Submitted 10/16; Revised 12/17; Published 6/18

Active Nearest-Neighbor Learning in Metric Spaces

Aryeh Kontorovich KARYEH@CS.BGU.AC.IL
Sivan Sabato SABATOS@CS.BGU.AC.IL
Department of Computer Science
Ben-Gurion University of the Negev
Beer Sheva 8499000, Israel

Ruth Urner RUTH@EECS.YORKU.CA
Lassonde School of Engineeging, EECS Department
York University
Toronto, ON, Canada

Editor: Andreas Krause

Abstract

We propose a pool-based non-parametric active learning algorithm for general metric spaces, called MArgin Regularized Metric Active Nearest Neighbor (MARMANN), which outputs a nearest-neighbor classifier. We give prediction error guarantees that depend on the noisy-margin properties of the input sample, and are competitive with those obtained by previously proposed passive learners. We prove that the label complexity of MARMANN is significantly lower than that of any passive learner with similar error guarantees. MARMANN is based on a generalized sample compression scheme, and a new label-efficient active model-selection procedure.

Keywords: Nearest-neighbors, active learning, metric spaces, non-parametric learning

1. Introduction

Active learning is a framework for reducing the amount of label supervision for prediction tasks. While labeling large amounts of data can be expensive and time-consuming, unlabeled data is often much easier to come by. In this paper we propose a non-parametric pool-based active learning algorithm for general metric spaces, which outputs a nearest-neighbor classifier.

In pool-based active learning (McCallum and Nigam, 1998), a collection of random examples is provided, and the algorithm can interactively query an oracle to label some of the examples. The goal is good prediction accuracy, while keeping the label complexity — that is, the number of labels queried — low. Our algorithm, MArgin Regularized Metric Active Nearest Neighbor (MARMANN), receives a pool of unlabeled examples in a general metric space, and outputs a variant of the 1-nearest-neighbor classifier, implemented by a 1-nearest-neighbor rule. The algorithm obtains a prediction error guarantee that depends on a noisy-margin property of the input sample, and has a provably smaller label complexity than any passive learner with a similar guarantee.

Active learning has been mostly studied in a parametric setting, in which learning takes place with respect to a fixed hypothesis class with a bounded capacity. There has also been some work on analyzing non-parametric active learning strategies under certain distributional assumptions (see Section 1.1 for more discussion on this). However, the question of whether active querying strategies can yield label savings for non-parametric methods in a general setting, without distributional assumptions, had not been analyzed prior to this work. Here, we provide a first demonstration that this is indeed possible. We discuss related work in detail in Section 1.1 below.

Our contributions. MARMANN is a new non-parametric pool-based active learning algorithm, which obtains an error guarantee competitive with that of a noisy-margin-based passive learner. Additionally, it provably uses significantly fewer labels in nontrivial regimes. As far as the authors are aware, this is the first non-parametric active learner for general metric spaces, which achieves competitive prediction error guarantees to the passive learner, while provably improving label complexity. The guarantees of MARMANN are given in Theorem 4 in Section 3. We further provide a passive learning lower bound (Theorem 5), which together with Theorem 4 shows that MARMANN can have a significantly reduced label complexity compared to any passive learner. The passive lower bound is more general than previous lower bounds, relies on a novel technique, and may be of independent interest. Additionally, we give an active label complexity lower bound (Theorem 6), which holds for any active learner with similar error guarantees as MARMANN. The proof of this active lower bound relies on a new No-Free-Lunch type result, which holds for active learning algorithms.

Our approach. Previous passive learning approaches to classification using nearest-neighbor rules under noisy-margin assumptions (Gottlieb et al., 2014b, 2017) provide statistical guarantees using sample compression bounds (Graepel et al., 2005). Their finite-sample guarantees depend on the number of noisy labels relative to an optimal margin scale.

A central challenge in the active setting is performing model selection to select a margin scale with a low label complexity. A key insight that we exploit in this work is that by designing a new labeling scheme for the compression set, we can construct the compression set and estimate its error with label-efficient procedures. We obtain statistical guarantees for this approach using generalization bounds for sample compression with side information.

We derive a label-efficient, as well as computationally efficient, active model-selection procedure. This procedure finds a good scale by estimating the sample error for some scales, using a small number of active querying rounds. Crucially, unlike cross-validation, our model-selection procedure does not require a number of labels that depends on the worst possible scale, nor does it test many scales. This allows our label complexity bounds to be low, and to depend only on the final scale selected by the algorithm. Our error guarantee is a constant factor over the error guarantee of the passive learner of Gottlieb et al. (2017). An approach similar to Gottlieb et al. (2017), proposed in Gottlieb et al. (2014a), has been shown to be Bayes consistent (Kontorovich and Weiss, 2015). The Bayes-consistency of the passive version of our approach has recently been established (Kontorovich et al., 2017).

Paper structure. Related work is discussed in Section 1.1. We lay down the preliminaries in Section 2. In Section 3 we provide our main result: Theorem 4, which gives error and label complexity guarantees for MARMANN. Additionally we state the passive and active lower bounds, Theorem 5 and Theorem 6. The rest of the paper is devoted to the description and analysis of MARMANN, and proof of the main results. Section 4 shows how MARMANN defines the nearest neighbor rule for a given scale, and Section 5 describes the model selection procedure of MARMANN. Theorem 4 is proved in Section 6, based on a framework for compression with side information. The passive lower bound in Theorem 5 is proved in Section 7. The active lower bound Theorem 6 is proved in Section 8. We conclude with a discussion in Section 9.

1.1 Related Work

The theory of active learning has received considerable attention in the past decade (e.g., Dasgupta, 2004; Balcan et al., 2007, 2009; Hanneke, 2011; Hanneke and Yang, 2015). Active learning theory has been mostly studied in a parametric setting (that is, learning with respect to a fixed hypothesis class with a bounded capacity). Benefits and limitations of various active querying strategies have been proven in the realizable setting (Dasgupta, 2004; Balcan et al., 2007; Gonen et al., 2013) as well as in the agnostic case (Balcan et al., 2009; Hanneke, 2011; Awasthi et al., 2014). It has also been shown that active queries can also be beneficial for regression tasks (Castro et al., 2005; Sabato and Munos, 2014). Further, an active model selection procedure has been developed for the parametric setting (Balcan et al., 2010).

The potential benefits of active learning for non-parametric settings are less well understood. Practical Bayesian graph-based active learning methods (Zhu et al., 2003; Wei et al., 2015) rely on generative model assumptions, and therefore come without distribution-free performance guarantees. From a theoretical perspective, the label complexity of graph based active learning has mostly been analyzed in terms of combinatorial graph parameters (Cesa-Bianchi et al., 2010; Dasarathy et al., 2015). While the latter work provides some statistical guarantees under specific conditions on the data-generating distribution, this type of analysis does not yield distribution-free statistical performance guarantees.

Castro et al. (2005); Castro and Nowak (2008) analyze minimax rates for non-parametric regression and classification respectively, for a class of distributions in Euclidean space, characterized by decision boundary regularity and noise conditions with uniform marginals. The paradigm of cluster-based active learning (Dasgupta and Hsu, 2008) has been shown to provide label savings under some distributional clusterability assumptions (Urner et al., 2013; Kpotufe et al., 2015). Dasgupta and Hsu (2008) showed that a suitable cluster-tree can yield label savings in this framework, and papers following up (Urner et al., 2013; Kpotufe et al., 2015) quantified the label savings under distributional clusterability assumptions. However, no active non-parametric strategy has been proposed so far that has label complexity guarantees for i.i.d. data from general distributions and general metric spaces. Here, we provide the first such algorithm and guarantees.

The passive nearest-neighbor classifier, introduced by Fix and Hodges (1951, 1989), is popular among theorists and practitioners alike (Fix and Hodges, 1989; Cover and Hart, 1967; Stone, 1977; Kulkarni and Posner, 1995; Boiman et al., 2008). This paradigm is applicable in general metric spaces, and its simplicity is an attractive feature for both implementation and analysis. When appropriately regularized — either by taking a majority vote among the k nearest neighbors (Stone, 1977; Devroye and Györfi, 1985; Zhao, 1987), or by enforcing a *margin* separating the classes (von Luxburg and Bousquet, 2004; Gottlieb et al., 2014a; Kontorovich and Weiss, 2015; Kontorovich et al., 2017) — this type of learner can be made Bayes-consistent. Another desirable property of nearest-neighbor-based methods is their ability to generalize at a rate that scales with the intrinsic data dimension, which can be much lower than that of the ambient space (Kpotufe, 2011; Gottlieb et al., 2014a, 2016; Chaudhuri and Dasgupta, 2014). Furthermore, margin-based regularization makes nearest neighbor classifiers ideally suited for sample compression, which yields a compact representation, faster classification runtime, and improved generalization performance (Gottlieb et al., 2014b). The resulting error guarantees can be stated in terms of the sample's noisy-margin, which depends on the distances between differently-labeled examples in the input sample.

Active learning strategies specific to nearest neighbor classification have recently received attention. It has been shown that certain active querying rules maintain Bayes consistency for nearest neighbor classification, while other, seemingly natural, rules do not lead to a consistent algorithm (Dasgupta, 2012). A selective querying strategy has been shown to be beneficial for nearest neighbors under covariate shift (Berlind and Urner, 2015), where one needs to adapt to a change in the data generating process. However, the querying rule in that work is based solely on information in the unlabeled data, to account for a shift in the distribution over the covariates. It does not imply any label savings in the standard learning setting, where training and test distribution are identical. In contrast, our current work demonstrates how an active learner can take label information into account, to reduce the label complexity of a general nearest neighbor method in the standard setting.

1.2 A Remark on Bayes-Consistency

We remark on the Bayes-consistency of the margin-based passive 1-NN methods. In Gottlieb et al. (2014a), a PAC-style generalization bound was given. At a given scale t, the algorithm first ensured t-separation of the sample by solving solving a minimum vertex cover problem to eliminate the t-blocking pairs. Following that, the hypothesis was constructed as a Lipschitz extension from the remaining sample; the latter is computationally implemented as a nearest neighbor classifier. Structural Risk Minimization (SRM) was used to select the optimal scale t. A very close variant of this learner was shown to be Bayes-consistent by Kontorovich and Weiss (2015). The only difference between the two is that the former analyzed the hypothesis complexity in terms of fat-shattering dimension while the latter via Rademacher averages. Thus, a margin-regularized 1-NN classifier was shown to be Bayes-consistent; however, no compression was involved.

A compression-based alternative to Lipschitz extension was proposed in Gottlieb et al. (2014b). The idea is again to ensure t-separation via vertex cover and then compress the remaining sample down to a t-net. We conjecture that this latter algorithm is also Bayes-consistent, but currently have no proof. If instead one considers a compression-based passive learner implemented as in this paper (by taking majority vote in each Voronoi region rather than enforcing t-separation via vertex cover), the resulting classifier is indeed Bayes-consistent, as was recently shown by Kontorovich et al. (2017).

2. Preliminaries

In this section we lay down the necessary preliminaries. We formally define the setting and necessary notation in Section 2.1. We discuss nets in metric spaces in Section 2.2, and present the guarantees of the compression-based passive learner of Gottlieb et al. (2017) in Section 2.3.

2.1 Setting and Notation

For positive integers n, denote $[n] := \{1, \dots, n\}$. We consider learning in a general metric space $(\mathcal{X}, \rho)$, where $\mathcal{X}$ is a set and ρ is the metric on $\mathcal{X}$. Our problem setting is that of classification of the instance space $\mathcal{X}$ into some finite label set $\mathcal{Y}$. Assume that there is some distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, and let $S \sim \mathcal{D}^m$ be a labeled sample of size m, where m is an integer. Denote the sequence of unlabeled points in S by $\mathbb{U}(S)$. We sometimes treat S and $\mathbb{U}(S)$ as multisets, since the order is unimportant. For a labeled multiset $S \subseteq \mathcal{X} \times \mathcal{Y}$ and $y \in \mathcal{Y}$, denote $S^y := \{x \mid (x, y) \in S\}$; in particular, $\mathbb{U}(S) = \cup_{y \in \mathcal{Y}} S^y$.

The error of a classifier $h : \mathcal{X} \to \mathcal{Y}$ on $\mathcal{D}$, for any fixed h, is denoted

$$\mathrm{err}(h, \mathcal{D}) := \mathbb{P}[h(X) \neq Y],$$

where $(X, Y) \sim \mathcal{D}$. The empirical error on a labeled sample S instantiates to

$$\mathrm{err}(h, S) = \frac{1}{|S|} \sum \mathbb{I}[h(X) \neq Y].$$

A passive learner receives a labeled sample S_{in} as input. An active learner receives the unlabeled part of the sample $U_{\mathrm{in}} := \mathbb{U}(S_{\mathrm{in}})$ as input, and is allowed to interactively select examples from U_{in} and request their label from S_{in}. In other words, the active learner iteratively selects an example and requests its label, wherein all the labels requested so far can be used to make the next selection.

When either learner terminates, it outputs a classifier $\hat{h} : \mathcal{X} \to \mathcal{Y}$, with the goal of achieving a low $\mathrm{err}(\hat{h}, \mathcal{D})$. An additional goal of the active learner is to achieve a performance competitive with that of the passive learner, while querying considerably fewer labels.

The diameter of a set $A \subseteq \mathcal{X}$ is defined by

$$\mathsf{diam}(A) := \sup_{a, a' \in A} \rho(a, a').$$

For a finite set $U = \{u_1, \ldots, u_{|U|}\} \subseteq \mathcal{X}$ with some fixed numbering of its elements,[1] denote the index of the closest point in U to $x \in \mathcal{X}$ by

$$\kappa(x, U) := \operatorname*{argmin}_{i : x_i \in U} \rho(x, x_i).$$

We assume here and throughout this work that when there is more than one minimizer for $\rho(x, x_i)$, ties are broken arbitrarily (but in a consistent and deterministic fashion). Any labeled sample $S = ((x_i, y_i))_{i \in [k]}$ naturally induces the 1-nearest-neighbor classifier $h^{\mathrm{nn}}_S : \mathcal{X} \to \mathcal{Y}$, via $h^{\mathrm{nn}}_S(x) := y_{\kappa(x, \mathbb{U}(S))}$. For a set $Z \subseteq \mathcal{X}$, denote by

$$\kappa(Z, U) := \{\kappa(z, U) \mid z \in Z\}$$

the set of all the indices $\kappa(z, U)$, as defined above. For $x \in \mathcal{X}$, and $t > 0$, denote by $\mathsf{ball}(x, t)$ the (closed) ball of radius t around x:

$$\mathsf{ball}(x, t) := \left\{x' \in \mathcal{X} \mid \rho(x, x') \leq t\right\}.$$

2.2 Nets

A set $A \subseteq \mathcal{X}$ is *t-separated* if $\inf_{a, a' \in A : a \neq a'} \rho(a, a') \geq t$. For $A \subseteq B \subseteq \mathcal{X}$, the set A is a *t-net* of B if A is t-separated and $B \subseteq \bigcup_{a \in A} \mathsf{ball}(a, t)$. Thus, A is a t-net of B if it is both a t-covering and a t-packing.

[1] Invoking the well-ordering principle, we may assume $\mathcal{X}$ to be well-ordered and then any $U \subseteq \mathcal{X}$ inherits the ordering of $\mathcal{X}$.

The size of a t-net of a metric space is strongly related to its *doubling dimension*. The doubling dimension is the effective dimension of the metric space, which controls generalization and runtime performance of nearest-neighbors (Kpotufe, 2011; Gottlieb et al., 2014a). It is defined as follows. Let $\lambda = \lambda(\mathcal{X})$ be the smallest number such that every ball in $\mathcal{X}$ can be covered by λ balls of half its radius, where all balls are centered at points of $\mathcal{X}$. Formally,

$$\lambda(\mathcal{X}) := \min\{\lambda \in \mathbb{N} : \forall x \in \mathcal{X}, r > 0, \quad \exists x_1, \ldots, x_\lambda \in \mathcal{X} : \mathsf{ball}(x, r) \subseteq \cup_{i=1}^{\lambda} \mathsf{ball}(x_i, r/2)\}.$$

Then the doubling dimension of $\mathcal{X}$ is defined by $\mathsf{ddim}(\mathcal{X}) := \log_2 \lambda$. In line with modern literature, we work in the low-dimension, large-sample regime, where the doubling dimension is assumed to be constant, and hence sample complexity and algorithmic runtime may depend on it exponentially. This exponential dependence is unavoidable, even under margin assumptions, as previous analyses (Kpotufe, 2011; Gottlieb et al., 2014a) indicate. Generalization bounds in terms of the doubling dimension of the hypothesis space were established in Bshouty et al. (2009), while runtime and generalization errors in terms of $\mathsf{ddim}(\mathcal{X})$ were given in Gottlieb et al. (2014a).

As shown in Gottlieb and Krauthgamer (2013), the doubling dimension is “almost hereditary” in the sense that for $A \subset \mathcal{X}$, we have $\mathsf{ddim}(A) \le c\mathsf{ddim}(\mathcal{X})$ for some universal constant $c \le 2$ (Feldmann et al., 2015, Lemma 6.6). In the works cited above, where generalization bounds are stated in terms of $\mathsf{ddim}(\mathcal{X})$, one can obtain tighter bounds in terms of $\mathsf{ddim}(\mathbb{U}(S))$ when the latter is substantially lower than that of the ambient space, and it is also possible to perform metric dimensionality reduction, as in Gottlieb et al. (2016).

