JMLR: Workshop and Conference Proceedings vol 40:1–14, 2015

Hierarchical Label Queries with Data-Dependent Partitions

Samory Kpotufe SAMORY@PRINCETON.EDU
Princeton University, Operations Research and Financial Engineering

Ruth Urner RURNER@TUEBINGEN.MPG.DE
MPI for Intelligent Systems, Empirical Inference Department, Tübingen

Shai Ben-David SHAI@CS.UWATERLOO.CA
University of Waterloo, Cheriton School of Computer Science

Abstract

Given a joint distribution $P_{X,Y}$ over a space $\mathcal{X}$ and a label set $\mathcal{Y} = \{0, 1\}$, we consider the problem of recovering the labels of an unlabeled sample with as few label queries as possible. The recovered labels can be passed to a passive learner, thus turning the procedure into an active learning approach.

We analyze a family of labeling procedures based on a hierarchical clustering of the data. While such labeling procedures have been studied in the past, we provide a new parametrization of $P_{X,Y}$ that captures their behavior in general low-noise settings, and which accounts for data-dependent clustering, thus providing new theoretical underpinning to practically used tools.

1. Introduction

Given a joint distribution over a space $\mathcal{X}$ and a label set $\mathcal{Y} = \{0, 1\}$, we consider the problem of recovering the labels of an unlabeled sample $X_{1:n} \triangleq \{X_i\}_1^n$ with as few label queries as possible, under a relaxed version of the *cluster-assumption*. Recovered labels can then be passed onto a supervised learner.

The cluster-assumption, generally stated, reads as follows: the data $X_{1:n}$, or the underlying data space $\mathcal{X}$, can be partitioned into clusters $\{C_i\}$ such that points in each C_i, are very likely to have the same label. If one knew the clusters $\{C_i\}$, then clearly the data $X_{1:n}$ could be cheaply labeled by asking for a few labels from each C_i to determine its dominating label. We refer the reader to an early rigorous work on the subject, Rigollet (2007), which analyzes possible gains in a setting where clusters are defined as high density regions of μ.

In practice however, clusters might not be apparent in data (e.g., μ is uniform on $\mathcal{X}$), or apparent clusters might not correspond to label boundaries. In the latter case, a procedure based on the cluster-assumption will mislabel the data, hence the assumption can result in worse prediction, which is hard to detect in such a setting with little label information.

A natural idea, recently analyzed in Dasgupta and Hsu (2008); Dasgupta (2011); Urner et al. (2013) , is to recursively *test* the cluster-assumption over a hierarchical partitioning of the data:

> Start by partitioning $X_{1:n}$ into large regions (or clusters), label these if they seem pure after a few label queries, otherwise refine the partition into smaller clusters and repeat.

This hierarchical approach relaxes the need for apparent clusters: few labels are requested if sufficiently large regions C of $\mathcal{X}$ admit a low-error label $Y(C)$, without these regions being known

a-priori. Furthermore, the procedure can be made *safe* by properly testing purity: if the cluster-assumption does not hold, i.e. large regions are highly unpure, it simply requests labels for most points in $X_{1:n}$ rather than mislabeling them.

The present work extends Dasgupta and Hsu (2008); Dasgupta (2011); Urner et al. (2013) and further elucidates the conditions guaranteeing the success of the above procedure. In particular, the guarantees of Dasgupta and Hsu (2008); Dasgupta (2011) are conditioned on the niceness of the sample but do not characterize the distributions $P_{X,Y}$ for which such nice samples are likely to arise; the subsequent work by Urner et al. (2013) derives the first distributional conditions guaranteeing low label complexity, but disallows partitioning procedures built using the data. We prove guarantees under more general conditions on $P_{X,Y}$, while allowing more practical families of hierarchical partitioning procedures (e.g. k-d trees).

Background and Overview of Results

The cluster-assumption is motivated by situations where labels are expensive to obtain. This is the same motivation as in active learning, and in fact the approach considered here can be viewed as an alternative to common active learners which operate by requesting labels in regions that increase confidence in a choice of hypothesis out of a class $\mathcal{H}$ (see e.g. Balcan et al. (2006); Dasgupta et al. (2007); Balcan et al. (2008); Hanneke (2009); Dasgupta (2011)). A practical advantage of this approach, which perhaps explains its seemingly high appeal with practitioners, is its ease of implementation. It has so far however received less theoretical attention.

