Hypothesis classes with a unique persistence diagram are NOT nonuniformly learnable

Nicholas Bishop*
University of Southampton
nb8g13@soton.ac.uk

Thomas Davies*
University of Southampton
t.o.m.davies@soton.ac.uk

Long Tran-Thanh
University of Warwick
long.tran-thanh@warwick.ac.uk

Abstract

Persistence-based summaries are increasingly integrated into deep learning through topological loss functions or regularisers. The implicit role of a topological term in a loss function is to restrict the class of functions in which we are learning (the hypothesis class) to those with a specific topology. Although doing so has had empirical success, to the best of our knowledge there exists no result in the literature that theoretically justifies this restriction. Given a binary classifier in the plane with a Morse-like decision boundary, we prove that the hypothesis class defined by restricting the topology of the possible decision boundaries to those with a unique persistence diagram results in a nonuniformly learnable class of functions. In doing so, we provide a statistical learning theoretic justification for the use of persistence-based summaries in loss functions.

Important: Since presenting this work at the TDA & Beyond workshop, we have found that the result is incorrect. We have left Sections 1-3 of this paper as they were at the workshop but have added Section 4, which explains why the result is incorrect and describes our (failed) attempts to find an additional condition on the hypothesis class that makes it nonuniformly learnable.

1 Introduction

Persistence diagrams concisely summarise the topology of an underlying dataset whilst offering strong theoretical guarantees. Diagrams and their embeddings, collectively referred to as persistence-based summaries, are increasingly being integrated into deep learning (Gabrielsson et al., 2020). Many techniques add persistence-based terms to loss functions, either seeking to integrate knowledge of topological priors or to regularise the learner by encouraging topological simplicity. In both cases, the implicit role of this term is to topologically restrict the class of functions from which the learnt function can be selected. Doing so has had empirical success, but to the best of our knowledge there are no results in the literature that give a theoretical justification for imposing a topological restriction on the hypothesis class in this manner. In the context of statistical learning theory, if a hypothesis class of functions is *learnable* then there are bounds on the number of sampled points required to probably approximately learn the best classifier. We show that, for a certain class of functions, restricting the hypothesis class using prior knowledge in the form of a persistence diagram D results in a nonuniformly learnable class of functions.

*NB and TD contributed equally to this work and should be considered joint first authors.

Topological Data Analysis and Beyond Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Specifically, given prior knowledge about the topology in the form of a persistence diagram D, topological loss terms implicitly restrict the decision boundaries to the class of functions

$$\mathcal{PH}^{-1}(D) = \{f : \mathcal{PH}(f) = D\},$$

where $\mathcal{PH}$ is the persistence map. This class of functions has been studied by Curry (2018). Given a real-valued function $f : [0,1] \to \mathbb{R}$, we identify its graph as the decision boundary of a binary classifier $h_f : [0,1] \times \mathbb{R} \to \{0,1\}$ by

$$h_f((x,y)) = \begin{cases} 0, & f(x) \leq y, \\ 1, & f(x) > y. \end{cases}$$

Thus we are studying the hypothesis class

$$\mathcal{H}_D = \{h_f : f \in \mathcal{PH}^{-1}(D)\}.$$

2 Background

2.1 Persistence-based loss functions

In the context of machine learning, the *persistence map*, given by

$$\mathcal{PH}_k : X \mapsto D,$$

takes a set of points $X \subset \mathbb{R}^d$ and maps them to a *persistence diagram* D: a multiset in the extended plane that concisely represents the k-persistent homology (roughly, the topology of the points at all scales). See Edelsbrunner and Harer (2010) for an introduction to computational topology. The persistence diagram has strong theoretical guarantees that make it appealing for use in data analysis: it is stable with respect to perturbations of the underlying points (Cohen-Steiner et al., 2005) and if the data X is sampled from some distribution μ then there are guarantees that the persistence diagram of X is close to the persistence diagram of the support of μ (Fasy et al., 2014).