Constructing a minimum size t-net for a general set B is NP-hard (Gottlieb and Krauthgamer, 2013). However, a simple greedy algorithm constructs a (not necessarily minimal) t-net in time $O(m^2)$ (Gottlieb et al., 2014b, Algorithm 1). There is also an algorithm for constructing a t-net in time $2^{O(\mathsf{ddim}(\mathcal{X}))} m \log(1/t)$ (Krauthgamer and Lee, 2004; Gottlieb et al., 2014b). The size of any t-net of a metric space $A \subseteq \mathcal{X}$ is at most

$$\lceil \mathsf{diam}(A)/t \rceil^{\mathsf{ddim}(\mathcal{X})+1} \tag{1}$$

(Krauthgamer and Lee, 2004). In addition, the size of any t-net is at most $2^{\mathsf{ddim}(A)+1}$ times the size of the minimal t-net, as the following easy lemma shows.

Lemma 1 (comparison of two nets) *Let $t > 0$ and suppose that M_1, M_2 are t-nets of $A \subseteq \mathcal{X}$. Then $|M_1| \le 2^{\mathsf{ddim}(A)+1} |M_2|$.*

Proof Suppose that $|M_1| \ge k|M_2|$ for some positive integer k. Since $M_1 \subseteq \bigcup_{x \in M_2} \mathsf{ball}(x, t)$, it follows from the pigeonhole principle that at least one of the points in M_2 must cover at least k points in M_1. Thus, suppose that $x \in M_2$ covers the set $Z = \{z_1, \ldots, z_l\} \subseteq M_1$, meaning that $Z \subseteq \mathsf{ball}(x, t)$, where $l = |Z| \ge k$. By virtue of belonging to the t-net M_1, the set Z is t-separated. Therefore Z is a t-net of Z. Since Z is contained in a t-ball, we have $\mathsf{diam}(Z) \le 2t$. It follows from Eq. (1) that $|Z| \le 2^{\mathsf{ddim}(A)+1}$, whence the claim. ■

Throughout the paper, we fix a deterministic procedure for constructing a t-net, and denote its output for a multiset $U \subseteq \mathcal{X}$ by $\mathsf{Net}(U, t)$. Let $\mathsf{Par}(U, t)$ be a partition of $\mathcal{X}$ into regions induced by $\mathsf{Net}(U, t)$, that is: for $\mathsf{Net}(U, t) = \{x_1, \ldots, x_N\}$, define $\mathsf{Par}(U, t) := \{P_1, \ldots, P_N\}$, where

$$P_i = \{x \in \mathcal{X} \mid \kappa(x, \mathsf{Net}(U, t)) = i\}.$$

For $t > 0$, let $\mathcal{N}(t) := |\mathsf{Net}(U_{\mathrm{in}}, t)|$ be the size of the t-net for the input sample.

2.3 Passive Compression-Based Nearest-Neighbors

Non-parametric binary classification admits performance guarantees that scale with the sample's noisy-margin (von Luxburg and Bousquet, 2004; Gottlieb et al., 2014a, 2017). The original margin-based methods of von Luxburg and Bousquet (2004) and Gottlieb et al. (2014a) analyzed the generalization performance via the technique of Lipschitz extension. Later, it was noticed in Gottlieb et al. (2014b) that the presence of a margin allows for compression — in fact, nearly optimally so.

We say that a labeled multiset S is (ν, t)-*separated*, for $\nu \in [0, 1]$ and $t > 0$ (representing a margin t with noise ν), if one can remove a ν-fraction of the points in S, and in the resulting multiset, any pair of differently labeled points is separated by a distance of at least t. Formally, we have the following definition.

Definition 2 *S is (ν, t)-separated if there exists a subsample $\tilde{S} \subseteq S$ such that*

1. *$|S \setminus \tilde{S}| \leq \nu |S|$ and*
2. *$\forall y_1 \neq y_2 \in \mathcal{Y}, a \in \tilde{S}^{y_1}, b \in \tilde{S}^{y_2}$, we have $\rho(a, b) \geq t$.*

For a given labeled sample S, denote by $\nu(t)$ the smallest value ν such that S is (ν, t)-separated. Gottlieb et al. (2017) propose a passive learner with the following guarantees[2] as a function of the separation of S. Setting $\alpha := m/(m - N)$, define the following form of a generalization bound:

$$\mathrm{GB}(\epsilon, N, \delta, m, k) := \alpha\epsilon + \frac{2}{3}\frac{(N+1)\log(mk) + \log(\frac{1}{\delta})}{m - N} + \frac{3}{\sqrt{2}}\sqrt{\frac{\alpha\epsilon((N+1)\log(mk) + \log(\frac{1}{\delta}))}{m - N}}.$$

Further, for an integer m and $\delta \in (0, 1)$, denote

$$G_{\min}(m, \delta) := \min_{t>0} \mathrm{GB}(\nu(t), \mathcal{N}(t), \delta, m, 1).$$

The quantity $G_{\min}(m, \delta)$ is small for datasets where we only need to remove few points to obtain a reasonably well separated subset. As an example, consider data generated by two well separated Gaussians (one generating the positively labeled points, and one generating the negatively labeled points). Then, most of the data points will be close to their respective means, but some will be farther, and may lie closer to the mean of the other Gaussian. Removing those few will result in a separated set.

Theorem 3 (Gottlieb et al. 2017) *Let m be an integer, $\delta \in (0, 1)$. There exists a passive learning algorithm that returns a nearest-neighbor classifier $h^{\mathrm{nn}}_{S_{\mathrm{pas}}}$, where $S_{\mathrm{pas}} \subseteq S_{\mathrm{in}}$, such that, with probability* $1 - \delta$,

$$\mathrm{err}(h^{\mathrm{nn}}_{S_{\mathrm{pas}}}, \mathcal{D}) \leq G_{\min}(m, \delta).$$

The bound above is data-dependent, meaning that the strength of the generalization guarantee depends on the quality of the random sample. Specifically, the passive algorithm of Gottlieb et al. (2017) generates S_{pas} of size approximately $\mathcal{N}(t)$ for the optimal scale $t > 0$ (found by searching over all scales), by removing the $|S_{\mathrm{in}}|\nu(t)$ points that obstruct the t-separation between different labels in S_{in}, and then selecting a subset of the remaining labeled examples to form S_{pas}, so that the

[2] The guarantees hold for the more general case of semimetrics.

examples are a t-net for S_{in} (not including the obstructing points). For the binary classification case ($|\mathcal{Y}| = 2$) an efficient algorithm is shown in Gottlieb et al. (2017). However, in the general multiclass case, it is not known how to find a minimal t-separation efficiently — a naive approach requires solving the NP-hard problem of vertex cover. Our approach, which we detail below, circumvents this issue, and provides an efficient algorithm also for the multiclass case.

3. Main Results

We propose a novel approach for generating a subset for a nearest-neighbor rule. This approach, detailed in the following sections, does not require finding and removing all the obstructing points in S_{in}, and can be implemented in an active setting using a small number of labels. The resulting active learning algorithm, MARMANN, has an error guarantee competitive with that of the passive learner, and a label complexity that can be significantly lower. We term the subset used by the nearest-neighbor rule a *compression set*.

Algorithm 1 MARMANN: MArgin Regularized Metric Active Nearest Neighbor

```
input  Unlabeled sample $U_{\text{in}}$ of size $m$, $\delta \in (0, 1)$.
  $\hat{t} \leftarrow$ SelectScale($\delta$).                 # SelectScale is given in Section 5, Alg. 4.
  $\hat{S} \leftarrow$ GenerateNNSet($\hat{t}$, $[\mathcal{N}(\hat{t})], \delta$).    # GenerateNNSet is given in Section 4, Alg. 2.
  Output $h^{\text{nn}}_{\hat{S}}$.
```

MARMANN, listed in Alg. 1, operates as follows. First, a scale $\hat{t} > 0$ is selected, by calling $\hat{t} \leftarrow \text{SelectScale}(\delta)$, where SelectScale is our model selection procedure. SelectScale has access to U_{in}, and queries labels from S_{in} as necessary. It estimates the generalization error bound GB for several different scales, and executes a procedure similar to binary search to identify a good scale. The binary search keeps the number of estimations (and thus requested labels) small. Crucially, our estimation procedure is designed to prevent the search from spending a number of labels that depends on the net size of the smallest possible scale t, so that the total label complexity of MARMANN depends only on the error of the selected $\hat{t}$. Second, the selected scale $\hat{t}$ is used to generate the compression set by calling $\hat{S} \leftarrow \text{GenerateNNSet}(\hat{t}, [\mathcal{N}(\hat{t})], \delta)$, where GenerateNNSet is our procedure for generating the compression set, and $[\mathcal{N}(\hat{t})] \equiv \{1, \ldots, \mathcal{N}(\hat{t})\}$. Our main result is the following guarantee for MARMANN.

Theorem 4 (Main result; Guarantee for MARMANN**)** *Let $S_{\text{in}} \sim \mathcal{D}^m$, where $m \geq \max(6, |\mathcal{Y}|)$, $\delta \in (0, \frac{1}{4})$. Let $h^{\text{nn}}_{\hat{S}}$ be the output of* MARMANN(U_{in}, δ)*, where $\hat{S} \subseteq \mathcal{X} \times \mathcal{Y}$, and let $\hat{N} := |\hat{S}|$. Let $\hat{h} := h^{\text{nn}}_{\hat{S}}$ and $\hat{\epsilon} := \text{err}(\hat{h}, S_{\text{in}})$, and denote $\hat{G} := \text{GB}(\hat{\epsilon}, \hat{N}, \delta, m, 1)$. With a probability of $1 - \delta$ over S_{in} and randomness of* MARMANN,

$$\text{err}(\hat{h}, \mathcal{D}) \leq 2\hat{G} \leq O\left(G_{\min}(m, \delta)\right),$$

and the number of labels from S_{in} requested by MARMANN *is at most*

$$O\left(\log^3(\frac{m}{\delta})\left(\frac{1}{\hat{G}}\log(\frac{1}{\hat{G}}) + m\hat{G}\right)\right). \tag{2}$$

Here the $O(\cdot)$ notation hides only universal multiplicative constants.

Our error guarantee is thus a constant factor over the error guarantee of the passive learner of (Gottlieb et al., 2017), given in Theorem 3. The constant factor that we derive in our analysis is in the order of 2000 (this can be seen in the proof of Theorem 15). Note that we did not focus on optimizing it, opting instead for a more streamlined analysis. As the lower bound in Theorem 6 shows, the additive term $m\hat{G}$ in Eq. (2) is essentially unavoidable. Whether the dependence on $1/\hat{G}$ is indeed necessary is currently an open problem.

To observe the advantages of MARMANN over a passive learner, consider a scenario in which the upper bound $G_{\min}$ of Theorem 3, as well as the Bayes error of $\mathcal{D}$, are of order $\tilde{\Theta}(1/\sqrt{m})$. Then $\hat{G} = \Theta(1/\sqrt{m})$ as well. Therefore, MARMANN obtains a prediction error guarantee of $\tilde{\Theta}(1/\sqrt{m})$, similarly to the passive learner, but it uses only $\tilde{\Theta}(\sqrt{m})$ labels instead of m. In contrast, the following result shows that no learner that selects labels uniformly at random from S_{in} can compete with MARMANN: Theorem 5 below shows that for any passive learner that uses $\tilde{\Theta}(\sqrt{m})$ random labels from S_{in}, there exists a distribution $\mathcal{D}$ with the above properties, for which the prediction error of the passive learner in this case is $\tilde{\Omega}(m^{-1/4})$, a decay rate which is almost quadratically slower than the $\tilde{O}(1/\sqrt{m})$ rate achieved by MARMANN. Thus, the guarantees of MARMANN cannot be matched by any passive learner.

Theorem 5 (Passive lower bound) *Let $m > 0$ be an integer, and suppose that $(\mathcal{X}, \rho)$ is a metric space such that for some $\bar{t} > 0$, there is a $\bar{t}$-net T of $\mathcal{X}$ of size $\Theta(\sqrt{m})$. Consider any passive learning algorithm that maps i.i.d. samples $S_\ell \sim \mathcal{D}^\ell$ from some distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, 1\}$, to functions $\hat{h}_\ell : \mathcal{X} \to \{-1, 1\}$. For any such algorithm and any $\ell = \tilde{\Theta}(\sqrt{m})$, there exists a distribution $\mathcal{D}$ such that:*

i. The Bayes error of $\mathcal{D}$ is $\Theta(1/\sqrt{m})$;

ii. With at least a constant probability, both of the following events occur:

(a) The passive learner achieves error $\mathrm{err}(\hat{h}_\ell, \mathcal{D}) = \tilde{\Omega}(m^{-1/4})$,

(b) $G_{\min}(m, \delta) = \tilde{\Theta}(1/\sqrt{m})$.

Furthermore, i. and ii. continue to hold when the learning algorithm has access to the full marginal distribution over $\mathcal{X}$.

Thus, MARMANN even improves over a semi-supervised learner: its label savings stem from actively selecting labels, and are not achievable by merely exploiting information from unlabeled data or by randomly selecting examples to label.

We deduce Theorem 5 from a more general result, which might be of independent interest. Theorem 18, given in Section 7, improves existing passive learning sample complexity lower bounds. In particular, our result removes the restrictions of previous lower bounds on the relationship between the sample size, the VC-dimension, and the noise level, which render existing bounds inapplicable to our parameter regime. The proof of Theorem 5 is given thereafter in Section 7, as a consequence of Theorem 18.

We further provide a label complexity lower bound, in Theorem 6 below, which holds for any active learner that obtains similar guarantees to those of MARMANN. The lower bound shows that any active learning algorithm which guarantees a multiplicative accuracy over $G_{\min}(m, \delta)$ has a label complexity which is $\tilde{\Omega}(mG_{\min}(m, \delta))$, for a wide range of values of $G_{\min}(m, \delta)$ — essentially, as long as $G_{\min}(m, \delta)$ is not trivially large or trivially small. This implies that the term $m\hat{G}$

in the upper bound of the label complexity of MARMANN in Theorem 4 cannot be significantly improved.

Theorem 6 (Active lower bound) *Let $\mathcal{X} = \mathbb{R}$, $\delta \in (0, 1/14)$. Let $C \geq 1$, and let $\mathcal{A}$ be an active learning algorithm that outputs $\hat{h}$. Suppose that for any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, if the input unlabeled sample is of size m, then $\mathrm{err}(\hat{h}, \mathcal{D}) \leq CG_{\min}(m, \delta)$ with probability at least $1 - \delta$. Then for any $\alpha \in (\frac{\log(m)+\log(28)}{8\sqrt{2m}}, \frac{1}{240C})$ there exists a distribution $\mathcal{D}$ such with probability at least $\frac{1}{28}$ over $S \sim \mathcal{D}^m$ and the randomness of $\mathcal{A}$, both of the following events hold:*

1. *$\alpha \leq G_{\min}(m, \delta) \leq 30\alpha$*
2. *$\mathcal{A}$ queries at least $\frac{1}{2}\left\lfloor \frac{mG_{\min}(m,\delta)-\log(\frac{m}{\delta})}{30\log(m)} \right\rfloor \equiv \tilde{\Omega}(mG_{\min}(m, \delta))$ labels.*

The proof of this lower bound is provided in Section 8. In the rest of the paper, the components of MARMANN are described in detail, and the main results are proved.

4. Active Nearest-Neighbor at a Given Scale

A main challenge for active learning in our non-parametric setting is performing model selection, that is, selecting a good scale t similarly to the passive learner of Gottlieb et al. (2017). In the passive supervised setting, the approach developed in several previous works (Gottlieb et al., 2014b; Kontorovich and Weiss, 2014; Gottlieb et al., 2014a; Kontorovich and Weiss, 2015) performs model selection by solving a minimum vertex cover problem for each considered scale t, so as to eliminate all of the *t-blocking pairs* — i.e., pairs of differently labeled points within a distance t. The passive algorithm generates a compression set by first finding and removing from S_{in} all points that obstruct (ν, t)-separation at a given scale $t > 0$. This incurs a computational cost but no significant sample complexity increase, aside from the standard logarithmic factor that comes from stratifying over data-dependent hierarchies (Shawe-Taylor et al., 1998).

While this approach works for passive learning, in the active setting we face a crucial challenge: estimating the error of a nearest-neighbor rule at scale t using a small number of samples. A key insight that we exploit in this work is that instead of eliminating the blocking pairs, one may simply relabel some of the points in the compression set, and this would also generate a low-error nearest neighbor rule. This new approach enables estimation of the sample accuracy of a (possibly relabeled) t-net by label-efficient active sampling. In addition, this approach is significantly simpler than estimating the size of the minimum vertex cover of the t-blocking graph. Moreover, we gain improved algorithmic efficiency, by avoiding the relatively expensive vertex cover procedure.

A small technical difference, which will be evident below, is that in this new approach, examples in the compression set might have a different label than their original label in S_{in}. Standard sample compression analysis (e.g. Graepel et al., 2005) assumes that the classifier is determined by a small number of labeled examples from S_{in}. This does not allow the examples in the compression set to have a different label than their original label in S_{in}. Therefore, we require a slight generalization of previous compression analysis (following previous works on compression, see details in Section 6.1), which allows adding side information to the compression set. This side information will be used to set the label of each of the examples in the compression set. The generalization incurs a small statistical penalty, which we quantify in Section 6, as a preliminary to proving Theorem 4.