Once the unlabeled data $X_{1:n}$ is fully labeled, say within error ϵ w.r.t. a fixed unknown labeling $Y_{1:n}$, it can be passed onto a passive learner, e.g., an empirical risk minimizer over some $\mathcal{H}$, which then returns a hypothesis with excess risk $O(\epsilon)$, assuming a sample size $n = \Omega\left(VC(\mathcal{H})/\epsilon^2\right)$. The goal of the labeler is to request less than the number of labels required in passive learning: it can be shown that, if the Bayes classifier is not in $\mathcal{H}$, even in low or no noise situations (i.e. $\mathbb{E}[Y|x]$ is close to 0 or 1), passive learning requires $\Omega\left(VC(\mathcal{H})/\epsilon^2\right)$ labels in order to return an $h \in \mathcal{H}$ with excess error $O(\epsilon)$ (Urner et al., 2013; Ben-David and Urner, 2014).

The procedure considered here was first properly formalized in Dasgupta and Hsu (2008); Dasgupta (2011). They show that, given data $X_{1:n}$ (with a fixed unknown labeling $Y_{1:n}$) and a hierarchical partitioning T of $X_{1:n}$, if most clusters (leaves) of some subtree T' are nearly pure (i.e. the minority label in a cluster occurs with proportion $o(\epsilon)$), then such a procedure guarantees a labeling error $O(\epsilon)$ w.r.t. $Y_{1:n}$, after $O(|T'|/\epsilon)$ label requests.

The subsequent work of Urner et al. (2013), rather than conditioning on the purity of the sample $X_{1:n}$ at hand, derives sufficient conditions on the distribution $P_{X,Y}$ that ensures similar (expected) label complexity. More precisely, in a deterministic setting ($\mathbb{E}[Y|x]$ is 0 or 1), they derive a nice parametrization $\Lambda(r)$ of $P_{X,Y}$, $\Lambda \in [0,1]$, called *Probabilistic Lipschitzness*, which encodes how likely Y is to change over regions of $\mathcal{X}$ of diameter r. They show that, given a partition T independent of the data, such a procedure guarantees a labeling error $O(\epsilon)$ while querying $O\left(\inf_r \quad |T_r|/\epsilon + \Lambda(r)/\epsilon^2\right)$ labels, T_r being the level of T with clusters of diameter less than r. The *niceness* parameter $\Lambda(r)$ is understood to decrease as $r \to 0$, while $|T_r|$ increases as $r \to 0$. Therefore fewer labels are requested when $\Lambda(r)$ decreases fast with r (i.e. the cluster-assumption holds over large unknown regions of $\mathcal{X}$), and $|T_r|$ slowly increases as $r \to 0$, i.e., the partitioning procedure T refines the data $X_{1:n}$ into small regions using small size partitions.

This dependence on how fast the partitioning tree T grows is intrinsic to the labeling approach and is illustrated in Figure 1. Unfortunately trees that are built independent of $X_{1:n}$ (e.g. dyadic partitions of $\mathcal{X}$) might yield unnecessarily large partitions. In fact it is common in practice to repeatedly cluster the data till one finds a tree that yields good data-quantization. A main motivation for the present paper is therefore to allow practical data-dependent trees, e.g., (randomized) k-d trees, RP trees, which are more likely to remain small by better fitting the data. However, allowing T to depend on $X_{1:n}$ introduces additional interdependencies between iterative steps of the hierarchical labeling procedure. This raises questions on the stability of such an instantiation w.r.t. to the random sample, and introduces new technical difficulties in the analysis.

Furthermore, while the cluster-assumption is inherently a low or no noise assumption (i.e. it assumes one label has low error), it is conceivable that some amount of noise might be tolerable by a labeling procedure. In particular, as in Dasgupta and Hsu (2008), if most clusters admit a labeling with error $o(\epsilon)$ for some target ϵ, then label savings should be possible, whether labels are deterministic or not. We therefore consider a general parametrization of $P_{X,Y}$ which allows for nondeterministic labels, i.e., allows for $\mathbb{E}[Y|x] \in [0,1]$.