When endowed with the Wasserstein metric the space of persistence diagrams is complete and separable (Mileyko et al., 2011), but does not admit an isometry into a Hilbert space (Bubenik and Wagner, 2020). As most machine learning workflows require a Hilbert structure, research has been done into embedding persistence diagrams, either into finite vectors (Kališnik, 2018; Fabio and Ferri, 2015; Chepushtanova et al., 2015) or functional summaries (Bubenik, 2015; Rieck et al., 2019). Persistence diagrams and their embeddings, collectively referred to as *persistence-based summaries*, have been integrated into deep learning via topological loss or regularisation terms. Chen et al. (2018) first use the sum of squares of robustness, a concept linked to persistence, to regularise learning algorithms. Gabrielsson et al. (2020) use functions defined on persistence diagrams to enforce a topology when learning, either by promoting topological simplicity to regularise or through integrating known topological priors. Hofer et al. (2019) and Moor et al. (2019) use persistence to integrate topological information into autoencoders. Clough et al. (2020) and Hu et al. (2019) use topological losses to improve image segmentation. All of these papers empirically demonstrate success, but none offer a theoretical justification for the inclusion of topological information in the loss function.

2.2 The fibre of the persistence map

We use work by Curry (2018) on the fibre of the persistence map to understand the hypothesis class $\mathcal{H}_D$. We say two continuous functions $f, g : [0,1] \to \mathbb{R}$ are *graph-equivalent* if there is an orientation preserving homeomorphism $\phi : [0,1] \to [0,1]$ such that $f = g \circ \phi$. We say a continuous function $f : [0,1] \to \mathbb{R}$ is *Morse-like* if it is graph equivalent to a piecewise linear function where every critical point is isolated and has a distinct critical value. Let $\mathcal{PH}_0 : \mathcal{M} \to \mathcal{D}$, where $\mathcal{M}$ is the set of Morse-like real-valued functions on the interval with local minima at $x = 0$ and $x = 1$, and $\mathcal{D}$ is the space of persistence diagrams. Then Curry (2018) tells us that for any persistence diagram D,

$$\mathcal{PH}_0^{-1}(D) = \bigcup_{i \in \mathcal{I}} \mathcal{H}_{f_i},$$

where $\mathcal{I}$ is a finite indexing set over some collection of functions $\{f_i\}_i$ and $\mathcal{H}_{f_i} = \{g \in \mathcal{M} : g \sim f_i\}$ where $\sim$ denotes graph equivalence.

2.3 Statistical learning theory

Consider the standard supervised machine learning setting, in which a learner, when given an input example $x \in \mathcal{X}$, from an input space $\mathcal{X}$, would like to accurately predict the corresponding target label $y \in \mathcal{Y}$, from an output space $\mathcal{Y}$. The learner has access to a finite sample, $S = \{(x_i, y_i)\}_{i=1}^m$, of training examples, sampled from a distribution, μ over $\mathcal{X} \times \mathcal{Y}$, of interest.

The goal of the learner is to select a hypothesis, $h : \mathcal{X} \to \mathcal{Y}$, belonging to the hypothesis set, $\mathcal{H}$, which achieves low expected error with respect to the distribution μ. The error of a given hypothesis rule h on a given example (x, y) is given via a loss function, $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, by evaluating $\ell(h(x), y)$. Common examples of loss functions include the square loss, used for least squares regression, and the hinge loss, used as surrogate loss in support vector machines. More formally, the goal of the learner is to select a hypothesis which attains low risk:

$$L_\mu(h) = \mathbb{E}_{(x,y)\sim\mu}\left[\ell(h(x), y)\right]$$

A learning algorithm is a, potentially randomised, mapping from samples to the hypothesis set $\mathcal{H}$. Of course, the learner would like to employ a learning algorithm which selects a hypothesis that attains low risk with high probability. These motivations are captured by the agnostic probably-approximately-correct (PAC) framework, first introduced by Valiant (1984):

Definition 2.1 (Agnostic PAC learnability (Valiant, 1984))**.** *A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$ and for every distribution μ over $\mathcal{X} \times \mathcal{Y}$, when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by μ, the algorithm returns a hypothesis h such that, with probability of at least $1 - \delta$ (over the choice of the m training examples),*

$$L_\mu(h) \leq \min_{h' \in \mathcal{H}} L_\mu(h') + \epsilon$$

We call any such algorithm an agnostic PAC learning algorithm for the set $\mathcal{H}$.