We now describe our approach to generating a compression set for a given scale $t > 0$. Recall that $\nu(t)$ is the smallest value for which S_{in} is (ν, t)-separated. We define two compression sets. The first one, denoted $S_{\mathrm{a}}(t)$, represents an ideal compression set, constructed (solely for the sake of analysis) so that it induces an empirical error of at most $\nu(t)$. Calculating $S_{\mathrm{a}}(t)$ might require many labels, thus it is only used for analysis purposes; the algorithm never constructs it. The second compression set, denoted $\hat{S}_{\mathrm{a}}(t)$, represents an approximation to $S_{\mathrm{a}}(t)$, which can be constructed using a small number of labels, and induces a sample error of at most $4\nu(t)$ with high probability.

We first define the ideal set $S_{\mathrm{a}}(t) := \{(x_1, y_1), \ldots, (x_N, y_N)\}$. The examples in $S_{\mathrm{a}}(t)$ are the points in $\mathsf{Net}(U_{\mathrm{in}}, t/2)$, and the label of each example is the majority label, out of the labels of the examples in S_{in} to which x_i is closest. Formally, $\{x_1, \ldots, x_N\} := \mathsf{Net}(U_{\mathrm{in}}, t/2)$, and for $i \in [N]$, $y_i := \mathrm{argmax}_{y \in \mathcal{Y}} |S^y \cap P_i|$, where $P_i = \{x \in \mathcal{X} \mid \kappa(x, \mathsf{Net}(U, t/2)) = i\} \in \mathsf{Par}(U_{\mathrm{in}}, t/2)$.

Lemma 7 *Let S be a labeled sample of size m, and let $\{P_1, \ldots, P_N\}$ be a partition of $\mathbb{U}(S)$, with $\max_i \mathsf{diam}(P_i) \le t$ for some $t \ge 0$. For $i \in [N]$, let $\Lambda_i := S^{y_i} \cap P_i$. Then*

$$\nu(t) \ge 1 - \frac{1}{m} \sum_{i \in [N]} |\Lambda_i|.$$

Proof Let $\tilde{S} \subseteq S$ be a subsample that witnesses the $(\nu(t), t)$-separation of S, so that $|\tilde{S}| \ge m(1 - \nu(t))$, and for any two points $(x, y), (x', y') \in \tilde{S}$, if $\rho(x, x') \le t$ then $y = y'$. Denote $\tilde{U} := \mathbb{U}(\tilde{S})$. Since $\max_i \mathsf{diam}(P_i) \le t$, for any $i \in [N]$ all the points in $\tilde{U} \cap P_i$ must have the same label in $\tilde{S}$. Therefore,

$$\exists y \in \mathcal{Y} \text{ s.t. } \tilde{U} \cap P_i \subseteq \tilde{S}^y \cap P_i.$$

Hence $|\tilde{U} \cap P_i| \le |\Lambda_i|$. It follows

$$|S| - \sum_{i \in [N]} |\Lambda_i| \le |S| - \sum_{i \in [N]} |\tilde{U} \cap P_i| = |S| - |\tilde{S}| = m \cdot \nu(t).$$

Dividing by m we get the statement of the lemma. ■

From Lemma 7, we get Cor. 8, which upper bounds the empirical error of $h^{\mathrm{nn}}_{S_{\mathrm{a}}(t)}$ by $\nu(t)$.

Corollary 8 *For every $t > 0$, $\mathrm{err}(h^{\mathrm{nn}}_{S_{\mathrm{a}}(t)}, S_{\mathrm{in}}) \le \nu(t)$.*

This corollary is immediate from Lemma 7, since for any $P_i \in \mathsf{Par}(U_{\mathrm{in}}, t/2)$, $\mathsf{diam}(P_i) \le t$, and

$$\mathrm{err}(h^{\mathrm{nn}}_{S_{\mathrm{a}}(t)}, S_{\mathrm{in}}) = 1 - \frac{1}{m} \sum_{i \in [N]} |\Lambda_i|.$$

Now, calculating $S_{\mathrm{a}}(t)$ requires knowing most of the labels in S_{in}. MARMANN constructs instead an approximation $\hat{S}_{\mathrm{a}}(t)$, in which the examples are the points in $\mathsf{Net}(U_{\mathrm{in}}, t/2)$ (so that $\mathbb{U}(\hat{S}_{\mathrm{a}}(t)) = \mathbb{U}(S_{\mathrm{a}}(t))$), but the labels are determined using a bounded number of labels requested from S_{in}. The labels in $\hat{S}_{\mathrm{a}}(t)$ are calculated by the simple procedure GenerateNNSet given in Alg. 2. The empirical error of the output of GenerateNNSet is bounded in Theorem 9 below.[3]

[3] In the case of binary labels ($|\mathcal{Y}| = 2$), the problem of estimating $S_{\mathrm{a}}(t)$ can be formulated as a special case of the benign noise setting for parametric active learning, for which tight lower and upper bounds are provided in Hanneke and Yang (2015). However, our case is both more general (as we allow multiclass labels) and more specific (as we are dealing with a specific "hypothesis class"). Thus we provide our own procedure and analysis.

A technicality in Alg. 2 requires explanation: In MARMANN, the generation of $\hat{S}_{\mathrm{a}}(t)$ will be split into several calls to GenerateNNSet, so that different calls determine the labels of different points in $\hat{S}_{\mathrm{a}}(t)$. Therefore GenerateNNSet has an additional argument I, which specifies the indices of the points in $\mathsf{Net}(U_{\mathrm{in}}, t/2)$ for which the labels should be returned this time. Crucially, if during the run of MARMANN, GenerateNNSet is called again for the same scale t and the same point in $\mathsf{Net}(U_{\mathrm{in}}, t/2)$, then GenerateNNSet returns the same label that it returned before, rather than recalculating it using fresh labels from S_{in}. This guarantees that despite the randomness in GenerateNNSet, the full $\hat{S}_{\mathrm{a}}(t)$ is well-defined within any single run of MARMANN, and is distributed like the output of GenerateNNSet$(t, [\mathcal{N}(t/2)], \delta)$, which is convenient for the analysis. Define

$$Q(m) := \lceil 18 \log(4m^3/\delta) \rceil . \tag{3}$$

Algorithm 2 GenerateNNSet(t, I, δ)

```
input  Scale $t > 0$, a target set $I \subseteq [\mathcal{N}(t/2)]$, confidence $\delta \in (0,1)$.
output  A labeled set $S \subseteq \mathcal{X} \times \mathcal{Y}$ of size $|I|$
  $\{x_1, \ldots, x_N\} \leftarrow \mathsf{Net}(U_{\mathrm{in}}, t/2)$, $\{P_1, \ldots, P_N\} \leftarrow \mathsf{Par}(U_{\mathrm{in}}, t/2)$, $S \leftarrow ()$
  for $i \in I$ do
    if $\hat{y}_i$ has not already been calculated for $U_{\mathrm{in}}$ with this value of $t$ then
      Draw $Q(m)$ points uniformly at random from $P_i$ and query their labels.
      Let $\hat{y}_i$ be the majority label observed in these $Q(m)$ queries.
    end if
    $S \leftarrow S \cup \{(x_i, \hat{y}_i)\}$.
  end for
  Output $S$
```

Theorem 9 *Let $\hat{S}_{\mathrm{a}}(t)$ be the output of* GenerateNNSet$(t, [\mathcal{N}(t/2)], \delta)$. *With a probability at least $1 - \frac{\delta}{2m^2}$, the following event, which we denote by $E(t)$, holds:*

$$\mathrm{err}(h^{\mathrm{nn}}_{\hat{S}_{\mathrm{a}}(t)}, S_{\mathrm{in}}) \leq 4\nu(t).$$

Proof By Cor. 8, $\mathrm{err}(h^{\mathrm{nn}}_{S_{\mathrm{a}}(t)}, S_{\mathrm{in}}) \leq \nu(t)$. In $S_{\mathrm{a}}(t)$, the labels assigned to each point in $\mathsf{Net}(U_{\mathrm{in}}, t/2)$ are the majority labels (based on S_{in}) of the points in the regions in $\mathsf{Par}(U_{\mathrm{in}}, t/2)$. As above, we denote the majority label for region P_i by $y_i := \mathrm{argmax}_{y \in \mathcal{Y}} |S^y \cap P_i|$. We now compare these labels to the labels $\hat{y}_i$ assigned by Alg. 2. Let $p(i) = |\Lambda_i|/|P_i|$ be the fraction of points in P_i which are labeled by the majority label y_i, where Λ_i is as defined in Lemma 7. Let $\hat{p}(i)$ be the fraction of labels equal to y_i out of those queried by Alg. 2 in round i. Let $\beta := 1/6$. By Hoeffding's inequality and union bounds, we have that with a probability of at least

$$1 - 2\mathcal{N}(t/2)\exp(-\frac{Q(m)}{18}) \geq 1 - \frac{\delta}{2m^2},$$

we have $\max_{i \in [\mathcal{N}(t/2)]} |\hat{p}(i) - p(i)| \leq \beta$. Denote this "good" event by E'. We now prove that $E' \Rightarrow E(t)$. Let $J = \{i \in [\mathcal{N}(t/2)] \mid \hat{p}(i) > \frac{1}{2}\}$. It can be easily seen that $\hat{y}_i = y_i$ for all $i \in J$. Therefore, for all x such that $\kappa(x, \mathbb{U}(S_{\mathrm{a}}(t))) \in J$, $h^{\mathrm{nn}}_{\hat{S}_{\mathrm{a}}(t)}(x) = h^{\mathrm{nn}}_{S_{\mathrm{a}}(t)}(x)$, and hence

$$\mathrm{err}(h^{\mathrm{nn}}_{S}, S_{\mathrm{in}}) \leq \mathbb{P}_{X \sim S_{\mathrm{in}}}[\kappa(X, \mathbb{U}(S_{\mathrm{a}}(t))) \notin J] + \mathrm{err}(h^{\mathrm{nn}}_{S_{\mathrm{a}}(t)}, U_{\mathrm{in}}).$$

The second term is at most $\nu(t)$ by Cor. 8, and it remains to bound the first term, on the condition that E' holds. We have $\mathbb{P}_{X\sim U}[\kappa(X, \mathbb{U}(S_{\mathrm{a}}(t))) \notin J] = \frac{1}{m}\sum_{i\notin J}|P_i|$. If E' holds, then for any $i \notin J$, $p(i) \leq \frac{1}{2} + \beta$, therefore

$$|P_i| - |\Lambda_i| = (1 - p(i))|P_i| \geq (\frac{1}{2} - \beta)|P_i|.$$

Recall that, by Lemma 7, $\nu(t) \geq 1 - \frac{1}{m}\sum_{i\in[\mathcal{N}(t/2)]}|\Lambda_i|$. Therefore,

$$\begin{aligned}\nu(t) &\geq 1 - \frac{1}{m}\sum_{i\in[\mathcal{N}(t/2)]}|\Lambda_i| \\ &= \frac{1}{m}\sum_{i\in[\mathcal{N}(t/2)]}(|P_i| - |\Lambda_i|) \\ &\geq \frac{1}{m}\sum_{i\notin J}(|P_i| - |\Lambda_i|) \\ &\geq \frac{1}{m}\sum_{i\notin J}(\frac{1}{2} - \beta)|P_i|.\end{aligned}$$

Thus, under E',

$$\mathbb{P}_{X\sim U}[\kappa(X, \mathbb{U}(S_{\mathrm{a}}(t))) \notin J] \leq \frac{\nu(t)}{\frac{1}{2} - \beta} = 3\nu(t).$$

It follows that under E', $\mathrm{err}(h_S^{\mathrm{nn}}, U_{\mathrm{in}}) \leq 4\nu(t)$. ■

5. Model Selection

We now show how to select the scale $\hat{t}$ that will be used to generate the output nearest-neighbor rule. The main challenge is to do this with a low label complexity: Generating the full classification rule for scale t requires a number of labels that depends on $\mathcal{N}(t)$, which might be very large. We would like the label complexity of MARMANN to depend only on $\mathcal{N}(\hat{t})$ (where $\hat{t}$ is the selected scale), which is of the order $m\hat{G}$. Therefore, during model selection we can only invest a bounded number of labels in each tested scale. In addition, to keep the label complexity low, we would like to avoid testing all scales. In Section 5.1 we describe how we estimate the error on a given scale. In Section 5.2 we provide a search procedure, resembling binary search, which uses the estimation procedure to select a single scale $\hat{t}$.

5.1 Estimating the Error at a Given Scale

For $t > 0$, let $\hat{S}_{\mathrm{a}}(t)$ be the compressed sample that MARMANN would generate if the selected scale were set to t. Our model selection procedure performs a search, similar to binary search, over the possible scales. For each tested scale t, the procedure estimates the empirical error $\epsilon(t) := \mathrm{err}(h^{\mathrm{nn}}_{\hat{S}_{\mathrm{a}}(t)}, S)$ within a certain accuracy, using an estimation procedure given below, called EstimateErr. EstimateErr outputs an estimate $\hat{\epsilon}(t)$ of $\epsilon(t)$, up to a given threshold $\theta > 0$, using labels requested from S_{in}.

To estimate the error, we sample random labeled examples from S_{in}, and check the prediction error of $h^{\mathrm{nn}}_{\hat{S}_{\mathrm{a}}(t)}$ on these examples. The prediction error of any fixed hypothesis h on a random labeled example from S_{in} is an independent Bernoulli variable with expectation $\mathrm{err}(h, S_{\mathrm{in}})$. EstimateErr is implemented using the following procedure, EstBer, which adaptively estimates the expectation of a Bernoulli random variable to an accuracy specified by the parameter θ, using a small number of random independent Bernoulli experiments. Let $B_1, B_2, \ldots \in \{0,1\}$ be i.i.d. Bernoulli random variables. For an integer n, denote $\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} B_i$. The estimation procedure EstBer is given in Alg. 3. We prove a guarantee for this procedure in Theorem 10. Note that we assume that the threshold parameter is in $(0,1]$, since for $\theta \geq 1$ one can simply output 1 using zero random draws to satisfy Theorem 10.

Algorithm 3 EstBer(θ, β, δ)

```
input  A threshold parameter $\theta \in (0,1]$, a budget parameter $\beta \geq 7$, confidence $\delta \in (0,1)$
  $S \leftarrow \{B_1, \ldots, B_4\}$
  $K \leftarrow \frac{4\beta}{\theta}\log(\frac{8\beta}{\delta\theta})$
  for $i = 3 : \lceil \log_2(\beta \log(2K/\delta)/\theta) \rceil$ do
    $n \leftarrow 2^i$
    $S \leftarrow S \cup \{B_{n/2+1}, \ldots, B_n\}$.
    if $\hat{p}_n > \beta \log(2n/\delta)/n$ then
      break
    end if
  end for
  Output $\hat{p}_n$.
```

The following theorem states that Alg. 3 essentially estimates p, the expectation of the i.i.d. Bernoulli variables $B_1, B_2, \ldots$, up to a multiplicative constant, except if p is smaller than a value proportional to the threshold θ, in which case the algorithm simply returns a value at most θ. Moreover, the theorem shows that the number of random draws required by the algorithm is inversely proportional to the maximum of the threshold θ and the expectation p. Thus, if p is very small, the number of random draws does not increase without bound. The parameter β controls the trade-off between the accuracy of estimation and the number of random draws.

Theorem 10 *Let $\delta \in (0,1)$, $\theta \in (0,1]$, $\beta \geq 7$. Let $B_1, B_2, \ldots \in \{0,1\}$ be i.i.d Bernoulli random variables with expectation p. Let p_o be the output of* EstBer(θ, β, δ). *The following holds with a probability of $1-\delta$, where $f(\beta) := 1 + \frac{8}{3\beta} + \sqrt{\frac{2}{\beta}}$.*

1. *If $p_o \leq \theta$, then $p \leq f(\beta)\theta$. Otherwise, $\frac{p}{f(\beta)} \leq p_o \leq \frac{p}{2-f(\beta)}$.*

2. *Let $\psi := \max(\theta, p/f(\beta))$. The number of random draws in* EstBer *is at most $\frac{4\beta \log(\frac{8\beta}{\delta\psi})}{\psi}$.*

Proof First, consider any single round i with $n = 2^i$. By the empirical Bernstein bound (Maurer and Pontil, 2009, Theorem 4), with a probability of $1 - \delta/n$, for $n \geq 8$, we have[4]

$$|\hat{p}_n - p| \leq \frac{8\log(2n/\delta)}{3n} + \sqrt{\frac{2\hat{p}_n \log(2n/\delta)}{n}}. \tag{4}$$

[4] This follows from Theorem 4 of Maurer and Pontil (2009) since $\frac{7}{3(n-1)} \leq \frac{8}{3n}$ for $n \geq 8$.

Define $g := (\beta + 8/3 + \sqrt{2\beta})$, so that $f(\beta) = g/\beta$. Conditioned on Eq. (4), there are two cases:

(a) $\hat{p}_n \leq \beta \log(2n/\delta)/n$. In this case, $p \leq g \log(2n/\delta)/n$.