We adopt a simple parametrization Γ that directly encodes how likely it is, under $P_{X,Y}$, that a labeling of a large region C of $\mathcal{X}$ has low error (equivalently $\mathbb{E}[Y|C]$ has margin away from $1/2$). Γ also captures how well a hierarchical procedure aligns with good regions of $P_{X,Y}$ in terms of the clusters it might produce on a random sample. Our bound on labels requested takes the form

$$O\left(\inf_r \quad |T_r| \cdot n \cdot \epsilon + n \cdot \Gamma(\tau_\epsilon, r)\right),$$

for an unlabeled sample size $n = \Omega(1/\epsilon^2)$, and a noise margin $\tau_\epsilon = 1/2 - O(\epsilon)$. $|T_r|$ is the data-dependent size of a partition fitted from the sample $X_{1:n}$. Γ is smallest for partitioning procedures likely to produce clusters well aligned with good regions of $\mathcal{X}$. However, it can be shown to be small independent of the partitioning procedure, under sufficient conditions on just $P_{X,Y}$. In our nondeterministic setting, we need two complementary such conditions. First, just as in Urner et al. (2013), we use Probabilistic Lipschiztness to capture how likely it is that $\mathbb{E}[Y|x] - 1/2$ changes sign over large regions of $\mathcal{X}$. However, this is no longer enough since $\mathbb{E}[Y|x] - 1/2$ can keep the same sign over some $C \subset \mathcal{X}$, but remain close to 0, in which case any labeling of C will have high error above our target ϵ. Therefore we also need to parametrize the level of noise in Y, i.e. how likely it is for $\mathbb{E}[Y|x]$ to be close to 0 or 1. We can then show that, under Probabilistic Lipschitzness and low-noise Tsybakov's conditions w.r.t. to ϵ (essentially large regions admit a labeling with error $o(\epsilon)$), the above bound on label-complexity recovers the form of the results of Urner et al. (2013). In particular, for a fixed ϵ, the results show a labeling complexity in the continuum between $O(C_T/\epsilon)$ and $O(1/\epsilon^2)$, as Lipschiztness decreases and noise level increases. Here C_T can be viewed as the size of the smallest partition of the tree T (first level).

Finally, to mitigate algorithmic stability issues arising from dependencies on data, we require that the partitioning procedure T be unlikely to produce clusters of high complexity. Relevant subsets of $\mathcal{X}$ can then be shown to have low-complexity allowing for the stability of labels over data-dependent clusters. This allows for high probability bounds on label requests, by properly conditioning on relevant events that decouple the interdependencies between labeling steps.

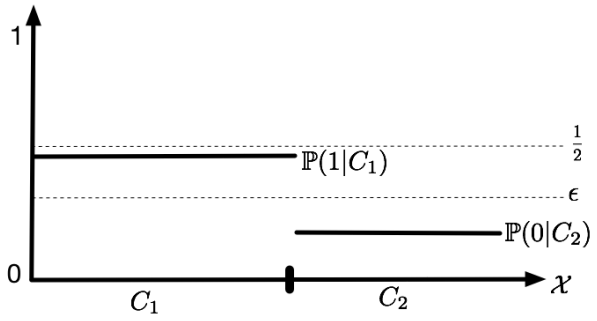

Figure 1: Two clusters C_1, C_2. Label 0 dominates in C_1 while label 1 dominates in C_2. The error of the best label is shown for both. For C_1, $\mathbb{P}(1|C_1)$ is close to $1/2$ far above ϵ, so there is no labeling of C_1 with error less than ϵ. A safe labeling procedure must request all labels for points in C_1 (or refine C_1). However, C_2 can be labeled 1 with error below ϵ. Detecting this requires $O(1/\epsilon)$ label requests. Thus, for total error $\sum_i Pr(C_i)\text{err}(Y(C_i)) \leq \epsilon$ on a partition $\{C_i\}$, a labeling procedure ends up requesting $O(1/\epsilon)$ labels on most clusters of the partition, hence the need for small partition sizes.

2. Preliminaries

2.1. Admissible Hierarchical Partitioning Procedures

We will consider the following type of hierarchical clustering of data into a tree T. The tree T is allowed to be randomized (e.g. randomized k-d trees). We let ρ denote this randomness.

Definition 1 *Given an unlabeled sample $X_{1:n} = \{X_i\}_{i=1}^n$, and random bits ρ, a* **hierarchical-clustering procedure** *is a function $T : (\rho; X_{1:n}) \mapsto T(\rho; X_{1:n}) \triangleq \{T_l(\rho; X_{1:n})\}_{l\in\mathbb{N}}$ mapping $(\rho; X_{1:n})$ to a hierarchical collection of partitions (clusterings) of the data. Formally, level l of such a tree, $T_l(\rho; X_{1:n})$, is a collection of disjoint subsets C of $\mathcal{X}$ such that*
• *For every cluster C, $C \cap X_{1:n} \neq \emptyset$, and diam$(C) \triangleq \sup_{x,x'\in C} \|x - x'\| \leq 2^{-l}$,*
• $X_{1:n} \subseteq \bigcup_{C\in T_l(\rho;X_{1:n})} C$.
• *$\forall l > 0$, every $C \in T_l(\rho; X_{1:n})$ has a* parent *$C' \in T_{l-1}(\rho; X_{1:n})$, s.t. $(C \cap X_{1:n}) \subset (C' \cap X_{1:n})$.*
We sometimes drop the parameters ρ and $X_{1:n}$ when these are understood from context.