In other words, a hypothesis set is agnostic PAC learnable if there exists a learning algorithm, such that for any (ϵ, δ), there is an $m \in \mathbb{N}$, such that when the algorithm receives at least m training examples, it returns a hypothesis rule that attains risk ϵ-close to the that of the best hypothesis in the set with probability $1 - \delta$. Of course, any learner would like to employ an agnostic PAC learning algorithm when one is available, as they give theoretical guarantees on the risk achieved as the sample size increases. Typically, m is referred to as the *sample complexity* necessary for a given algorithm to guarantee an ϵ-efficient hypothesis with probability $1 - \delta$.

In the context of binary classification, we have $\mathcal{Y} = \{0, 1\}$. In such a setting, a useful property to consider when analysing the PAC learnability of a hypothesis set $\mathcal{H}$ is the Vapnik-Chervonenkis dimension (VC-dimension) of $\mathcal{H}$.

Definition 2.2 (Shattering)**.** *Let $C = \{c_1, \ldots, c_m\} \subset \mathcal{X}$. Moreover, $\mathcal{H}_C$ be the restriction of $\mathcal{H}$ to C:*

$$\mathcal{H}_C = \{(h(c_1), \ldots, h(c_m)) \ : \ h \in \mathcal{H}\}$$

Then we say that $\mathcal{H}$ shatters C if $|\mathcal{H}_C| = 2^{|C|}$.

Definition 2.3 (VC-dimension (Vapnik, 2000))**.** *The VC-dimension of a hypothesis class $\mathcal{H}$, denoted by VCdim($\mathcal{H}$) is the maximal size of a set $C \subset \mathcal{X}$, which can be shattered by $\mathcal{H}$. If $\mathcal{H}$ shatters sets of arbitrary size, we say that $\mathcal{H}$ has infinite VC-dimension.*

In fact, it is well-known that VC-dimension characterises the PAC learnability of a hypothesis class in the case of binary classification. That is, a hypothesis set $\mathcal{H}$ is PAC learnable if and only if it has finite VC-dimension (Vapnik, 2000).

In many cases we desire learnability guarantees for complex hypothesis sets which do not have finite VC-dimension. In such cases, we can consider a relaxation of the PAC framework, in which we can derive weaker, but still useful, guarantees for generalisation:

Definition 2.4 (Nonuniform learnability (Benedek and Itai, 1988))**.** *A hypothesis class $\mathcal{H}$ is nonuniformly learnable if there exists a learning algorithm, A, and a function $m_{\mathcal{H}}^{NUL} : (0,1)^2 \times \mathcal{H} \to \mathbb{N}$ such that, for every $\epsilon, \delta \in (0,1)$ and for every $h \in \mathcal{H}$, if $m \geq m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h)$ then for every distribution μ, with probability at least $1 - \delta$ over the choice of $S \sim \mu^m$, it holds that:*

$$L_\mu(A(S)) \leq L_\mu(h) + \epsilon$$

Similarly to the framework of PAC learning, nonuniform learnability of hypothesis classes for binary classification can be characterised through VC-dimension:

Theorem 1 (Necessary and sufficient conditions for nonuniform learnability (Benedek and Itai, 1988))**.** *A hypothesis class $\mathcal{H}$ of binary classifiers is nonuniformly learnable if and only if it is a countable union of agnostic PAC learnable hypothesis classes.*

In what follows, we will illustrate how persistence diagrams can help in constructing hypothesis classes which are nonuniformly learnable.

3 The class of functions with a unique persistence diagram is nonuniformly learnable

Following previous work on the fibre of the persistence map, we concern ourselves with Morse-like real-valued functions on the unit interval with local minima at $x = 0, 1$, and the 0th persistence diagram. Although this is clearly a restricted setting, it provides an initial justification for the use of topology in learning. Curry (2018) tells us that $\mathcal{PH}_0^{-1}(D)$ is finite when we introduce graph equivalence on the set of functions. Moreover, Theorem 1 tells us that a hypothesis class of binary classifiers is nonuniformly learnable if and only if it is a countable union of agnostic PAC learnable hypothesis classes. Therefore, to show that restricting a hypothesis class of functions to those with a unique persistence diagram is nonuniformly learnable, we are required to show that the class of graph equivalent functions is agnostic PAC learnable, or equivalently, that it has finite VC-dimension.

Theorem 2. *Let $f : [0, 1] \to \mathbb{R}$ be a Morse-like real-valued function on the unit interval with local minima at $x = 0, 1$, and $\mathcal{H}_f$ be the class of functions that are graph equivalent to f. Then* $\mathrm{VCdim}(\mathcal{H}_f) = 0$.