(b) $\hat{p}_n > \beta \log(2n/\delta)/n$. In this case, $n \geq \beta \log(2n/\delta)/\hat{p}_n$. Thus, by Eq. (4),

$$|\hat{p}_n - p| \leq \hat{p}_n(\frac{8}{3\beta} + \sqrt{2/\beta}) = \hat{p}_n(g/\beta - 1).$$

Therefore

$$\frac{\beta p}{g} \leq \hat{p}_n \leq \frac{p}{2 - g/\beta}.$$

Taking a union bound on all the rounds, we have that the guarantee holds for all rounds with a probability of at least $1 - \delta$.

Condition now on the event that these guarantees all hold. First, we prove the label complexity bound. Note that since $\beta \geq 7$, $K \geq 28$, thus we have $2\log(2K) > 8$, therefore there is always at least one round. Let n_o be the value of n in the last round that the algorithm runs, and let $p_o = \hat{p}_{n_o}$. Let i such that $n_o = 2^{i+1}$, thus the algorithm stops during round $i + 1$. This implies $\hat{p}_n \leq \beta \log(2n/\delta)/n$ for $n = 2^i$, therefore case (a) holds for n, which means $n \leq g \log(2n/\delta)/p$. It follows that $n \leq 2g\log(4g/\delta)/p$, therefore $n_o \leq 4g \log(4g/(\delta p))/p$. In addition, the number of random draws in the algorithm is n_0, which is bounded by

$$n_0 \leq 2^{\lceil \log_2(\beta \log(2K/\delta)/\theta) \rceil} \leq 2 \cdot 2^{\log_2(\beta \log(2K/\delta)/\theta)} \leq 2\beta \log(2K/\delta)/\theta.$$

Therefore we have the following bound on the number of random draws:

$$n_o \leq \min\left(\frac{2\beta \log(2K/\delta)}{\theta}, \frac{4g \log(4g/(\delta p))}{p}\right).$$

Plugging in the definition of K and substituting $\beta \cdot f(\beta)$ for g yields

$$\begin{aligned} n_0 \leq & \beta \min\left(\frac{2\log(\frac{8\beta}{\delta\theta}\log(\frac{8\beta}{\delta\theta}))}{\theta}, \frac{4f(\beta)\log(\frac{4\beta f(\beta)}{\delta p})}{p}\right) \leq \\ & \beta \min\left(\frac{4\log(\frac{8\beta}{\delta\theta})}{\theta}, \frac{4f(\beta)\log(\frac{4\beta f(\beta)}{\delta p})}{p}\right) \leq \\ & 4\beta \min\left(\frac{1}{\theta}(\log(\frac{8\beta}{\delta}) + \log(\frac{1}{\theta})), \frac{f(\beta)}{p}(\log(\frac{4\beta}{\delta}) + \log(\frac{f(\beta)}{p}))\right). \end{aligned}$$

Using the definition of ψ, we get that the number of draws is at most $\frac{4\beta\log(\frac{8\beta}{\delta\psi})}{\psi}$.

Next, we prove the accuracy of p_o (item 1 in the theorem statement) by considering two cases.

(I) If $p_o > \beta \log(2n_o/\delta)/n_o$, then case (b) above holds for n_o, thus

$$\frac{\beta p}{g} \leq p_o \leq \frac{p}{2 - g/\beta}.$$

In addition, if $p_o \leq \theta$, the LHS implies $p \leq f(\beta)\theta$. Thus item 1 in the theorem statement holds in this case.

(II) If $p_o \leq \beta \log(2n_o/\delta)/n_o$, then EstBer could not have ended by breaking out of the loop, thus it ran until the last round. Therefore $n_o \geq \beta \log(2K/\delta)/\theta$. In addition, case (a) holds for n_0, therefore

$$p \leq \frac{g \log(2n_o/\delta)}{n_o} \leq \frac{g\theta \log(2n_0/\delta)}{\beta \log(2K/\delta)}. \tag{5}$$

Now, for any possible value of n_o,

$$n_o \leq 2\beta \log(2K/\delta)/\theta \leq K.$$

The first inequality follows from the bound on i in EstBer, and the second inequality holds since, as defined in EstBer, $K \geq \frac{4\beta}{\theta} \log(\frac{8\beta}{\theta\delta})$. Since $n_0 \leq K$, Eq. (5) implies that

$$p \leq \frac{g\theta}{\beta} = f(\beta)\theta.$$

In addition, we have

$$p_o \leq \beta \log(2n_o/\delta)/n_o \leq \frac{\theta \log(2n_0/\delta)}{\log(2K/\delta)} \leq \theta.$$

Therefore in this case, necessarily $p_o \leq \theta$ and $p \leq f(\beta)\theta$, which satisfies item 1 in the theorem statement.

In both cases item 1 holds, thus the theorem is proved. ■

The procedure EstimateErr(t, θ, δ) is then implemented as follows:

- Call EstBer$(\theta, 52, \delta/(2m^2))$, where the random variables B_i are independent copies of the Bernoulli variable

$$B := \mathbb{I}[h^{\mathrm{nn}}_{\hat{S}_{\mathrm{a}}(t)}(X) \neq Y]$$

  and $(X, Y) \sim S_{\mathrm{in}}$.

- To draw a single B_i, sample a random pair (x', y') from S_{in}, set

$$i := \kappa(x', \mathsf{Net}(U_{\mathrm{in}}, t/2)),$$

  and get $S \leftarrow$ GenerateNNSet$(t, \{i\}, \delta)$. This returns $S = ((x_i, \hat{y}_i))$ where $\hat{y}_i$ is the label of x_i in $\hat{S}_{\mathrm{a}}(t)$. Then $B_i := \mathbb{I}[\hat{y}_i \neq y']$. Note that B_i is indeed distributed like B, and $\mathbb{E}[B] = \epsilon(t)$. Note further that this call to GenerateNNSet$(t, \{i\}, \delta)$ uses $Q(m)$ label queries. Therefore the overall label complexity of a single draw of a B_i is $Q(m) + 1$.

Cor. 11 gives a guarantee for the accuracy and label complexity of EstimateErr. The proof is immediate from Theorem 10, by setting $\beta = 52$, which implies $f(\beta) \leq 5/4$.

Corollary 11 *Let $t, \theta > 0$ and $\delta \in (0, 1)$, and let $\hat{\epsilon}(t) \leftarrow$* EstimateErr$(t, \theta, \delta)$*. Let $Q(m)$ as defined in Eq. (3) The following properties hold with a probability of $1 - \frac{\delta}{2m^2}$ over the randomness of* EstimateErr *(and conditioned on $\hat{S}_{\mathrm{a}}(t)$).*

1. *If $\hat{\epsilon}(t) \leq \theta$, then $\epsilon(t) \leq 5\theta/4$. Otherwise,*
$$\frac{4\epsilon(t)}{5} \leq \hat{\epsilon}(t) \leq \frac{4\epsilon(t)}{3}.$$

2. *Let $\psi' := \max(\theta, \epsilon(t))$. The number of labels that* EstimateErr *requests is at most*
$$\frac{260(Q(m)+1)\log(\frac{1040m^2}{\delta\psi'})}{\psi'}.$$

To derive item 2. above from Theorem 10, note that for $\beta = 52$,

$$\psi' = \max(\theta, \epsilon(t)) \leq f(\beta)\max(\theta, \epsilon(t)/f(\beta)) = f(\beta)\psi \leq \frac{5}{4}\psi,$$

where ψ is as defined in Theorem 10. Below we denote the event that the two properties in Cor. 11 hold for t by $V(t)$.

5.2 Selecting a Scale

The model selection procedure SelectScale, given in Alg. 4, implements its search based on the guarantees in Cor. 11. First, we introduce some notation. We would like MARMANN to obtain a generalization guarantee that is competitive with $G_{\min}(m, \delta)$. Denote

$$\phi(t) := \frac{(\mathcal{N}(t)+1)\log(m) + \log(\frac{1}{\delta})}{m}, \tag{6}$$

and let

$$G(\epsilon, t) := \epsilon + \frac{2}{3}\phi(t) + \frac{3}{\sqrt{2}}\sqrt{\epsilon\phi(t)}.$$

Note that for all ϵ, t,

$$\mathrm{GB}(\epsilon, \mathcal{N}(t), \delta, m, 1) = \frac{m}{m - \mathcal{N}(t)} G(\epsilon, t).$$

When referring to $G(\nu(t), t)$, $G(\epsilon(t), t)$, or $G(\hat{\epsilon}(t), t)$ we omit the second t for brevity.

Instead of directly optimizing $G(\nu(t))$, we will select a scale based on our estimate $G(\hat{\epsilon}(t))$ of $G(\epsilon(t))$. Let Dist denote the set of pairwise distances in the unlabeled dataset U_{in} (note that $|\mathrm{Dist}| < \binom{m}{2}$). We remove from Dist some distances, so that the remaining distances have a net size $\mathcal{N}(t)$ that is monotone non-increasing in t. We also remove values with a very large net size. Concretely, define

$$\mathrm{Dist}_{\mathrm{mon}} := \mathrm{Dist} \setminus \{t \mid \mathcal{N}(t) + 1 > m/2\} \setminus \{t \mid \exists t' \in \mathrm{Dist}, t' < t \text{ and } \mathcal{N}(t') < \mathcal{N}(t)\}.$$

Then for all $t, t' \in \mathrm{Dist}_{\mathrm{mon}}$ such that $t' < t$, we have $\mathcal{N}(t') \geq \mathcal{N}(t)$. The output of SelectScale is always a value in $\mathrm{Dist}_{\mathrm{mon}}$. The following lemma shows that it suffices to consider these scales.

Lemma 12 *Assume $m \geq 6$ and let $t^*_m \in \mathrm{argmin}_{t \in \mathrm{Dist}} G(\nu(t))$. If $G_{\min}(m, \delta) \leq 1/3$ then $t^*_m \in \mathrm{Dist}_{\mathrm{mon}}$.*

Algorithm 4 SelectScale(δ)

```
input $\delta \in (0,1)$
output Scale $\hat{t}$
$\mathcal{T} \leftarrow \mathrm{Dist}_{\mathrm{mon}}$,        # $\mathcal{T}$ maintains the current set of possible scales
while $\mathcal{T} \neq \emptyset$ do
    $t \leftarrow$ the median value in $\mathcal{T}$            # break ties arbitrarily
    $\hat{\epsilon}(t) \leftarrow \mathsf{EstimateErr}(t, \phi(t), \delta)$.
    if $\hat{\epsilon}(t) < \phi(t)$ then
        $\mathcal{T} \leftarrow \mathcal{T} \setminus [0,t]$ # go right in the binary search
    else if $\hat{\epsilon}(t) > \frac{11}{10}\phi(t)$ then
        $\mathcal{T} \leftarrow \mathcal{T} \setminus [t,\infty)$ # go left in the binary search
    else
        $t_0 \leftarrow t, \mathcal{T}_0 \leftarrow \{t_0\}$.
        break from loop
    end if
end while
if $\mathcal{T}_0$ was not set yet then
    If the algorithm ever went to the right, let $t_0$ be the last value for which this happened, and
    let $\mathcal{T}_0 := \{t_0\}$. Otherwise, $\mathcal{T}_0 := \emptyset$.
end if
Let $\mathcal{T}_L$ be the set of all $t$ that were tested and made the search go left
Output $\hat{t} := \mathrm{argmin}_{t \in \mathcal{T}_L \cup \mathcal{T}_0} G(\hat{\epsilon}(t))$
```

Proof Assume by way of contradiction that $t_m^* \in \mathrm{Dist} \setminus \mathrm{Dist}_{\mathrm{mon}}$. First, since $G(\nu(t_m^*)) \leq G_{\min}(m,\delta) \leq 1/3$ we have

$$\frac{\mathcal{N}(t_m^*)+1}{m-\mathcal{N}(t_m^*)} \log(m) \leq \frac{1}{2}.$$

Therefore, since $m \geq 6$, it is easy to verify $\mathcal{N}(t_m^*)+1 \leq m/2$. Therefore, by definition of $\mathrm{Dist}_{\mathrm{mon}}$ there exists a $t \leq t_m^*$ with $\phi(t) < \phi(t_m^*)$. Since $\nu(t)$ is monotone over all of $t \in \mathrm{Dist}$, we also have $\nu(t) \leq \nu(t_m^*)$. Now, $\phi(t) < \phi(t_m^*)$ and $\nu(t) \leq \nu(t_m^*)$ together imply that $G(\nu(t)) < G(\nu(t_m^*))$, a contradiction. Hence, $t_m^* \in \mathrm{Dist}_{\mathrm{mon}}$. ■

SelectScale follows a search procedure similar to binary search, however the conditions for going right and for going left are not exhaustive, thus it is possible that neither condition holds. The search ends either when neither conditions hold, or when no additional scale should be tested. The final output of the algorithm is based on minimizing $G(\hat{\epsilon}(t))$ over some of the values tested during search.

For $c > 0$, define

$$\gamma(c) := 1 + \frac{2}{3c} + \frac{3}{\sqrt{2c}} \text{ and } \tilde{\gamma}(c) := \frac{1}{c} + \frac{2}{3} + \frac{3}{\sqrt{2c}}.$$

For all $t, \epsilon > 0$ we have the implications

$$\epsilon \geq c\phi(t) \;\Rightarrow\; \gamma(c)\epsilon \geq G(\epsilon,t) \quad \text{and} \quad \phi(t) \geq c\epsilon \;\Rightarrow\; \tilde{\gamma}(c)\phi(t) \geq G(\epsilon,t). \tag{7}$$

The following lemma uses Eq. (7) to show that the estimate $G(\hat{\epsilon}(t))$ is close to the true $G(\epsilon(t))$.

Lemma 13 *Let $t > 0$, $\delta \in (0,1)$, and suppose that* $\mathsf{SelectScale}$ *calls* $\hat{\epsilon}(t) \leftarrow \mathsf{EstimateErr}(t, \phi(t), \delta)$. *Suppose that $V(t)$ as defined in Cor. 11 holds. Then*

$$\frac{1}{6}G(\hat{\epsilon}(t)) \leq G(\epsilon(t)) \leq 6.5G(\hat{\epsilon}(t)).$$

Proof Under $V(t)$, we have that if $\hat{\epsilon}(t) < \phi(t)$ then $\epsilon(t) \leq \frac{5}{4}\phi(t)$. In this case,

$$G(\epsilon(t)) \leq \tilde{\gamma}(4/5)\phi(t) \leq 4.3\phi(t),$$

by Eq. (7). Therefore

$$G(\epsilon(t)) \leq \frac{3 \cdot 4.3}{2} G(\hat{\epsilon}(t)).$$

In addition, $G(\epsilon(t)) \geq \frac{2}{3}\phi(t)$ (from the definition of G), and by Eq. (7) and $\tilde{\gamma}(1) \leq 4$,

$$\phi(t) \geq \frac{1}{4}G(\hat{\epsilon}(t)).$$

Therefore $G(\epsilon(t)) \geq \frac{1}{6}G(\hat{\epsilon}(t))$. On the other hand, if $\hat{\epsilon}(t) \geq \phi(t)$, then by Cor. 11

$$\frac{4}{5}\epsilon(t) \leq \hat{\epsilon}(t) \leq \frac{4}{3}\epsilon(t).$$

Therefore $G(\hat{\epsilon}(t)) \leq \frac{4}{3}G(\epsilon(t))$ and $G(\epsilon(t)) \leq \frac{5}{4}G(\hat{\epsilon}(t))$. Taking the worst-case of both possibilities, we get the bounds in the lemma. ■

The next theorem bounds the label complexity of $\mathsf{SelectScale}$. Let $\mathcal{T}_{\mathrm{test}} \subseteq \mathrm{Dist}_{\mathrm{mon}}$ be the set of scales that are tested during $\mathsf{SelectScale}$ (that is, their $\hat{\epsilon}(t)$ was estimated).