The requirement that the diameters of the clusters decrease may appear to exclude trees such as randomized k-d trees where only the diameter of the data might decrease. Such are easily converted to satisfy the Definition 1 (Section 5.2). The *statistical complexity* of a tree T, is captured by both the number of clusters produced at each level of the tree, and the complexity of the cluster shapes.

Definition 2 *The hierarchical-clustering $T(\rho; X_{1:n})$ has* **tree-growth rate** *$\kappa \geq 1$ on the sample $X_{1:n}$ if, $\forall l \in \mathbb{N}$, $\quad |T_l(\rho; X_{1:n})| \leq 2^{\kappa l}$.*

Note that since $X_{1:n}$ contains at most n points, for every such tree, from some level l on, there are at most n clusters in each level and each cluster contains a single point of $X_{1:n}$. Next, we characterize the complexity of the cluster shapes. Intuitively, clusters of *simple* shape yield better generalization.

Definition 3 (VC of clusters) *Consider a hierarchical-clustering procedure T. For given random bits ρ, let $\mathcal{C}_T(\rho) \triangleq \bigcup_{X_{1:n} \in \mathcal{X}^n} \{C \in T_l(\rho; X_{1:n}) : l \in \mathbb{N}\}$ denote the class of all possible clusters of $T(\rho; \cdot)$. Let VC$(\mathcal{C}_T(\rho))$ denote the Vapnik-Chervonenkis dimension of this class. Given $0 < \delta < 1$, the procedure T has* **VC-cluster dimension** *$V_{T,\delta}$ if $\mathbb{P}_\rho\left(VC(\mathcal{C}_T(\rho)) \leq V_{T,\delta}\right) \geq 1 - \delta$.*

2.2. Distributional properties

We consider a binary classification problem where the r.v.s $X \in \mathcal{X}$, and $Y \in \{0, 1\}$, are jointly distributed. The space $\mathcal{X}$ has diameter 1, i.e. $\sup_{x,x' \in \mathcal{X}} \|x - x'\| = 1$. We denote the marginal measure on $\mathcal{X}$ by μ, and $\mathcal{X}$ can be viewed as the support of μ. The following quantities measure the noise in Y: $\eta(x) \triangleq \mathbb{P}(1|x)$ and for any $C \subset \mathcal{X}$, $\eta_C \triangleq \mathbb{P}(1|x \in C)$.

The behavior of the algorithm will depend on how well labels cluster under the given hierarchical procedure. Thus, we want to capture the *clusterability* of labels in terms of properties on the unknown distribution $P_{X,Y}$ and the choice of clustering procedures. While it might be unlikely that labels cluster well on large clusters, it is reasonable to expect that, for most procedures T, labels might cluster well in sufficiently small clusters. For instance, suppose the Bayes decision boundary is sufficiently smooth, then in most small clusters, independent of shape, one label will dominate. We use the following parametrization to capture the above ideas.

Definition 4 (Clusterability of labels) *Let $\tau > 0$ and $0 < r \leq 1$. Fix a hierarchical clustering procedure T. For any point $x \in \mathcal{X}$, let $\mathcal{C}_{T,r}(x)$ denote the set of clusters $C \in \cup_\rho \mathcal{C}_T(\rho)$ containing x and of diameter $\sup_{x,x' \in C} \|x - x'\|$ at most r. Set $\Gamma(\tau, r) \triangleq \mathbb{P}_X\left(\exists C \in \mathcal{C}_{T,r}(X),\ |\eta_C - 1/2| < \tau\right)$. Notice that the functions $\Gamma(\tau, \cdot)$ and $\Gamma(\cdot, r)$ are nondecreasing, respectively for τ and r fixed.*

Thus the function $\Gamma(\cdot, \cdot)$ (that exists whenever the sets in $\cup_\rho \mathcal{C}_T(\rho)$ are measurable), parametrizes the pairing $P_{X,Y}, T$. We want $\Gamma(\tau, \cdot)$ and $\Gamma(\cdot, r)$ to decrease fast as $r \to 0$ or $\tau \to 0$. If $P_{X,Y}$ is sufficiently benign, then Γ decreases fast for admissible tree-procedures T. Lemma 10 shows that the rate of decrease can be upper bounded under low noise and Lipschitzness type conditions on η.