Proof. Since f is a continuous real-valued function on a closed interval of $\mathbb{R}$, then, by the boundedness theorem, there is an (attained) upper bound M of f. Let $g : [0, 1] \to \mathbb{R}$ be graph equivalent to f, i.e., there exists an orientation preserving homeomorphism $\phi : [0, 1] \to [0, 1]$ such that $g = f \circ \phi$. Then $g(x) = f(\phi(x)) \leq M$ for all $x \in [0, 1]$. Thus, given our classifier, h_g, we have that $h_g((x, y)) \neq 0$ whenever $y > M$. Since g is an arbitrary function that is graph equivalent to f, $\mathcal{H}_f$ cannot shatter every set of size 1, so has VC-dimension 0. □

Therefore we have the following result:

Theorem 3. *Given a persistence diagram D, the hypothesis class of binary classifiers with decision boundaries defined by Morse-like real-valued functions on the interval with local minima at $x = 0, 1$ given by*

$$\mathcal{H}_D = \{h_f : f \in \mathcal{PH}_0^{-1}(D)\}.$$

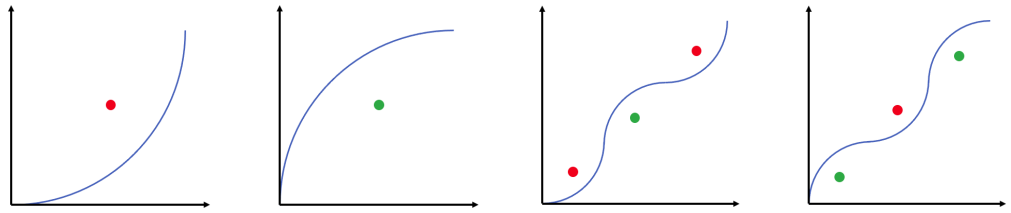
is nonuniformly learnable.

Proof. Theorem 6.12 by Curry (2018) shows that $\mathcal{PH}_0^{-1}(D) = \bigcup_{i \in \mathcal{I}} \mathcal{H}_{f_i}$ for some finite indexing set $\mathcal{I}$. Meanwhile, Theorem 2 shows that each $\mathcal{H}_{f_i}$ has finite VC-dimension, and is therefore agnostic PAC learnable. In addition, Theorem 1 shows that the countable union of agnostic PAC learnable hypothesis classes is nonuniformly learnable. Therefore, $\mathcal{H}_D$ is nonuniformly learnable. □

4 A counterexample to nonuniform learnability

Our error came from a misunderstanding of the VC dimension (Definition 2.3). In order to show that a hypothesis class $\mathcal{H}$ has VC dimension k, we believed we had to show that $\mathcal{H}$ cannot shatter every set of k points (as we showed in Theorem 2). This led us to believe that $\mathrm{VCdim}(\mathcal{H}_f) = 0$. However, you in fact need to show that (i) there exists a set of size k that can be shattered by $\mathcal{H}$ and (ii) *every* set of size $k + 1$ cannot be shattered by $\mathcal{H}$ (Shalev-Shwartz and Ben-David, 2014). Therefore to prove that $\mathrm{VCdim}(\mathcal{H}_f) > 0$ we just need to find one point that can be shattered by $\mathcal{H}_f$. Recall that two functions are graph equivalent if there is an orientation preserving homeomorphism between them, and a continuous function f is Morse-like if it is graph equivalent to a piecewise linear function where every critical point is isolated and has a distinct critical value. By applying orientation-preserving transformations, we can 'stretch and squeeze' the x axis, but we can never

change the number of critical points or their values. Imagine the most simple case: a point to be classified lying between two critical points that are connected with a line. We can easily find orientation-preserving homeomorphisms that allow us to classify that point as either 0 or 1, as shown in Figure 1(a). Therefore we have shattered that point. In fact, if there are arbitrarily many points lying on a line between two critical points, then we can shatter them, as demonstrated in Figure 1(b). We formalise this below.



(a) Shattering one point

(b) Shattering several points

Figure 1: We can shatter arbitrarily many points between two critical points using graph equivalent functions, proving that the VC dimension of $\mathcal{H}_f$ is unbounded.