Theorem 14 *Suppose that the event $V(t)$ defined in Cor. 11 holds for all $t \in \mathcal{T}_{\mathrm{test}}$ for the calls $\hat{\epsilon}(t) \leftarrow \mathsf{EstimateErr}(t, \phi(t), \delta)$. If the output of* $\mathsf{SelectScale}$ *is $\hat{t}$, then the number of labels requested by* $\mathsf{SelectScale}$ *is at most*

$$9620|\mathcal{T}_{\mathrm{test}}|(Q(m)+1)\frac{1}{G(\epsilon(\hat{t}))}\log(\frac{38480m^2}{\delta G(\epsilon(\hat{t}))}).$$

Proof The only labels used by the procedure are those used by calls to $\mathsf{EstimateErr}$. Let $\psi_t := \max(\phi(t), \epsilon(t))$, and $\psi_{\min} := \min_{t \in \mathcal{T}_{\mathrm{test}}} \psi_t$. Denote also $\hat{\psi_t} := \max(\phi(t), \hat{\epsilon}(t))$. From Cor. 11 we have that the total number of labels in all the calls to $\mathsf{EstimateErr}$ in $\mathsf{SelectScale}$ is at most

$$\sum_{t \in \mathcal{T}_{\mathrm{test}}} \frac{260(Q(m)+1)\log(\frac{1040m^2}{\delta\psi_t})}{\psi_t} \leq |\mathcal{T}_{\mathrm{test}}| \frac{260(Q(m)+1)\log(\frac{1040m^2}{\delta\psi_{\min}})}{\psi_{\min}}. \tag{8}$$

We now lower bound $\psi_{\min}$ using $G(\epsilon(\hat{t}))$. By Lemma 13 and the choice of $\hat{t}$,

$$G(\epsilon(\hat{t})) \leq 6.5G(\hat{\epsilon}(\hat{t})) = 6.5 \min_{t \in \mathcal{T}_L \cup T_0} G(\hat{\epsilon}(t)).$$

From the definition of G, for any $t > 0$,

$$G(\hat{\epsilon}(t)) \leq \gamma(1) \max(\phi(t), \hat{\epsilon}(t)) \leq 25\hat{\psi}_t.$$

Therefore

$$G(\epsilon(\hat{t})) \leq 25 \min_{t \in \mathcal{T}_L \cup \mathcal{T}_0} \hat{\psi}_t. \tag{9}$$

We will show a similar upper bound when minimizing over all of $\mathcal{T}_{\text{test}}$, not just over $\mathcal{T}_L \cup \mathcal{T}_0$. This is trivial if $\mathcal{T}_{\text{test}} = \mathcal{T}_L \cup \mathcal{T}_0$. Consider the case $\mathcal{T}_L \cup \mathcal{T}_0 \subsetneq \mathcal{T}_{\text{test}}$. For any $t \in \mathcal{T}_{\text{test}}$, we have one of:

- The search went left on t (step 8), hence $t \in \mathcal{T}_L$.
- The search went nowhere on t and the loop broke (step 11), hence $t = t_0 \in \mathcal{T}_0$.
- The search went right on t (step 6) and this was the last value for which this happened, hence $t = t_0 \in \mathcal{T}_0$.
- The search went right on t (step 6) and this was *not* the last value for which this happened. Hence $t \in \mathcal{T}_{\text{test}} \setminus (\mathcal{T}_L \cup \mathcal{T}_0)$.

Set some $t_1 \in \mathcal{T}_{\text{test}} \setminus (\mathcal{T}_L \cup \mathcal{T}_0)$. Since the search went right on t_1, then t_0 also exists, since the algorithm did go to the right for some t (see step 15). Since the binary search went right on t_1, we have $\hat{\epsilon}(t_1) \leq \phi(t_1)$. Since the binary search did *not* go left on t_0 (it either broke from the loop or went right), $\hat{\epsilon}(t_0) \leq \frac{11}{10}\phi(t_0)$.

In addition, $t_0 \geq t_1$ (since the search went right at t_1, and t_0 was tested later than t_1), thus $\phi(t_0) \leq \phi(t_1)$ (since $t_0, t_1 \in \text{Dist}_{\text{mon}}$). Therefore,

$$\hat{\psi}_{t_0} = \max(\phi(t_0), \hat{\epsilon}(t_0)) \leq \frac{11}{10}\phi(t_0) \leq \frac{11}{10}\phi(t_1) = \frac{11}{10}\max(\phi(t_1), \hat{\epsilon}(t_1)) = \hat{\psi}_{t_1}.$$

It follows that for any such t_1,

$$\min_{t \in \mathcal{T}_L \cup \mathcal{T}_0} \hat{\psi}_t \leq \frac{11}{10}\hat{\psi}_{t_1}.$$

Therefore

$$\min_{t \in \mathcal{T}_L \cup \mathcal{T}_0} \hat{\psi}_t \leq \frac{11}{10} \min_{t \in \mathcal{T}_{\text{test}}} \hat{\psi}_t.$$

Therefore, by Eq. (9)

$$G(\epsilon(\hat{t})) \leq 27.5 \min_{t \in \mathcal{T}_{\text{test}}} \hat{\psi}_t.$$

By Cor. 11, $\hat{\epsilon}(t) \leq \max(\phi(t), 4\epsilon(t)/3)$, therefore $\hat{\psi}_t \leq \frac{4}{3}\psi_t$. Therefore $G(\epsilon(\hat{t})) \leq 37\psi_{\min}$. Therefore, from Eq. (8), the total number of labels is at most

$$9620|\mathcal{T}_{\text{test}}|(Q(m)+1)\frac{1}{G(\epsilon(\hat{t}))}\log(\frac{38480m^2}{\delta G(\epsilon(\hat{t}))}).$$

■

The following theorem provides a competitive error guarantee for the selected scale $\hat{t}$.

Theorem 15 *Suppose that $V(t)$ and $E(t)$, defined in Cor. 11 and Theorem 9, hold for all values $t \in \mathcal{T}_{\mathrm{test}}$, and that $G_{\min}(m, \delta) \leq 1/3$. Then* SelectScale *outputs $\hat{t} \in \mathrm{Dist}_{\mathrm{mon}}$ such that*

$$\mathrm{GB}(\epsilon(\hat{t}), \mathcal{N}(\hat{t}), \delta, m, 1) \leq O(G_{\min}(m, \delta)),$$

Where the $O(\cdot)$ notation hides only universal multiplicative constants.

The full proof of this theorem is given below. The idea of the proof is as follows: First, we show (using Lemma 13) that it suffices to prove that $G(\nu(t^*_m)) \geq O(G(\hat{\epsilon}(\hat{t})))$ to derive the bound in the theorem. Now, SelectScale ends in one of two cases: either $\mathcal{T}_0$ is set within the loop, or $\mathcal{T} = \emptyset$ and $\mathcal{T}_0$ is set outside the loop. In the first case, neither of the conditions for turning left and turning right holds for t_0, so we have $\hat{\epsilon}(t_0) = \Theta(\phi(t_0))$ (where Θ hides numerical constants). We show that in this case, whether $t^*_m \geq t_0$ or $t^*_m \leq t_0$, $G(\nu(t^*_m)) \geq O(G(\hat{\epsilon}(t_0)))$. In the second case, there exist (except for edge cases, which are also handled) two values $t_0 \in \mathcal{T}_0$ and $t_1 \in \mathcal{T}_L$ such that t_0 caused the binary search to go right, and t_1 caused it to go left, and also $t_0 \leq t_1$, and $(t_0, t_1) \cap \mathrm{Dist}_{\mathrm{mon}} = \emptyset$. We use these facts to show that for $t^*_m \geq t_1$, $G(\nu(t^*_m)) \geq O(G(\hat{\epsilon}(t_1)))$, and for $t^*_m \leq t_0$, $G(\nu(t^*_m)) \geq O(G(\hat{\epsilon}(t_0)))$. Since $\hat{t}$ minimizes over a set that includes t_0 and t_1, this gives $G(\nu(t^*_m)) \geq O(G(\hat{\epsilon}(\hat{t})))$ in all cases.

Proof First, note that it suffices to show that there is a constant C, such that for the output $\hat{t}$ of SelectScale, we have $G(\epsilon(\hat{t})) \leq CG(\nu(t^*_m))$. This is because of the following argument: From Lemma 12 we have that if $G_{\min}(m, \delta) \leq 1/3$, then $t^*_m \in \mathrm{Dist}_{\mathrm{mon}}$. Now

$$G_{\min}(m, \delta) = \frac{m}{m - \mathcal{N}(t^*_m)} G(\nu(t^*_m)) \geq G(\nu(t^*_m)).$$

And, if we have the guarantee on $G(\epsilon(\hat{t}))$ and $G_{\min}(m, \delta) \leq 1/3$ we will have

$$\mathrm{GB}(\epsilon(\hat{t}), \mathcal{N}(\hat{t}), \delta, m, 1) = \frac{m}{m - \mathcal{N}(\hat{t})} G(\epsilon(\hat{t})) \leq 2G(\epsilon(\hat{t})) \leq 2CG(\nu(t^*_m)) \leq 2CG_{\min}(m, \delta). \tag{10}$$

We now prove the existence of such a guarantee and set C. Denote the two conditions checked in SelectScale during the binary search by Condition 1: $\hat{\epsilon}(t) < \phi(t)$ and Condition 2: $\hat{\epsilon}(t) > \frac{11}{10}\phi(t)$. The procedure ends in one of two ways: either $\mathcal{T}_0$ is set within the loop (Case 1), or $\mathcal{T} = \emptyset$ and $\mathcal{T}_0$ is set outside the loop (Case 2). We analyze each case separately.

In Case 1, none of the conditions 1 and 2 hold for t_0. Therefore

$$\phi(t_0) \leq \hat{\epsilon}(t_0) \leq \frac{11}{10}\phi(t_0).$$

Therefore, by Eq. (7),

$$\phi(t_0) \geq G(\hat{\epsilon}(t_0))/\tilde{\gamma}(\frac{10}{11}).$$

By Cor. 11, since $\hat{\epsilon}(t_0) > \phi(t_0)$,

$$\frac{3}{4}\phi(t_0) \leq \frac{3}{4}\hat{\epsilon}(t_0) \leq \epsilon(t_0) \leq \frac{5}{4}\hat{\epsilon}(t_0) \leq \frac{55}{40}\phi(t_0).$$

Suppose $t^*_m \geq t_0$, then

$$G(\nu(t^*_m)) \geq \nu(t^*_m) \geq \nu(t_0) \geq \frac{1}{4}\epsilon(t_0) \geq \frac{3}{16}\phi(t_0).$$

here we used $\epsilon(t_0) \leq 4\nu(t_0)$ by Theorem 9. Therefore, from Eq. (7) and Lemma 13,

$$G(\nu(t^*_m)) \geq \frac{3}{16}\phi(t_0) \geq \frac{3}{16\tilde{\gamma}\left(\frac{40}{55}\right)}G(\epsilon(t_0)) \geq \frac{\frac{1}{2}}{16\tilde{\gamma}(\frac{40}{55})}G(\hat{\epsilon}(t_0)).$$

Now, suppose $t^*_m < t_0$, then

$$G(\nu(t^*_m)) \geq \frac{2}{3}\phi(t^*_m) \geq \frac{2}{3}\phi(t_0) \geq \frac{2}{3\tilde{\gamma}(\frac{10}{11})}G(\hat{\epsilon}(t_0)).$$

In this inequality we used the fact that $t^*_m, t_0 \in \mathrm{Dist}_{\mathrm{mon}}$, hence $\phi(t^*_m) \geq \phi(t_0)$. Combining the two possibilities for t^*_m, we have in Case 1,

$$G(\hat{\epsilon}(t_0)) \leq \max(32\tilde{\gamma}(\frac{40}{55}), \frac{3\tilde{\gamma}(\frac{10}{11})}{2})G(\nu(t^*_m)).$$

Since $\hat{t}$ minimizes $G(\hat{\epsilon}(t))$ on a set that includes t_0, we have, using Lemma 13

$$G(\epsilon(\hat{t})) \leq 6.5G(\hat{\epsilon}(\hat{t})) \leq 6.5G(\hat{\epsilon}(t_0)).$$

Therefore, in Case 1,

$$G(\epsilon(\hat{t})) \leq 6.5\max(32\tilde{\gamma}(\frac{40}{55}), \frac{3\tilde{\gamma}(\frac{10}{11})}{2})G(\nu(t^*_m)). \tag{11}$$

In Case 2, the binary search halted without satisfying Condition 1 nor Condition 2 and with $\mathcal{T} = \emptyset$. Let t_0 be as defined in this case in SelectScale (if it exists), and let t_1 be the smallest value in $\mathcal{T}_L$ (if it exists). At least one of these values must exist. If both values exist, we have $t_0 \leq t_1$ and $(t_0, t_1) \cap \mathrm{Dist}_{\mathrm{mon}} = \emptyset$.

If t_0 exists, it is the last value for which the search went right. We thus have $\hat{\epsilon}(t_0) < \phi(t_0)$. If $t^*_m \leq t_0$, from condition 1 on t_0 and Eq. (7) with $\tilde{\gamma}(1) \leq 4$,

$$G(\nu(t^*_m)) \geq \frac{2}{3}\phi(t^*_m) \geq \frac{2}{3}\phi(t_0) \geq \frac{1}{6}G(\hat{\epsilon}(t_0)).$$

Here we used the monotonicity of ϕ on $t^*_m, t_0 \in \mathrm{Dist}_{\mathrm{mon}}$, and Eq. (7) applied to condition 1 for t_0.

If t_1 exists, the search went left on t_1, thus $\hat{\epsilon}(t_1) > \frac{11}{10}\phi(t_1)$. By Cor. 11, it follows that $\hat{\epsilon}(t_1) \leq \frac{4}{3}\epsilon(t_1)$. Therefore, if $t^*_m \geq t_1$,

$$G(\nu(t^*_m)) \geq \nu(t^*_m) \geq \nu(t_1) \geq \frac{1}{4}\epsilon(t_1) \geq \frac{3}{16}\hat{\epsilon}(t_1) \geq \frac{3}{16\gamma(11/10)}G(\hat{\epsilon}(t_1)).$$

Here we used $\epsilon(t_1) \leq 4\nu(t_1)$ by Theorem 9 and Eq. (7). Combining the two cases for t^*_m, we get that if t_0 exists and $t^*_m \leq t_0$, or t_1 exists and $t^*_m \geq t_1$,

$$G(\nu(t^*_m)) \geq \min(\frac{1}{6}, \frac{3}{16\gamma(11/10)}) \min_{t \in T_E} G(\hat{\epsilon}(t)).$$

where we define $T_E = \{t \in \{t_0, t_1\} \mid t \text{ exists}\}$. We now show that this covers all possible values for t^*_m: If both t_0, t_1 exist, then since $(t_0, t_1) \cap \mathrm{Dist}_{\mathrm{mon}} = \emptyset$, it is impossible to have $t^*_m \in (t_0, t_1)$.

If only t_0 exists, then the search never went left, which means $t_0 = \max(\mathrm{Dist}_{\mathrm{mon}})$, thus $t^*_m \leq t_0$. If only t_1 exists, then the search never went right, which means $t_1 = \min(\mathrm{Dist}_{\mathrm{mon}})$, thus $t^*_m \geq t_1$.

Since $\hat{t}$ minimizes $G(\hat{\epsilon}(t))$ on a set that has T_E as a subset, we have, using Lemma 13 $G(\epsilon(\hat{t})) \leq 6.5G(\hat{\epsilon}(\hat{t})) \leq 6.5 \min_{t\in T_E} G(\hat{\epsilon}(t))$. Therefore in Case 2,

$$G(\nu(t^*_m)) \geq \frac{1}{6.5}\min\left(\frac{1}{6}, \frac{3}{16\gamma(11/10)}\right) G(\epsilon(\hat{t}). \tag{12}$$

From Eq. (11) and Eq. (12) we get that in both cases

$$G(\nu(t^*_m)) \geq \frac{1}{6.5}\min\left(\frac{1}{6}, \frac{3}{16\gamma(11/10)}, \frac{2}{3\tilde{\gamma}(10/11)}, \frac{1}{32\tilde{\gamma}(\frac{40}{55})}\right) G(\epsilon(\hat{t})) \geq G(\epsilon(\hat{t}))/865.$$

Combining this with Eq. (10) we get the statement of the theorem. ■

6. Bounding the Label Complexity of MARMANN

We are now almost ready to prove Theorem 4. Our last missing piece is quantifying the effect of side information on the generalization of sample compression schemes in Section 6.1. We then prove Theorem 4 in Section 6.2.

6.1 Sample Compression with Side Information

It appears that compression-based generalization bounds were independently discovered by Littlestone and Warmuth (1986) and Devroye et al. (1996); some background is given in Floyd and Warmuth (1995). As noted in Section 4, our algorithm relies on a generalized sample compression scheme, which requires side information. This side information is used to represent the labels of the sample points in the compression set. A similar idea appears in Floyd and Warmuth (1995) for hypotheses with short description length. Here we provide a generalization that is useful for the analysis of MARMANN.

Let Σ be a finite alphabet, and define a mapping $\mathsf{Rec}_N : (\mathcal{X} \times \mathcal{Y})^N \times \Sigma^N \to \mathcal{Y}^{\mathcal{X}}$.[5] This is a *reconstruction* function mapping a labeled sequence of size N with side information $T \in \Sigma^N$ to a classifier. For $I \subseteq [|S|]$, denote by $S[I]$ the subsequence of S indexed by I. For a labeled sample S, define the set of possible hypotheses reconstructed from a compression of S of size N with side information in Σ: $\mathcal{H}_N(S) := \left\{h : \mathcal{X} \to \mathcal{Y} \mid h = \mathsf{Rec}_N(S[I], T), I \in [m]^N, T \in \Sigma^N\right\}$. The following result closely follows the sample compression arguments in Graepel et al. (2005, Theorem 2), and Gottlieb et al. (2017, Theorem 6), but incorporates side information.

Theorem 16 *Let m be an integer and $\delta \in (0,1)$. Let $S \sim \mathcal{D}^m$. With probability at least $1-\delta$, if there exist $N < m$ and $h \in \mathcal{H}_N(S)$ with $\epsilon := \mathrm{err}(h, S) \leq \frac{1}{2}$, then* $\mathrm{err}(h, \mathcal{D}) \leq \mathrm{GB}(\epsilon, N, \delta, m, |\Sigma|)$.