3. Labeling Procedure

Algorithm 1 defines a family of labeling procedures indexed by the particular hierarchical clustering T instantiated. It proceeds level by level by refining each cluster if it is deemed unpure. To detect purity of a cluster, it requests $O(1/\epsilon)$ labels from the sample points in the cluster and declares it pure if the minority label is at most $O(\epsilon)$. For robustness, it only labels clusters containing at least $n \cdot \epsilon$ points, where n itself needs to be sufficiently large w.r.t. tree complexity $V_{T,\delta}$ (see Theorem 7).

4. Results

4.1. Error bound

Definition 5 (Underlying labeling)

- *Given an unlabeled sequence $X_{1:n}$, $Y_{1:n} \doteq \{Y_i\}_1^n$ denotes a labeling of $X_{1:n}$ (drawn by $P_{Y|X}$; so given X_i, Y_i is a Bernoulli random variable).*
- *The labeling error of a sequence of labels ${Y_{1:n}}' \doteq \{Y_i'\}_1^n$ on $Y_{1:n}$ is the average disagreement between $Y_{1:n}$ and ${Y_{1:n}}'$. Namely, $\frac{1}{n}\sum_i \mathbb{1}_{\{Y_i \neq Y_i'\}}$.*

Algorithm 1 Labeler $(X_{1:n}, T(\rho; X_{1:n}), \epsilon, \delta)$

Initialize: $l = 0$, Active cluster set $\mathcal{C}^l = T_0(\rho; X_{1:n})$
for $l = 0, 1, 2, \ldots, \infty$ **do**
  $\delta_l \leftarrow \delta / \left|\mathcal{C}^l\right| 2^{l+1}$
  $n_l(\epsilon) \leftarrow 9 \ln(8/\delta_l)/\epsilon$
  for each $C \in \mathcal{C}^l$ **do**
    if $\mu_n(C) < \epsilon$ **then**
      Request all labels for points in $C \cap X_{1:n}$, and skip to the next cluster in $\mathcal{C}^l$
    $S \equiv$ labeled sample from $C \cap X_{1:n}$ (with replacement) of size $n_l(\epsilon)$
    // At any time in the procedure, the same label is returned each time the same X is sampled
    $\hat{\eta}_S \leftarrow$ probablity of label 1 over S
    if $\min\{\hat{\eta}_S, 1 - \hat{\eta}_S\} \leq \epsilon/3$ **then**
      Label all $C \cap X_{1:n}$ with the majority label from S
    else
      Add children of C to $\mathcal{C}^{l+1}$.
  if all of $X_{1:n}$ is labeled **then**
    return labeled sample $\mathcal{X}$

The number of errors that the labeler makes is measured with respect to the above underlying labeling (Definition 5). We show that with high probability Algorithm 1 makes few label errors. The proof is in the appendix (long version)

Theorem 6 (Labeling error) *Let $0 < \epsilon, \delta < 1$. Suppose $Y_{1:n} \doteq \{Y_i\}_1^n$ is defined as in Definition 5. Let ${Y_{1:n}}' \doteq \{Y_i'\}_1^n$ be the labeling obtained by Algorithm 1. With probability at least $1 - \delta$, the labeling error of the algorithm, i.e. $\frac{1}{n}\sum_i \mathbb{1}_{\{Y_i \neq Y_i'\}}$, is at most ϵ.*

Since the labeling returned by the procedure has error at most ϵ with respect to i.i.d labels $Y_{1:n}$, the labeled sample $(X_{1:n}, {Y_{1:n}}')$ can therefore be used as input to a noise tolerant supervised learner, whose generalization error is then worsened by at most $c\epsilon$ for some small constant c. Consider for instance the case of ERM over a hypothesis class $\mathcal{H}$. Given $\Omega(VC(\mathcal{H})/\epsilon^2)$ unlabeled samples, we will then return a hypothesis with error at most $\inf_{h\in\mathcal{H}} \mathrm{err}(h) + 2\epsilon$ while requesting a number of labels potentially much less than $O(VC(\mathcal{H})/\epsilon^2)$.

