\mathcal{P}, \mathcal{Q})$ is based on the sum of lenghts of segments in Cross-Barcode$_1(P, Q)$, see section 2.7 for details.

The sum of lengths of segments in Cross-Barcode$_1(P, Q)$ has an interesting interpretation via the Earth Moving Distance. Namely, it is easy to prove that EM-Distance between the Relative Living Time histogram for Cross-Barcode$_1(P, Q)$ and the histogram of the empty barcode, multiplied by the parameter α_{max} from the definition of RLT, see e.g. (Khrulkov & Oseledets, 2018), for α_{max} bigger than all the "death" times in H_1, coincides with the sum of lengths of segments in H_1. This ensures the standard stability properties of this quantity.

Our metrics can be applied in two settings: to a pair of distributions $\mathcal{P}_{\text{data}}$, $\mathcal{Q}_{\text{model}}$, in which case we denote our score MTop-Div(D,M) and to a pair of distribtions $\mathcal{Q}_{\text{model}}$, $\mathcal{P}_{\text{data}}$, in which case our score is denoted MTop-Div(M,D). These two variants of the Cross-Barcode, and of the MTop-Divergency are related to the concepts of precision and recall, we leave exploration of this analogy for further research.

2.7. Algorithm

To calculate the score that evaluates the similitude between two distributions, we employ the following algorithm. First, we compute Cross-Barcode$_1(P, Q)$ on point clouds P, Q of sizes b_P, b_Q sampled from the two distributions $\mathcal{P}$, $\mathcal{Q}$. For this we calculate the matrices m_P, $m_{P,Q}$ of pairwise distances within the cloud P and between clouds P and Q. Then the algorithm constructs the Vietoris-Rips filtered simplicial complex from the matrix $m_{(P\cup Q)/Q}$ which is the matrix of pairwise distances in $P \cup Q$ with the pairs of points from cloud Q block replaced by zeroes and with other blocks given by m_P, $m_{P,Q}$. Next step is to calculate the barcode of the constructed filtered simplicial complex. This step and the previous step constructing the filtered complex from the matrix $m_{(P\cup Q)/Q}$ can be done using one of the fast scripts[1], some of them are optimized for GPU acceleration, e.g.(Zhang et al., 2020). The calculation of barcode is based on the persistence algorithm (Barannikov, 1994). Next, one of the numerical characteristcs of Cross-Barcode$_1(P, Q)$ is computed. Using numerical tests, the H1sum characteristic was picked as the principal, however depending on the situation, other characteristics can also be interesting to compute, like H1max, RLT or H0sum. Then this experiment is run a sufficient number of times to obtain the mean value of the picked characteristic. In our experiments we have found that for common datasets the number of times from 10 to 100 is generally sufficient. For comparison of scores for different datasets, the characteristics measured in units of length are normalized by the mean distance to the closest neighbor in a sample of 1000 points from $\mathcal{P}_{\text{data}}$. Our method is summarized in the Algorithms 1 and 2.

[1] Persistent Homology.Computation (wiki)

Algorithm 1 Cross-Barcode$_i(P, Q)$

Input: $m[P, P], m[P, Q]$: matrices of pairwise distances within point cloud P, and between point clouds P and Q
Require: VR(M): function computing the filtered complex from pairwise distances matrix M
Require: B(c, i): function computing persistence intervals of filtered complex c in dimension i
$b_Q \leftarrow$ number of columns in matrix $m[P, Q]$
$m[Q, Q] \leftarrow$ zeroes(b_Q, b_Q)
$M \leftarrow \begin{pmatrix} m[P,P] & m[P,Q] \\ m[P,Q] & m[Q,Q] \end{pmatrix}$
Cross-Barcode$_i \leftarrow$ B(VR$(M), i)$
Return: list of intervals **Cross-Barcode**$_i(P, Q)$ representing "births" and "deaths" of topological discrepancies

Algorithm 2 MTop-Divergency$(\mathcal{P}, \mathcal{Q})$, see section 2.7 for details, default suggested values: $b_{\mathcal{P}} = 1000$, $b_{\mathcal{Q}} = 10000$, $n = 100$