Proposition 3.1. *Let $f : [0, 1] \to \mathbb{R}$ be a Morse-like real-valued function on the unit interval with local minima at $x = 0, 1$, and $\mathcal{H}_f$ be the class of functions that are graph equivalent to f. Then $\mathrm{VCdim}(\mathcal{H}_f) = \infty$ and $\mathcal{H}_D$ is not (nonuniformly) learnable.*

Proof. Let $\mathcal{H}_f$ be the class of classifiers with decision boundaries graph equivalent to some piecewise linear function f. Let n points lie equally spaced on the straight line between two adjacent critical points of f. By introducing inflection points using orientation-preserving homeomorphisms we can shatter the points, as shown in Figure 1(b). Note that we only introduce inflection points, not stationary points, as the number and values of critical points must remain constant in $\mathcal{H}_f$. Since we can shatter an arbitrary number of points, $\mathrm{VCdim}(\mathcal{H}_f) = \infty$. The non-learnability of $\mathcal{H}_D$ follows immediately. □

Note that since we only need one counterexample, this is sufficient, but by the pigeonhole principle we can force an arbitrary number of points to be strictly between two adjacent critical points for any $\mathcal{H}_f$ (where they are strictly between the two critical points because any point above or below a critical point cannot be shattered).

Our counterexample to $\mathcal{H}_f$ came about when we introduced inflection points. As these allow our classifier to change from convex to concave without introducing additional critical points, this lets graph equivalent functions be extremely expressive. This gave us the idea that perhaps restricting the number of inflection points allowed within $\mathcal{H}_f$ could lead to a finite VC dimension. Such a restriction is realistic: the number of inflection points is finite in many common learning algorithms, including neural networks. However, there remains sets of points that can be shattered, as we show in the following proposition.

Proposition 3.2. *Let f and $\mathcal{H}_f$ be as in Proposition 4.5.1, and additionally suppose that the number of inflection points (i.e., points where the function changes from convex to concave) is finite. Then it remains the case that $\mathrm{VCdim}(\mathcal{H}_f) = \infty$ and $\mathcal{H}_D$ is not (nonuniformly) learnable.*

Proof. Without loss of generality, consider a section of f that is increasing and concave, i.e., it lies between two critical or inflection points $(x_0, y_0), (x_{n+1}, y_{n+1})$. Place n points $(x_1, y_1), \ldots, (x_n, y_n)$ between them so that $2x_i = x_{i-1}$ and $y_i = 2y_{i-1}$ for all $i = 1, \ldots, n+1$. Define a new list of points by

$$(\alpha_i, \beta_i) = \begin{cases} (x_i + \epsilon, y_i), & \text{if } (x_i, y_i) \text{ is labelled } 1, \\ (x_i - \epsilon, y_i), & \text{if } (x_i, y_i) \text{ is labelled } 0, \end{cases}$$

for some small $\epsilon > 0$. The decision boundary defined by connecting $\{(\alpha_i, \beta_i)\}_i$ into a piecewise linear function can attain any labelling of the n points, and so shatters them. Furthermore, it is increasing, has no critical points, and is concave (an example is shown in Figure 2). Therefore n points can be shattered by such a function for arbitrary n, and $\mathrm{VCdim}(\mathcal{H}_f) = \infty$. □

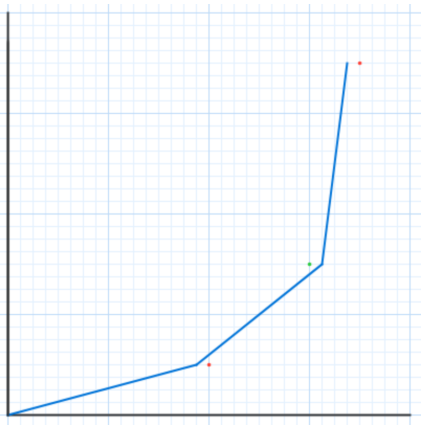


Figure 2: Bounding the number of inflection points does not prevent an arbitrary number of points from being shattered.

In fact, we could not think of any reasonable conditions on the hypothesis class that would make them learnable.

5 Conclusion

We initially believed that our work provided the first learning theoretic justification for integrating persistence-based summaries into loss functions for deep learning. Although we've now shown that topologically restricted hypothesis classes are *not* nonuniformly learnable, the existence of pathological examples that prevent learning theoretic bounds being achieved does not mean that topological loss functions have no practical value. On the contrary, we've seen that the literature demonstrates they can still provide valuable additional information (Section 2.1).