Proof We recall a result of Dasgupta and Hsu (2008, Lemma 1): if $\hat{p} \sim \mathrm{Bin}(n,p)/n$ and $\delta > 0$, then the following holds with probability at least $1-\delta$:

$$p \leq \hat{p} + \frac{2}{3n}\log\frac{1}{\delta} + \sqrt{\frac{9\hat{p}(1-\hat{p})}{2n}\log\frac{1}{\delta}}. \tag{13}$$

[5] If $\mathcal{X}$ is infinite, replace $\mathcal{Y}^{\mathcal{X}}$ with the set of measurable functions from $\mathcal{X}$ to $\mathcal{Y}$.

Now fix $N < m$, and suppose that $h \in \mathcal{H}_{\mathcal{N}}(S)$ has $\hat{\epsilon} \leq \frac{1}{2}$. Let $I \in [m]^N, T \in \Sigma^N$ such that $h = \mathsf{Rec}_N(S[I], T)$. We have $\mathrm{err}(h, S[[m] \setminus I]) \leq \frac{\hat{\epsilon} m}{m-N} = \theta\hat{\epsilon}$. Substituting into (13) $p := \mathrm{err}(h, \mathcal{D})$, $n := m - N$ and $\hat{p} := \mathrm{err}(h, S[[m] \setminus I]) \leq \theta\hat{\epsilon}$, yields that for a fixed $S[I]$ and a random $S[[m] \setminus I] \sim \mathcal{D}^{m-N}$, with probability at least $1 - \delta$,

$$\mathrm{err}(h, \mathcal{D}) \leq \theta\hat{\epsilon} + \frac{2}{3(m-N)} \log\frac{1}{\delta} + \sqrt{\frac{9\theta\hat{\epsilon}}{2(m-N)} \log\frac{1}{\delta}}. \tag{14}$$

To make (14) hold simultaneously for all $(I, T) \in [m]^N \times \Sigma^N$, divide δ by $(m|\Sigma|)^N$. To make the claim hold for all $N \in [m]$, stratify (as in Graepel et al. (2005, Lemma 1)) over the (fewer than) m possible choices of N, which amounts to dividing δ by an additional factor of m. ■

For MARMANN, we use the following sample compression scheme with $\Sigma = \mathcal{Y}$. Given a subsequence $S' := S[I] := (x'_1, \ldots, x'_N)$ and $T = (t_1, \ldots, t_N) \in \mathcal{Y}^N$, the reconstruction function $\mathsf{Rec}_N(S[I], T)$ generates the nearest-neighbor rule induced by the labeled sample $\psi(S', T) := ((x'_i, t_i))_{i \in [N]}$. Formally, $\mathsf{Rec}_N(S', T) = h^{\mathrm{nn}}_{\psi(S',T)}$. Note the slight abuse of notation: formally, the y_i in $S_{\mathrm{a}}(t)$ should be encoded as side information T, but for clarity, we have opted to "relabel" the examples $\{x_1, \ldots, x_N\}$ as dictated by the majority in each region. The following corollary is immediate from Theorem 16 and the construction above.

Theorem 17 *Let $m \geq |\mathcal{Y}|$ be an integer, $\delta \in (0, \frac{1}{4})$. Let $S_{\mathrm{in}} \sim \mathcal{D}^m$. With probability at least $1 - \delta$, if there exist $N < m$ and $S \subseteq (\mathcal{X} \times \mathcal{Y})^N$ such that $\mathbb{U}(S) \subseteq U_{\mathrm{in}}$ and $\epsilon := \mathrm{err}(h^{\mathrm{nn}}_S, S_{\mathrm{in}}) \leq \frac{1}{2}$, then $\mathrm{err}(h^{\mathrm{nn}}_S, \mathcal{D}) \leq \mathrm{GB}(\epsilon, N, \delta, m, |\mathcal{Y}|) \leq 2\mathrm{GB}(\epsilon, N, 2\delta, m, 1)$.*

If the compression set includes only the original labels, the compression analysis of Gottlieb et al. (2017) gives the bound $\mathrm{GB}(\epsilon, N, \delta, m, 1)$. Thus the effect of allowing the labels to change is only logarithmic in $|\mathcal{Y}|$, and does not appreciably degrade the prediction error.

6.2 Proof of Theorem 4

The proof of the main theorem, Theorem 4, which gives the guarantee for MARMANN, is almost immediate from Theorem 17, Theorem 9, Theorem 15 and Theorem 14.

Proof [of Theorem 4] We have $|\mathrm{Dist}_{\mathrm{mon}}| \leq \binom{m}{2}$. By a union bound, the events $E(t)$ and $V(t)$ of Theorem 9 and Cor. 11 hold for all $t \in \mathcal{T}_{\mathrm{test}} \subseteq \mathrm{Dist}_{\mathrm{mon}}$ with a probability of at least $1 - \delta/2$. Under these events, we have by Theorem 15 that if $G_{\min}(m, \delta) \leq 1/3$,

$$\mathrm{GB}(\epsilon(\hat{t}), \mathcal{N}(\hat{t}), \delta, m, 1) \leq O\left(\min_t \mathrm{GB}(\nu(t), \mathcal{N}(t), \delta, m, 1)\right).$$

By Theorem 17, with a probability at least $1 - \delta/2$, if $\epsilon(\hat{t}) \leq \frac{1}{2}$ then

$$\mathrm{err}(\hat{h}, \mathcal{D}) \leq 2\mathrm{GB}(\epsilon(\hat{t}), \mathcal{N}(\hat{t}), \delta, m, 1).$$

The statement of the theorem follows. Note that the statement trivially holds for $G_{\min}(m, \delta) \geq 1/3$ and for $\epsilon(\hat{t}) \geq \frac{1}{2}$, thus these conditions can be removed. To bound the label complexity, note that the total number of labels used by MARMANN is at most the number of labels used by SelectScale plus the number of labels used by GenerateNNSet when the final compression set is generated.

By Theorem 14, since $Q(m) = O(\log(m/\delta))$, the number of labels used by SelectScale is at most

$$O\left(|\mathcal{T}_{\text{test}}|\frac{\log^2(m/\delta)}{G(\epsilon(\hat{t}))}\log\left(\frac{1}{G(\epsilon(\hat{t})}\right)\right).$$

In addition,

$$G(\epsilon(\hat{t})) \geq \text{GB}(\epsilon(\hat{t}), \mathcal{N}(\hat{t}), \delta, m, 1) = \hat{G}.$$

The number of tested scales in SelectScale is bounded by

$$|\mathcal{T}_{\text{test}}| \leq \lfloor\log_2(|\text{Dist}_{\text{mon}}|) + 1\rfloor \leq 2\log_2(m)$$

Therefore the number of labels used by SelectScale is

$$O\left(\frac{\log^3(m/\delta)}{\hat{G}}\log\left(\frac{1}{\hat{G}}\right)\right).$$

The number of labels used by GenerateNNSet is at most $Q(m)\mathcal{N}(\hat{t})$, where $Q(m) \leq O(\log(m/\delta)$, and from the definition of $\hat{G}$, $\mathcal{N}(\hat{t}) \leq O(m\hat{G}/\log(m))$. Summing up the number of labels used by SelectScale and the number used by GenerateNNSet, this gives the bound in the statement of the theorem. ■

7. Passive Learning Lower Bounds

Theorem 5 lower bounds the performance of a passive learner who observes a limited number ℓ of random labels from S_{in}. The number ℓ is chosen so that it is of the same order as the number of labels MARMANN observes for the case analyzed in Section 3. We deduce Theorem 5 from a more general result pertaining to the sample complexity of passive learning. The general result is given as Theorem 18 in Section 7.1. The proof of Theorem 5 is provided in Section 7.2.

We note that while the lower bounds below assume that the passive learner observes only the random labeled sample of size ℓ, in fact their proofs hold also if the algorithm has access to the full unlabeled sample of size m of which S_ℓ is sampled. This is because the lower bound is based on requiring the learner to distinguish between distributions that all have the same marginal. Under this scenario, access to unlabeled examples does not provide any additional information to the learner.

7.1 A General Lower Bound

In this section we show a general sample complexity lower bound for passive learning, which may be of independent interest. We are aware of two existing lower bounds for agnostic PAC with bounded Bayes error: Devroye et al. (1996, Theorem 14.5) and Audibert (2009, Theorem 8.8). Both place restrictions on the relationship between the sample size, VC-dimension, and Bayes error level, which render them inapplicable as stated to some parameter regimes, including the one needed for proving Theorem 5.

Let $\mathcal{H}$ be a hypothesis class with VC-dimension d and suppose that $\mathcal{L}$ is a passive learner[6] mapping labeled samples $S_\ell = (X_i, Y_i)_{i\in[\ell]}$ to hypotheses $\hat{h}_\ell \in \mathcal{H}$. For any distribution $\mathcal{D}$ over

[6] We allow $\mathcal{L}$ access to an independent internal source of randomness.

$\mathcal{X} \times \{-1,1\}$, define the *excess risk* of $\hat{h}_\ell$ by

$$\Delta(\hat{h}_\ell, \mathcal{D}) := \operatorname{err}(\hat{h}_\ell, \mathcal{D}) - \inf_{h \in \mathcal{H}} \operatorname{err}(h, \mathcal{D}).$$

Let $\mathscr{D}(\eta)$ be the collection of all *η-bounded agnostic error* distributions $\mathcal{D}$ over $\mathcal{X} \times \{-1,1\}$ that satisfy $\inf_{h\in\mathcal{H}} \operatorname{err}(h,\mathcal{D}) \le \eta$. We say that $Z \in \{-1,1\}$ has Rademacher distribution with parameter $b \in [-1,1]$, denoted $Z \sim \mathrm{R}_b$, if

$$\mathbb{P}[Z = 1] = 1 - \mathbb{P}[Z = -1] = \frac{1}{2} + \frac{b}{2}.$$

All distributions on $\{-1,1\}$ are of this form. For $k \in \mathbb{N}$ and $b \in [0,1]$, define the function

$$\operatorname{bayes}(k,b) = \tfrac{1}{2}\left(1 - \tfrac{1}{2}\left\|\mathrm{R}_b^k - \mathrm{R}_{-b}^k\right\|_1\right),$$

where $\mathrm{R}_{\pm b}^k$ is the corresponding product distribution on $\{-1,1\}^k$ and $\frac{1}{2}\left\|\cdot\right\|_1$ is the total variation norm. This expression previously appeared in Berend and Kontorovich (2015, Equation (25)) in the context of information-theoretic lower bounds; the current terminology was motivated in Kontorovich and Pinelis (2016), where various precise estimates on $\operatorname{bayes}(\cdot)$ were provided. In particular, the function $\operatorname{ba\check{y}es}(\kappa, b)$ was defined as follows: for each fixed $b \in [0,1]$, $\operatorname{ba\check{y}es}(\cdot, b)$ is the largest convex minorant on $[0,\infty)$ of the function $\operatorname{bayes}(\cdot, b)$ on $\{0,1,\dots\}$. It was shown in Kontorovich and Pinelis (2016, Proposition 2.8) that $\operatorname{ba\check{y}es}(\cdot, b)$ is the linear interpolation of $\operatorname{bayes}(\cdot, b)$ at the points $0,1,3,5,\dots$.

Theorem 18 *Let $0 < \eta < \frac{1}{2}$, $\ell \ge 1$, and $\mathcal{H}$ be a hypothesis class with VC-dimension $d > 1$. Then, for all $0 < b, p < 1$ satisfying*

$$p\left(\frac{1}{2} - \frac{b}{2}\right) \le \eta, \tag{15}$$

we have

$$\inf_{\hat{h}_\ell} \sup_{\mathcal{D}\in\mathscr{D}(\eta)} \mathop{\mathbb{E}}_{\mathcal{D}^\ell}\left[\Delta(\hat{h}_\ell, \mathcal{D})\right] \quad \ge \quad bp\,\operatorname{ba\check{y}es}(\ell p/(d-1), b). \tag{16}$$

Furthermore, for $0 \le u < 1$,

$$\inf_{\hat{h}_\ell} \sup_{\mathcal{D}\in\mathscr{D}(\eta)} \mathbb{P}\left[\Delta(\hat{h}_\ell, \mathcal{D}_{\sigma,b,p}) > bpu\right] \quad > \quad \operatorname{ba\check{y}es}(\ell p/(d-1), b) - u. \tag{17}$$

Proof This proof uses ideas from Devroye et al. (1996, Theorem 14.5), Anthony and Bartlett (1999, Theorem 5.2) and Kontorovich and Pinelis (2016, Theorem 2.2).

We will construct adversarial distributions supported on a shattered subset of size d, and hence there is no loss of generality in taking $\mathcal{X} = [d]$ and $\mathcal{H} = \{-1,1\}^{\mathcal{X}}$. A random distribution $\mathcal{D}_{\sigma,b,p}$ over $\mathcal{X} \times \{-1,1\}$, parametrized by a random $\sigma \sim \operatorname{Unif}(\{-1,1\}^{d-1})$ and scalars $b, p \in (0,1)$ to be specified later, is defined as follows. The point $x = d \in \mathcal{X}$ gets a marginal weight of $1-p$ where p is a parameter to be set; the remaining $d-1$ points each get a marginal weight of $p/(d-1)$:

$$\mathop{\mathbb{P}}_{X\sim\mathcal{D}_{\sigma,b,p}}[X = d] = 1 - p, \qquad \mathop{\mathbb{P}}_{X\sim\mathcal{D}_{\sigma,b,p}}[X < d] = \frac{p}{d-1}. \tag{18}$$

The distribution of Y conditional on X is given by $\mathbb{P}_{(X,Y)\sim\mathcal{D}_{\sigma,b,p}}[Y=1\,|\,X=d]=1$ and

$$\mathop{\mathbb{P}}_{(X,Y)\sim\mathcal{D}_{\sigma,b,p}}[Y=\pm1\,|\,X=j<d]=\frac{1}{2}\pm\frac{b\sigma_j}{2}. \tag{19}$$

Suppose that $(X_i,Y_i)_{i\in[\ell]}$ is a sample drawn from $\mathcal{D}^{\ell}_{\sigma,b,p}$. The assumption that $\mathcal{D}_{\sigma,b,p}\in\mathscr{D}(\eta)$ implies that b and p must satisfy the constraint (15).

A standard argument (e.g., Anthony and Bartlett (1999) p. 63 display (5.5)) shows that, for any hypothesis $\hat{h}_\ell$,

$$\begin{aligned}\Delta(\hat{h}_\ell,\mathcal{D}_{\sigma,b,p}) &= \operatorname{err}(\hat{h}_\ell,\mathcal{D}_{\sigma,b,p})-\inf_{h\in\mathcal{H}}\operatorname{err}(h,\mathcal{D}_{\sigma,b,p})\\ &= \mathop{\mathbb{P}}_{X\sim\mathcal{D}_{\sigma,b,p}}[X=d,\hat{h}_\ell(X)\neq1]+b\mathop{\mathbb{P}}_{X\sim\mathcal{D}_{\sigma,b,p}}[X<d,\hat{h}_\ell(X)\neq\sigma(X)]\\ &\geq b\mathop{\mathbb{P}}_{X\sim\mathcal{D}_{\sigma,b,p}}[X<d,\hat{h}_\ell(X)\neq\sigma(X)]\\ &= bp\mathop{\mathbb{P}}_{X\sim\mathcal{D}_{\sigma,b,p}}[\hat{h}_\ell(X)\neq\sigma(X)|X<d]. \end{aligned}\tag{20}$$

It follows from Kontorovich and Pinelis (2016, Theorems 2.1, 2.5) that

$$\begin{aligned}\mathop{\mathbb{E}}_{\sigma}\mathop{\mathbb{P}}_{X\sim\mathcal{D}_{\sigma,b,p}}[\hat{h}_\ell(X)\neq\sigma(X)|X<d] &\geq \mathop{\mathbb{E}}_{N\sim\operatorname{Bin}(\ell,p/(d-1))}[\operatorname{bayes}(N,b)] \qquad (21)\\ &\geq \mathop{\mathbb{E}}_{N\sim\operatorname{Bin}(\ell,p/(d-1))}[\operatorname{ba\check{y}es}(N,b)]\\ &\geq \operatorname{ba\check{y}es}(\mathbb{E}[N],b)=\operatorname{ba\check{y}es}(\ell p/(d-1),b),\end{aligned}$$

where the second inequality holds because $\operatorname{ba\check{y}es}$ is, by definition, a convex minorant of bayes, and the third follows from Jensen's inequality. Combined with (20), this proves (16).

To show (17), define the random variable

$$Z=Z(\sigma,\mathcal{L})=\mathop{\mathbb{P}}_{X\sim\mathcal{D}_{\sigma,b,p}}[\hat{h}_\ell(X)\neq\sigma(X)|X<d].$$

Since $Z\in[0,1]$, Markov's inequality implies

$$\mathbb{P}[Z>u]\geq\frac{\mathbb{E}[Z]-u}{1-u}>\mathbb{E}[Z]-u,\qquad 0\leq u<1.$$

Now (20) implies that $\Delta(\hat{h}_\ell,\mathcal{D}_{\sigma,b,p})\geq bpZ$ and hence, for $0\leq u<1$,

$$\begin{aligned}\inf_{\hat{h}_\ell}\sup_{\mathcal{D}\in\mathscr{D}(\eta)}\mathbb{P}\left[\Delta(\hat{h}_\ell,\mathcal{D}_{\sigma,b,p})>bpu\right] &= \inf_{\hat{h}_\ell}\sup_{\mathcal{D}\in\mathscr{D}(\eta)}\mathbb{P}[Z>u]\\ &> \inf_{\hat{h}_\ell}\sup_{\mathcal{D}\in\mathscr{D}(\eta)}\mathbb{E}[Z]-u\\ &\geq \frac{1}{bp}\inf_{\hat{h}_\ell}\sup_{\mathcal{D}\in\mathscr{D}(\eta)}\mathbb{E}[\Delta(\hat{h}_\ell,\mathcal{D}_{\sigma,b,p})]-u\\ &\geq \operatorname{ba\check{y}es}(\ell p/(d-1),b)-u.\end{aligned}$$

■

7.2 Proof of Theorem 5

We break down the proof into several steps. For now, we assume that the labeled examples are sampled i.i.d. as per the classic PAC setup. At the end, we show how to extend the proof to the semi-supervised setting.