Input: $X_{\mathcal{P}}$, $X_{\mathcal{Q}}$: $N_{\mathcal{P}} \times D$, $N_{\mathcal{P}} \times D$ arrays representing datasets
for $j = 1$ **to** n **do**
  $P_j \leftarrow$ random choice($X_{\mathcal{P}}$,$b_{\mathcal{P}}$)
  $Q_j \leftarrow$ random choice($X_{\mathcal{Q}}$,$b_{\mathcal{Q}}$)
  $\mathcal{B}_j \leftarrow$ list of intervals Cross-Barcode$_1(P_j, Q_j)$ calculated by Algorithm 1
  $mtd_j \leftarrow$ sum of lengths of all intervals in $\mathcal{B}_j$
end for
MTop-Divergency$(\mathcal{P}, \mathcal{Q}) \leftarrow$ mean(mtd)
$r \leftarrow$ mean distance to the closest neighbor in a sample of 1000 points from $\mathcal{P}_{\text{data}}$
Normalized MTop-Divergency$(\mathcal{P}, \mathcal{Q}) \leftarrow$ mean$(mtd)/r$
Return: numbers **MTop-Divergency**$(\mathcal{P}, \mathcal{Q})$, and **Normalized MTop-Divergency**$(\mathcal{P}, \mathcal{Q})$ representing disrepancy between the distributions $\mathcal{P}, \mathcal{Q}$

Complexity. The Algorithm 1 requires computation of the two matrices of pairwise distances $m[P, P]$, $m[P, Q]$ for a pair of samples $P \in \mathbb{R}^{b_{\mathcal{P}} \times D}$, $Q \in \mathbb{R}^{b_{\mathcal{Q}} \times D}$ involving $O(b_{\mathcal{P}}^2 D)$ and $O(b_{\mathcal{P}} b_{\mathcal{Q}} D)$ operations. After that, the complexity of the computation of barcode does not depend on the dimension D of the data. Generally the persistence algorithm is at worst cubic in the number of simplices involved. In practice, the boundary matrix is sparse in our case and thanks also to the GPU optimization, the computation of cross-barcode takes similar time as in the previous step on datasets of big dimensionality. Since only the discrepancies in manifold topology are calculated, the results are quite

robust and a relatively low number of iterations is needed to obtain accurate results. Since the algorithm scales linearly with D it can be applied to the most recent datasets with D up to 10^7. For example, for $D = 3,15 \times 10^6$, and batch sizes $b_{\mathcal{P}} = 10^3, b_{\mathcal{Q}} = 10^4$, on NVIDIA TITAN RTX and 40Intel(R) Xeon(R) CPU 2.20GHz, the time for GPU accelerated calculation of pairwise distances was 15 seconds, and GPU-accelerated calculation of Cross-Barcode$_1(P,Q)$ took 30 seconds.

3. Conclusions

We have described a new framework for the evaluation of quality of GANs. We have introduced a novel tool, Cross-Barcode$_i(P,Q)$, which tracks multiscale topology discrepancies between manifolds on which the distributions are concentrated. Various numerical characteristics of Cross-Barcode(P,Q) provide qualitative and quantitative measures for the evaluation of generative models. Based on Cross-Barcode, we introduce the Manifold Topology Divergence score (MTop-Divergence)

References

Barannikov, S. Framed Morse complexes and its invariants. *Adv. Soviet Math.*, 22:93–115, 1994.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks, 2014.

Hatcher, A. *Algebraic topology*. 2005.

Khrulkov, V. and Oseledets, I. Geometry score: A method for comparing generative adversarial networks. *arXiv preprint arXiv:1802.02664*, 2018.

Moraleda, R. R., Valous, N. A., Xiong, W., and Halama, N. *Computational Topology for Biomedical Image and Data Analysis: Theory and Applications*. CRC Press, 2019.

Whitehead, G. W. *Elements of homotopy theory*, volume 61. Springer Science & Business Media, 1968.

Zhang, S., Xiao, M., and Wang, H. Gpu-accelerated computation of vietoris-rips persistence barcodes. In *36th International Symposium on Computational Geometry (SoCG 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.