References

Gyora M. Benedek and Alon Itai. Nonuniform learnability. In Timo Lepistö and Arto Salomaa, editors, *Automata, Languages and Programming*, pages 82–92, Berlin, Heidelberg, 1988. Springer Berlin Heidelberg. ISBN 978-3-540-39291-0.

Peter Bubenik. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research*, 16(3):77–102, 2015. URL `http://jmlr.org/papers/v16/bubenik15a.html`.

Peter Bubenik and A. Wagner. Embeddings of persistence diagrams into hilbert spaces. *J. Appl. Comput. Topol.*, 4:339–351, 2020.

Chao Chen, Xiuyan Ni, Qinxun Bai, and Yusu Wang. Toporeg: A topological regularizer for classifiers. *CoRR*, abs/1806.10714, 2018. URL `http://arxiv.org/abs/1806.10714`.

Sofya Chepushtanova, Tegan Emerson, Eric Hanson, Michael Kirby, Francis Motta, Rachel Neville, Chris Peterson, Patrick Shipman, and Lori Ziegelmeier. Persistence images: An alternative persistent homology representation. 07 2015.

J. Clough, N. Byrne, I. Oksuz, V. A. Zimmer, J. A. Schnabel, and A. King. A topological loss function for deep-learning based image segmentation using persistent homology. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. volume 37, pages 263–271, 01 2005. doi: 10.1007/s00454-006-1276-5.

Justin Curry. The fiber of the persistence map for functions on the interval. *Journal of Applied and Computational Topology*, 2:301–321, 2018.

Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. 01 2010. ISBN 978-0-8218-4925-5. doi: 10.1007/978-3-540-33259-6_7.

Barbara Di Fabio and Massimo Ferri. Comparing persistence diagrams through complex vectors. In *Image Analysis and Processing — ICIAP 2015*, pages 294–305. Springer International Publishing, 2015. doi: 10.1007/978-3-319-23231-7_27. URL https://doi.org/10.1007/978-3-319-23231-7_27.

Brittany Terese Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, and A. Singh. Confidence sets for persistence diagrams. *Annals of Statistics*, 42:2301–2339, 2014.

Rickard Brüel Gabrielsson, Bradley J. Nelson, Anjan Dwaraknath, and Primoz Skraba. A topology layer for machine learning. volume 108 of *Proceedings of Machine Learning Research*, pages 1553–1563, Online, 26–28 Aug 2020. PMLR. URL http://proceedings.mlr.press/v108/gabrielsson20a.html.

Christoph Hofer, Roland Kwitt, Marc Niethammer, and Mandar Dixit. Connectivity-optimized representation learning via persistent homology. volume 97 of *Proceedings of Machine Learning Research*, pages 2751–2760, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/hofer19a.html.

Xiaoling Hu, Fuxin Li, Dimitris Samaras, and Chao Chen. Topology-preserving deep image segmentation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5657–5668. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8803-topology-preserving-deep-image-segmentation.pdf.

Sara Kališnik. Tropical coordinates on the space of persistence barcodes. *Foundations of Computational Mathematics*, 19(1):101–129, January 2018. doi: 10.1007/s10208-018-9379-y. URL https://doi.org/10.1007/s10208-018-9379-y.

Yuriy Mileyko, Sayan Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. *Inverse Problems - INVERSE PROBL*, 27, 12 2011. doi: 10.1088/0266-5611/27/12/124007.

M. Moor, Max Horn, Bastian Alexander Rieck, and K. Borgwardt. Topological autoencoders. *ArXiv*, abs/1906.00722, 2019.

Bastian Alexander Rieck, F. Sadlo, and H. Leitte. Topological machine learning with persistence indicator functions. *ArXiv*, abs/1907.13496, 2019.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014. ISBN 1107057132.

L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, November 1984. ISSN 0001-0782. doi: 10.1145/1968.1972. URL https://doi.org/10.1145/1968.1972.

Vladimir N. Vapnik. *Bounds on the Rate of Convergence of Learning Processes*, pages 69–91. Springer New York, New York, NY, 2000. ISBN 978-1-4757-3264-1. doi: 10.1007/978-1-4757-3264-1_4. URL https://doi.org/10.1007/978-1-4757-3264-1_4.