(i) Defining a family of adversarial distributions. Let T be a $\bar{t}$-net of $\mathcal{X}$ of size $\Theta(\sqrt{m})$ and $\eta = \Theta(1/\sqrt{m})$. For any passive learning algorithm mapping i.i.d. samples of size $\ell = \tilde{\Theta}(\sqrt{m})$ to hypotheses $\hat{h}_\ell : \mathcal{X} \to \{-1, 1\}$, we construct a random adversarial distribution $\mathcal{D}_{\sigma,b,p}$ with agnostic error η. We accomplish this via the construction described in the proof of Theorem 18, with $|T| = d = \Theta(\sqrt{m})$. The marginal distribution over $T = \{x_1, \dots, x_d\}$ puts a mass of $1 - p$ on $x_d \in T$ and spreads the remaining mass uniformly over the other points, as in (18). The "heavy" point has a deterministic label and the remaining "light" points have noisy labels drawn from a random distribution with symmetric noise level b, as in (19). We proceed to choose b and p; namely,

$$p = \frac{d-1}{2\ell}\sqrt{\eta} = \tilde{\Theta}(m^{-1/4}), \qquad b = 1 - \frac{2\eta}{p} = 1 - \tilde{\Theta}(m^{-1/4}),$$

which makes the constraint in (15) hold with equality; this means that the agnostic error is exactly η and in particular, establishes (i).

(ii.a) Lower-bounding the passive learner's error. Our choice of p implies that $\ell p/(d-1) = \sqrt{\eta}/2 =: \kappa < 1$. For this range of κ, Kontorovich and Pinelis (2016, Proposition 2.8) implies that $\check{\text{bayes}}(\kappa, b) = \frac{1}{2}(1 - \kappa b) = \Theta(1)$. Choosing $u = \frac{1}{4}(1 - \kappa b) = \Theta(1)$ in (17), Theorem 18 implies that

$$\inf_{\hat{h}_\ell} \sup_{\mathcal{D} \in \mathscr{D}(\eta)} \mathbb{P}[\Delta(\hat{h}_\ell, \mathcal{D}) > \tilde{\Omega}(m^{-1/4})] > \Omega(1).$$

In more formal terms, there exist constants $c_0, c_1 > 0$ such that

$$\inf_{\hat{h}_\ell} \sup_{\mathcal{D} \in \mathscr{D}(\eta)} \mathbb{P}[\Delta(\hat{h}_\ell, \mathcal{D}) > c_0 p] > c_1. \tag{22}$$

(ii.b) Upper-bounding $\nu(\bar{t})$. To establish (ii.b), it suffices to show that for $(X_i, Y_i)_{i \in [\ell]} \sim \mathcal{D}^\ell_{\sigma,b,p}$, we will have $\nu(\bar{t}) = O(m^{-1/2})$ with sufficiently high probability. Indeed, the latter condition implies the requisite upper bound on $\min_{t>0:\mathcal{N}(t)<m} \mathrm{GB}(\nu(t), \mathcal{N}(t), \delta, m, 1)$, while (i) implies the lower bound, since the latter quantity cannot be asymptotically smaller than the Bayes error (which coincides with the agnostic error for $\mathcal{H} = \{-1, 1\}^{\mathcal{X}}$).

Recall that the $\bar{t}$-net points $\{x_1, \dots, x_{d-1}\}$ are the "light" ones (i.e., each has weight $p/(d-1)$) and define the random sets $J_j \subset [\ell]$ by

$$J_j = \{i \in [\ell] : X_i = x_j\}, \qquad j \in [d-1].$$

In words, J_j consists of the indices i of the sample points for which X_i falls on the net point x_j. For $y \in \{-1, 1\}$, put $\tau_j^y = \sum_{i \in J_j} \mathbb{I}[Y_i = y]$ and define the *minority count* ξ_j at the net point x_j by

$$\xi_j \;=\; \min_{y \in \{-1,1\}} \tau_j^y \;=\; \tfrac{1}{2}(|\tau_j^+ + \tau_j^-| - |\tau_j^+ - \tau_j^-|).$$

Observe that by virtue of being a $\bar{t}$-net, T is $\bar{t}$-separated and hence the only contribution to $\nu(\bar{t})$ is from the minority counts (to which the "heavy" point x_d does not contribute due to its deterministic label):

$$\nu(\bar{t}) = \frac{1}{\ell}\sum_{j=1}^{d-1}\xi_j.$$

Now

$$\mathbb{E}\,|\tau_j^+ + \tau_j^-| = \mathbb{E}\,|J_j| = \frac{\ell p}{d-1} = \Theta(m^{-1/4})$$

and

$$\begin{aligned}\mathbb{E}\,|\tau_j^+ - \tau_j^-| &= \mathop{\mathbb{E}}_{\sigma_j}\mathbb{E}[|\tau_j^+ - \tau_j^-||\sigma_j] \\ &\geq \mathop{\mathbb{E}}_{\sigma_j}\left|\mathbb{E}[\tau_j^+ - \tau_j^-|\sigma_j]\right|.\end{aligned}$$

Computing

$$\mathbb{E}[\tau_j^+|\sigma_j = +1] = (\frac{1}{2}+\frac{b}{2})\frac{\ell p}{d-1}, \qquad \mathbb{E}[\tau_j^-|\sigma_j = +1] = (\frac{1}{2}-\frac{b}{2})\frac{\ell p}{d-1},$$

with an analogous calculation when conditioning on $\sigma_j = -1$, we get

$$\mathbb{E}\,|\tau_j^+ - \tau_j^-| \geq \frac{b\ell p}{d-1}$$

and hence

$$\begin{aligned}\mathbb{E}[\xi_j] &\leq \frac{1}{2}\left(\frac{\ell p}{d-1} - b\frac{\ell p}{d-1}\right) \\ &= (1-b)\frac{\ell p}{2(d-1)} = \frac{2\eta}{p}\cdot\frac{\ell p}{2(d-1)} = \frac{\eta\ell}{d-1}.\end{aligned}$$

It follows that

$$\begin{aligned}\mathbb{E}[\nu(\bar{t})] &= \frac{1}{\ell}\sum_{j=1}^{d-1}\mathbb{E}[\xi_j] \\ &\leq \frac{d-1}{\ell}\cdot\frac{\eta\ell}{d-1} = \eta = \Theta(m^{-1/2}).\end{aligned}$$

To give tail bounds on $\nu(\bar{t})$, we use Markov's inequality: for all $c_2 > 0$,

$$\mathbb{P}[\nu(\bar{t}) > c_2\,\mathbb{E}[\nu(\bar{t})] \leq \frac{1}{c_2}.$$

Choosing c_2 sufficiently large that $1 - 1/c_2 > c_1$ (the latter from (22)) implies the existence of constants $c_0, c_2, c_3 > 0$ such that

$$\inf_{\hat{h}_\ell}\sup_{\mathcal{D}\in\mathscr{D}(\eta)}\mathbb{P}\left[\left(\Delta(\hat{h}_\ell,\mathcal{D}) > c_0 p\right)\wedge(\nu(\bar{t}) \leq c_2\eta)\right] > c_3.$$

Since $p = \tilde{\Theta}(m^{-1/4})$ and $\eta = \Theta(m^{-1/2})$, this establishes (ii) and concludes the proof of Theorem 5.

(iii) Extending to the semi-supervised setting. Providing the learner with the exact weights of $\mathcal{X} = [d]$ under our adversarial distribution does not give it any additional power. Indeed, the information-theoretic excess risk lower bound in Eq. (21) hinges on the fact that to estimate $\sigma(x)$ with some desired certainty, the point x must be sampled some minimal number of times. The marginal probability of x does not enter that calculation, and hence knowing its value does not afford the learner an improved performance.

8. Active Learning Lower Bound

We now prove the active learning lower bound stated in Theorem 6. To prove the theorem, we first prove a result which is similar to the classical No-Free-Lunch theorem, except it holds for active learning algorithms. The proof follows closely the proof of the classical No-Free-Lunch theorem given in Shalev-Shwartz and Ben-David (2014, Theorem 5.1), with suitable modifications.

Theorem 19 *Let $\beta \in [0, \frac{1}{2})$, and m be an integer. Let $\mathcal{A}$ any active learning algorithm over a finite domain $\mathcal{X}$ which gets as input a random labeled sample $S \sim \mathcal{D}^m$ (with hidden labels) and outputs $\hat{h}$. If $\mathcal{A}$ queries fewer than $\mathcal{X}/2$ labels from S, then there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$ such that*

- *Its marginal on $\mathcal{X}$ is uniform, and for each $x \in \mathcal{X}$, $\mathbb{P}[Y = 1 \mid X = x] \in \{\beta, 1 - \beta\}$.*
- $\mathbb{E}[\mathrm{err}(\hat{h}, \mathcal{D})] \geq \frac{1}{4}$.

Proof Let $\mathcal{F} = \{f_1, \ldots, f_T\}$ be the set of possible functions $f_i : \mathcal{X} \to \{0, 1\}$. Let $\mathcal{D}_i$ to be a distribution with a uniform marginal over $\mathcal{X}$, and $\mathbb{P}_{(X,Y)\sim\mathcal{D}_i}[Y = 1 \mid X = x] = f_i(x)(1 - \beta) + (1 - f_i(x))\beta$. Consider the following random process: First, draw an unlabeled sample $X = (x_1, \ldots, x_m)$ i.i.d. from $\mathcal{D}_X^m$. Then, draw $B = (b_1, \ldots, b_m)$ independently from a Bernoulli distribution with $\mathbb{P}[b_i = 1] = \beta$. For $i \in [T]$, let $S^i(X, B) = ((x_1, y_1), \ldots, (x_m, y_m))$ such that x_i are set by X, and $y_i = f_i(x)$ if $b_i = 0$ and $1 - f_i(x)$ otherwise. Clearly, $S^i(X, B)$ is distributed according to $\mathcal{D}_i^m$. Let $h_i(S)$ be the output of $\mathcal{A}$ when the labeled sample is S. Denote by $\hat{h}_i$ the (random) output of $\mathcal{A}$ when the sample is drawn from $\mathcal{D}_i$. Clearly

$$\mathbb{E}[\mathrm{err}(\hat{h}_i, \mathcal{D}_i)] = \mathop{\mathbb{E}}_{X,B}[\mathrm{err}(h_i(S(X, B)), \mathcal{D}_i)].$$

Therefore (as in (5.4) in Shalev-Shwartz and Ben-David (2014)), for some j, X, B it holds that

$$\mathbb{E}[\mathrm{err}(\hat{h}_j, \mathcal{D}_j)] \geq \frac{1}{T}\sum_{i=1}^{T} \mathbb{E}[\mathrm{err}(\hat{h}_i, \mathcal{D}_i)] \geq \frac{1}{T}\sum_{i=1}^{T} \mathrm{err}(h_i(S(X, B)), \mathcal{D}_i). \tag{23}$$

Fix X, B, j as above, and denote for brevity $h_i := h_i(S(X, B))$. Let V_i be the set of examples $x \in \mathcal{X}$ for which that $\mathcal{A}$ does not observe their label if the labeled sample is $S^i(X, B)$ (this includes both examples that are not in the sample at all as well as examples that are in the sample but their label is not requested by $\mathcal{A}$). We have $|V_i| > |\mathcal{X}|/2$ by assumption. Then (as in Eq. (5.6) therein)

$$\frac{1}{T}\sum_{i=1}^{T} \mathrm{err}(h_i, \mathcal{D}_i) \geq \frac{1}{T}\sum_{i=1}^{T} \frac{1}{2|V_i|} \sum_{x \in V_i} \mathbb{I}[h_i(x) \neq f_i(x)]. \tag{24}$$

Since $\mathcal{A}$ is active, it selects which examples to request, which can depend on the labels observed by $\mathcal{A}$ so far. Therefore, V_i can be different for different i. However, an argument similar to that of the No-Free-Lunch theorem for the passive case still goes through, as follows.

Let i, i' such that $f_i(x) = f_{i'}(x)$ for all $x \notin V_i$, and $f_i(x) = 1 - f_{i'}(x)$ for all $x \in V_i$. Since X, B are fixed, $\mathcal{A}$ observes the same labels for all $x \notin V_i$ for both $S^{i'}(X, B)$ and $S^i(X, B)$, thus all its decisions and requests are identical for both samples, and so $V_i = V_{i'}$, and $h_i = h_{i'}$. Therefore, it is possible to partition T into $T/2$ pairs of indices i, i' such that for each such pair,

$$\begin{aligned}
&\frac{1}{2|V_i|}\sum_{x\in V_i}\mathbb{I}[h_i(x) \neq f_i(x)] + \frac{1}{2|V_{i'}|}\sum_{x\in V_{i'}}\mathbb{I}[h_{i'}(x) \neq f_{i'}(x)] \\
&= \frac{1}{2|V_i|}\sum_{x\in V_i}\mathbb{I}[h_i(x) \neq f_i(x)] + \mathbb{I}[h_i(x) \neq 1 - f_i(x)] \\
&= \frac{1}{2}.
\end{aligned}$$

Therefore, $\frac{1}{T}\sum_{i=1}^{T} \mathrm{err}(h_i, \mathcal{D}_i) \geq \frac{1}{4}$. Therefore, from Eq. (24), $\frac{1}{T}\sum_{i=1}^{T} \mathrm{err}(h_i, \mathcal{D}_i) \geq \frac{1}{4}$. Combining this with Eq. (23), it follows that $\mathbb{E}[\mathrm{err}(\hat{h}_j, \mathcal{D}_j)] \geq \frac{1}{4}$. ■

We will also make use of the following simple lemma.

Lemma 20 *Let $\beta \in [0, 1]$. Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \{0, 1\}$ such that for $(X, Y) \sim \mathcal{D}$, for any x in the support of $\mathcal{D}$, $\mathbb{P}[Y = 1 \mid X = x] \in \{\beta, 1 - \beta\}$. Let N be the size of the support of $\mathcal{D}$. Let $S \sim \mathcal{D}^m$. Denote by n_x the number of sample pairs (x', y') in S where $x' = x$, and let n_x^+ be the number of sample pairs (x', y') where $x' = x$ and $y' = 1$. Let $\hat{p}_x^+ = n_x^+/n_x$ (or zero if $n_x = 0$). Then*

$$2\beta(1-\beta)(m - N) \leq \sum_{x\in\mathcal{X}} \mathbb{E}[2n_x\hat{p}_x^+(1 - \hat{p}_x^+)] \leq 2\beta(1-\beta)m.$$

Proof We have

$$\mathbb{E}[2n_x\hat{p}_x^+(1 - \hat{p}_x^+)] = \sum_{i=1}^{\infty} \mathbb{P}[n_x = i] \cdot i \cdot \mathbb{E}[2\hat{p}_x^+(1 - \hat{p}_x^+) \mid n_x = i].$$

Note that $\mathbb{E}[2\hat{p}_x^+(1 - \hat{p}_x^+) \mid n_x = 1] = 0$. For $i > 1$, let $y_1, \ldots, y_i$ be the labels of the examples that are equal to x in S, then

$$\sum_{j,k\in[i]} \mathbb{I}[y_k \neq y_j] = 2n_x^+(i - n_x^+) = i^2 \cdot 2\hat{p}_x^+(1 - \hat{p}_x^+).$$

Therefore, letting $(X_1, Y_1), (X_2, Y_2) \sim \mathcal{D}^2$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[2\hat{p}_x^+(1 - \hat{p}_x^+) \mid n_x = i] &= \frac{1}{i^2}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[\sum_{j,k\in[n_x]} \mathbb{I}[y_k \neq y_j] \mid n_x = i] \\
&= \frac{i^2 - i}{i^2}\mathbb{P}[Y_1 \neq Y_2 \mid X_1 = X_2 = x] \\
&= 2(1 - \frac{1}{i})\beta(1-\beta),
\end{aligned}$$

Thus

$$\begin{aligned}\mathbb{E}[2n_x\hat{p}_x^+(1-\hat{p}_x^+)] &= 2\beta(1-\beta)\sum_{i=2}^{\infty}(i-1)\,\mathbb{P}[n_x=i]\\ &= 2\beta(1-\beta)(\mathbb{E}[n_x]+\mathbb{P}[n_x=0]-1).\end{aligned}$$

To complete the proof, sum over all x in the support of $\mathcal{D}$, and note that $\sum_x \mathbb{E}[n_x] = m$, and $\sum_x(\mathbb{P}[n_x=0]-1) \in [-N,0]$. ■

We now prove our lower bound, stated in Theorem 6, on the number of queries required by any active learning with competitive guarantees similar to ours.

Proof [of Theorem 6] Let $N = \left\lfloor \frac{m\alpha - \log(\frac{m}{\delta})}{\log(m)} \right\rfloor$. Let $\beta = 8\alpha \le \frac{1}{2}$. Consider a marginal distribution $\mathcal{D}_X$ over $\mathcal{X}$ which is uniform over N points $1,\ldots,N \in \mathbb{R}$. Consider the following family of distributions: $\mathcal{D}$ such that its marginal over $\mathcal{X}$ is $\mathcal{D}_X$, and for each $x \in \mathcal{X}$, $\mathbb{P}[Y=1 \mid X=x] \in \{\beta, 1-\beta\}$. Thus the Bayes optimal error for each of these distributions is β.

Let $S \sim \mathcal{D}^m$. If one example in S is changed, $\nu(\frac{1}{2})$ changes by at most $1/m$. Hence, by McDiarmid's inequality (McDiarmid, 1989), with probability at least $1-\frac{1}{28}$, $\left|\nu(\frac{1}{2}) - \mathbb{E}[\nu(\frac{1}{2})]\right| \le \sqrt{\frac{\log(28)}{2m}}$. Denote the event that this holds E_M. Since $\beta = 8\alpha \ge \frac{\log(m)+\log(28)}{\sqrt{2m}}$, it follows that under E_M,

$$|\nu(1/2) - \mathbb{E}[\nu(1/2)]| \le \beta/8. \tag{25}$$

We now bound $\mathbb{E}[\nu(\frac{1}{2})]$. Using the notation $\hat{p}_x^+, n_x, n_x^+$ as in Lemma 20, we have

$$\nu(1/2) = \frac{1}{m}\sum_{x\in\mathcal{X}}\min(n_x^+, n_x - n_x^+) = \frac{1}{m}\sum_{x\in\mathcal{X}} n_x \min(p_x^+, 1-p_x^+)$$

Also, for all $p \in [0,1]$, $\min(p,1-p) \le 2p(1-p) \le 2\min(p,1-p)$. Therefore

$$\frac{1}{2m}\sum_{x\in\mathcal{X}} 2n_x p_x^+(1-p_x^+) \le \nu(1/2) \le \frac{1}{m}\sum_{x\in\mathcal{X}} 2n_x p_x^+(1-p_x^+).$$

By Lemma 20, it follows that

$$\frac{m-N}{m}\beta(1-\beta) \le \mathbb{E}[\nu(1/2)] \le 2\beta(1-\beta).$$

Since $N \le m/2$ and $\beta \in [0,\frac{1}{2}]$, $\mathbb{E}[\nu(\frac{1}{2})] \in (\beta/4, 2\beta)$. Combining this with Eq. (25), we get that under E_M, $\alpha = \beta/8 \le \nu(\frac{1}{2}) \le \frac{17}{8}\beta = 17\alpha$.

Now, we bound $G_{\min}$ from above and below assuming E_M holds. Denote

$$G(t) := \mathrm{GB}(\nu(t), \mathcal{N}(t), \delta, m, 1).$$

To establish a lower bound on $G_{\min}(m,\delta)$, note that $G_{\min}(m,\delta) = \min_{t>0} G(t) \ge \min_{t>0}\nu(t)$. For $t \in (0,\frac{1}{2})$, $\nu(t) = \nu(\frac{1}{2})$ (since the distances between any two distinct points in S is at least 1). In addition, since ν is monotonically increasing, we have $\nu(t) \ge \nu(\frac{1}{2})$ for $t \ge \frac{1}{2}$. Hence $\min_{t>0}\nu(t) \ge \nu(\frac{1}{2}) \ge \beta/8 = \alpha$.

To show an upper bound on $G_{\min}(m,\delta)$, we upper bound $G(\frac{1}{2})$. Note that $\mathcal{N}(\frac{1}{2}) \leq N$. Recall the definition of $\phi(t)$ in Eq. (6). We have

$$\phi(\frac{1}{2}) = \frac{(N+1)\log(m) + \log(\frac{1}{\delta})}{m} \leq \alpha.$$

Then, since $\nu(\frac{1}{2}) \leq 17\alpha$,

$$G(1/2) \leq \frac{m}{m-N}(\nu(\frac{1}{2}) + \frac{2}{3}\alpha + \frac{3}{\sqrt{2}}\sqrt{\nu(\frac{1}{2})\alpha}) \leq 30\alpha.$$

In the last inequality we used the fact that $\frac{m}{m-N} \leq 10/9$. So if E_M holds, $G_{\min}(m,\delta) \leq G(\frac{1}{2}) \leq 30\alpha$.

From the assumption on $\mathcal{A}$, with probability at least $1-\delta$, we have $\mathrm{err}(\hat{h},\mathcal{D}) \leq CG_{\min}(m,\delta) \leq 30C\alpha \leq 1/8$ (since $\alpha \leq \frac{1}{240C}$). Let $E_L(\mathcal{D})$ denote the event that $\mathcal{A}$ queries fewer than $N/2$ labels, where the probability is over the randomness of S and $\mathcal{A}$. Let h' be the output of an algorithm that behaves like $\mathcal{A}$ in cases where $E_L(\mathcal{D})$ holds, and queries at most $N/2$ otherwise. By Theorem 19, there exists some $\mathcal{D}$ in the family of distributions such that $\mathbb{E}[\mathrm{err}(h',\mathcal{D})] \geq \frac{1}{4}$. By Markov's inequality, $\mathbb{P}[\mathrm{err}(h',\mathcal{D}) \geq \frac{1}{8}] \geq 1/7$. Also, $\mathbb{P}[h' = \hat{h}] \geq \mathbb{P}[E_L(\mathcal{D})]$. Therefore

$$\mathbb{P}[\mathrm{err}(\hat{h},\mathcal{D}) \geq 1/8] \geq \mathbb{P}[\mathrm{err}(h',\mathcal{D}) \geq 1/8] + \mathbb{P}[E_L(\mathcal{D})] - 1 = \mathbb{P}[E_L(\mathcal{D})] - 6/7.$$

Therefore $\mathbb{P}[E_L(\mathcal{D})] - 6/7 \leq \mathbb{P}[\mathrm{err}(\hat{h},\mathcal{D}) \geq \frac{1}{8}] \leq \delta$. Since by assumption $\delta \leq 1/14$, it follows that $\mathbb{P}[E_L(\mathcal{D})] \leq 6/7 + \delta \leq 13/14$. It follows that with a probability of at least $1/14$, the negation of $E_L(\mathcal{D})$ holds. Since also E_M holds with probability at least $1 - \frac{1}{28}$, it follows that with a probability of at least $\frac{1}{28}$, both E_M and the negation of $E_L(\mathcal{D})$ hold. Now, as shown above, E_M implies the bounds on $G_{\min}(m,\delta)$ (item 1 in the theorem statement). In addition, the negation of $E_L(\mathcal{D})$ implies that $\mathcal{A}$ queries at least $N/2 = \frac{1}{2}\left\lfloor \frac{m\alpha - \log(\frac{m}{\delta})}{\log(m)} \right\rfloor$ labels (item 2 in the theorem statement). This completes the proof. ■

9. Discussion

We have presented an efficient fully empirical proximity-based non parametric active learner. Our approach provides competitive error guarantees for general distributions, in a general metric space, while keeping label complexity significantly lower than any passive learner with the same guarantees. MARMANN yields fully empirical error estimates, easily computable from finite samples. This is in contrast with classic techniques, that present bounds and rates that depend on unknown distribution-dependent quantities.

An interesting question is whether the guarantees can be related to the Bayes error of the distribution. Our error guarantees give a constant factor over the error guarantees of Gottlieb et al. (2017). A variant of this approach (Gottlieb et al., 2014a) was shown to be Bayes-consistent (Kontorovich and Weiss, 2015), and we conjecture that this holds also for the algorithm of Gottlieb et al. (2017). The passive component of our learning algorithm is indeed Bayes-consistent (Kontorovich et al., 2017). Since in our analysis MARMANN achieves a constant factor over the error of the passive learner, Bayes-consistency of the active learner cannot be inferred from our present techniques; we leave this problem open for future research.

Another important issue is one of efficient implementation. We mentioned that the naive $O(m^2)$ runtime for constructing a t-net may be improved to $2^{O(\mathrm{ddim}(\mathcal{X}))} m \log(1/t)$, as shown in Krauthgamer and Lee (2004); Gottlieb et al. (2014b). The fast t-net construction was the algorithmic work-horse of Gottlieb et al. (2014a,b, 2017) and inspired the passive component of our learner. We note that implementing even this passive component efficiently is far from trivial; this formed the core of a Master's thesis (Korsunsky, 2017). The remaining obstacle to making our algorithm fully practical is the magnitude of some of the constants. We believe these to be artifacts of the proof and intend to bring them down to manageable values in future work.

Acknowledgments

Acknowledgements Sivan Sabato was partially supported by the Israel Science Foundation (grant No. 555/15). Aryeh Kontorovich was partially supported by the Israel Science Foundation (grants No. 1141/12 and 755/15) and Yahoo Faculty and Paypal awards. We thank Lee-Ad Gottlieb and Dana Ron for the helpful discussions, and the referees for carefully reading the manuscript and their helpful suggestions.

References

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1999.

Jean-Yves Audibert. Fast learning rates in statistical inference through aggregation. *Ann. Statist.*, 37(4):1591–1646, 08 2009.

Pranjal Awasthi, Maria-Florina Balcan, and Philip M. Long. The power of localization for efficiently learning linear separators with noise. In *Symposium on Theory of Computing, STOC 2014*, pages 449–458, 2014.

Maria-Florina Balcan, Andrei Z. Broder, and Tong Zhang. Margin based active learning. In *Proceedings of the 20th Annual Conference on Learning Theory, COLT 2007*, pages 35–50, 2007.

Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. *J. Comput. Syst. Sci.*, 75(1), 2009.

Maria-Florina Balcan, Steve Hanneke, and Jennifer Wortman Vaughan. The true sample complexity of active learning. *Machine Learning*, 80(2-3):111–139, 2010.

Daniel Berend and Aryeh Kontorovich. A finite sample analysis of the naive bayes classifier. *Journal of Machine Learning Research*, 16:1519–1545, 2015.

Christopher Berlind and Ruth Urner. Active nearest neighbors in changing environments. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, pages 1870–1879, 2015.

Oren Boiman, Eli Shechtman, and Michal Irani. In defense of nearest-neighbor based image classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2008*, 2008.

Nader H. Bshouty, Yi Li, and Philip M. Long. Using the doubling dimension to analyze the generalization of learning algorithms. *Journal of Computer and System Sciences*, 75(6):323 – 335, 2009.

Rui M Castro and Robert D Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008.

Rui M. Castro, Rebecca Willett, and Robert D. Nowak. Faster rates in regression via active learning. In *Advances in Neural Information Processing Systems, NIPS 2005*, pages 179–186, 2005.

Nicolò Cesa-Bianchi, Claudio Gentile, Fabio Vitale, and Giovanni Zappella. Active learning on trees and graphs. In *Proceedings of rhe 23rd Conference on Learning Theory, COLT 2010*, pages 320–332, 2010.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems, NIPS 2014*, pages 3437–3445, 2014.

Thomas M. Cover and Peter E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

Gautam Dasarathy, Robert D. Nowak, and Xiaojin Zhu. S2: an efficient graph based active learning algorithm with application to nonparametric classification. In *Proceedings of the 28th Annual Conference on Learning Theory, COLT 2015*, pages 503–522, 2015.

Sanjoy Dasgupta. Analysis of a greedy active learning strategy. In *Advances in Neural Information Processing Systems, NIPS 2004*, pages 337–344, 2004.

Sanjoy Dasgupta. Consistency of nearest neighbor classification under selective sampling. In *Proceedings of the 25th Annual Conference on Learning Theory, COLT 2012*, pages 18.1–18.15, 2012.

Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th International Conference on Machine Learning, ICML 2008*, pages 208–215, 2008.

Luc Devroye and László Györfi. *Nonparametric density estimation: the L_1 view*. Wiley Series in Probability and Mathematical Statistics: Tracts on Probability and Statistics. John Wiley & Sons, Inc., New York, 1985.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

Andreas Emil Feldmann, Wai Shing Fung, Jochen Könemann, and Ian Post. A $(1 + \epsilon)$-embedding of low highway dimension graphs into bounded treewidth graphs. *CoRR*, abs/1502.04588, 2015. URL `http://arxiv.org/abs/1502.04588`.

Evelyn Fix and Jr. Hodges, J. L. Report Number 4, Project Number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

Evelyn Fix and Jr. Hodges, J. L. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review / Revue Internationale de Statistique*, 57(3):pp. 238–247, 1989.

Sally Floyd and Manfred K. Warmuth. Sample compression, learnability, and the vapnik-chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995.

Alon Gonen, Sivan Sabato, and Shai Shalev-Shwartz. Efficient active learning of halfspaces: an aggressive approach (extended abstract ICML 2013). *Journal of Machine Learning Research*, 14 (1):2583–2615, 2013.

Lee-Ad Gottlieb and Robert Krauthgamer. Proximity algorithms for nearly doubling spaces. *SIAM J. Discrete Math.*, 27(4):1759–1769, 2013.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient classification for metric data (extended abstract COLT 2010). *IEEE Transactions on Information Theory*, 60(9):5750–5759, 2014a.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Near-optimal sample compression for nearest neighbors. In *Advances in Neural Information Processing Systems, NIPS 2014*, pages 370–378, 2014b.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Adaptive metric dimensionality reduction (extended abstract ALT 2013). *Theoretical Computer Science*, pages 105–118, 2016.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Nearly optimal classification for semimetrics (extended abstract AISTATS 2016). *Journal of Machine Learning Research*, 2017.

Thore Graepel, Ralf Herbrich, and John Shawe-Taylor. Pac-bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59(1-2):55–76, 2005.

Steve Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011.

Steve Hanneke and Liu Yang. Minimax analysis of active learning. *Journal of Machine Learning Research*, 16:3487–3602, 2015.

Aryeh Kontorovich and Iosif Pinelis. Exact lower bounds for the agnostic probably-approximately-correct (PAC) machine learning model. *CoRR*, abs/1606.08920, 2016.

Aryeh Kontorovich and Roi Weiss. Maximum margin multiclass nearest neighbors. In *Proceedings of the 31st International Conference on Machine Learning, ICML 2014*, pages 892–900, 2014.

Aryeh Kontorovich and Roi Weiss. A bayes consistent 1-nn classifier. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015*, 2015.

Aryeh Kontorovich, Sivan Sabato, and Roi Weiss. Nearest-neighbor sample compression: Efficiency, consistency, infinite dimensions. In *Advances in Neural Information Processing Systems, NIPS 2017*, pages 1572–1582, 2017.

Yevgeni Korsunsky. Classifying processes by their system call traces using 1-nearest-neighbor with compression. Master's thesis, Ben-Gurion University Computer Science Department, 2017. URL `https://tinyurl.com/Korsunsky-msc`.

Samory Kpotufe. k-nn regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems, NIPS 2011*, pages 729–737, 2011.

Samory Kpotufe, Ruth Urner, and Shai Ben-David. Hierarchical label queries with data-dependent partitions. In *Proceedings of the 28th Annual Conference on Learning Theory, COLT 2015*, pages 1176–1189, 2015.

Robert Krauthgamer and James R. Lee. Navigating nets: Simple algorithms for proximity search. In *15th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 791–801, January 2004.

Sanjeev R. Kulkarni and S. E. Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41(4):1028–1039, 1995.

Nick Littlestone and Manfred K. Warmuth. Relating data compression and learnability. unpublished, 1986.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample-variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory, COLT 2009*, 2009.

Andrew McCallum and Kamal Nigam. Employing EM and pool-based active learning for text classification. In *Proceedings of the 15th International Conference on Machine Learning, ICML 1998*, pages 350–358, 1998.

Colin McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics, volume 141 of LMS Lecture Notes Series*, pages 148–188. Morgan Kaufmann Publishers, San Mateo, CA, 1989.

Sivan Sabato and Rémi Munos. Active regression by stratification. In *Advances in Neural Information Processing Systems, NIPS 2014*, pages 469–477, 2014.

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

John Shawe-Taylor, Peter L. Bartlett, Robert C. Williamson, and Martin Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5): 1926–1940, 1998.

Charles J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–620, 1977.

Ruth Urner, Sharon Wulff, and Shai Ben-David. PLAL: cluster-based active learning. In *Proceedings of the 26th Annual Conference on Learning Theory, COLT 2013*, pages 376–397, 2013.

Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with Lipschitz functions. *Journal of Machine Learning Research*, 5:669–695, 2004.

Kai Wei, Rishabh K. Iyer, and Jeff A. Bilmes. Submodularity in data subset selection and active learning. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, pages 1954–1963, 2015.

Lin Cheng Zhao. Exponential bounds of mean error for the nearest neighbor estimates of regression functions. *J. Multivariate Anal.*, 21(1):168–178, 1987.

Xiaojin Zhu, John Lafferty, and Zoubin Ghahramani. Combining active learning and semi-supervised learning using gaussian fields and harmonic functions. In *ICML 2003 workshop*, pages 58–65, 2003.