4.2. Label query bound

We now present and discuss our main result, namely bounding the number of label queries that our procedure will make. The main idea behind the label-complexity analysis is to consider the number of labels requested at any given level and bound the numbers of labels yet to request. This idea is borrowed from Urner et al. (2013). However, the analysis in Urner et al. (2013) only handles the case where labels are deterministic and provides bounds in expectation. It can therefore avoid the various inter-dependencies introduced by the labeling decisions and the fact that the clusters C can depend on the data $X_{1:n}$.

Here we give a high probability result where the main technicality is in handling the various interdependencies. This is done by properly conditioning on key events and bounding the VC-dimension of some important subsets of the support $\mathcal{X}$. The theorem is proved in Section 6.

Theorem 7 *Let $0 < \epsilon, \delta < 1/2$. The following holds with probability at least $1 - 8\delta$. Suppose the hierarchical-clustering $T(\rho; X_{1:n})$ has tree-growth rate κ on $X_{1:n}$. Let $\alpha_n = (\ln 2n + \ln(8/\delta))/n$, and assume $n \geq 81\,(16 V_{T,\delta} \ln 2n + \ln(8/\delta))/\epsilon^2$. Let $\tau_\epsilon \triangleq 1/2 - \epsilon/162$. The number of labels*

requested by Algorithm 1 is at most

$$\inf_{l\in\mathbb{N}} \quad 2^{\kappa l}(2\kappa l)\cdot n\cdot\epsilon + n\cdot\left(\Gamma\left(\tau_\epsilon, 2^{-l}\right) + \sqrt{\Gamma\left(\tau_\epsilon, 2^{-l}\right)\cdot\alpha_n} + \alpha_n\right).$$

The bound in Theorem 7 depends on the complexity of the clustering procedure T, as captured by $V_{T,\delta}$, and the size of the clustering produced on the data $X_{1:n}$, as captured by κ. The infimum is taken over all levels of l of the tree, so the bound is best when Γ decreases quickly in l. Γ expectedly depends on ϵ since the inherent assumption in hierarchical labeling is that there exists regions C that admit a label $Y(C)$ of error $o(\epsilon)$. In the worst case when Γ is large, i.e., the relaxed cluster assumption fails to hold, the procedure never queries more than n points, so is safe.

To best understand the above bound, we instantiate it in the next section under more common noise parametrizations, namely Tsybakov's low noise conditions and Probabilistic Lipschitzness. This is followed by a discussion of values of κ that might be expected in practice with data-dependent trees such as RP-tree and other randomized k-d trees.

5. Instantiation of results

5.1. Clusterability of labels under common data assumptions

The following definition captures the distribution of the noise margin $|\eta - 1/2|$.

Definition 8 (Noise level) *For $\tau > 0$, define $\Delta(\tau) \doteq \mathbb{P}_X\left(|\eta(X) - 1/2| < \tau\right)$. Note that Δ is nondecreasing.*

The *Tsybakov low-noise condition* is a probabilistic relaxation of the stronger condition that $\Delta(\tau) = 0$ for all τ less than some τ_0 (Massart and Nedelec (2006)). It simply sates that $\Delta(\tau) \le c\tau^\beta$ for some $c, \beta > 0$ (Mammen and Tsybakov (1999); Tsybakov (2004)). The Tsybakov condition becomes stronger for large values of β.

The next definition relaxes Lipschitzness. Recall that a function η is L-Lipschitz if, for all $x, x' \in \mathcal{X}$, $|\eta(x) - \eta(x')| \le L\cdot\|x - x'\|$.

Definition 9 (Probabilistic Lipschitzness (PL) on η) *For $\lambda > 0$ define $\Lambda(\lambda) \doteq \mathbb{P}_X\left(\mathbb{P}_{X'}\left(|\eta(X) - \eta(X')| > \frac{1}{\lambda}\|X - X'\|\right) > 0\right)$. Note that Λ is nondecreasing.*

In Urner et al. (2013) label complexity reductions of active learning for VC-classes were shown for the above condition with $\Lambda(\lambda) \le \lambda^\alpha$, $\alpha > 0$, (under the further assumption of deterministic Y). This PL condition gets stronger with larger values of α.

The next lemma shows how the parametrization Γ for the cluster-assumption can be bounded in terms of the above two noise parameters.

Lemma 10 (Γ in terms of Δ and Λ) *Let $\tau, r > 0$. We have $\Gamma(\tau, r) \le \inf_{\lambda>0}\Delta(\tau + r/\lambda) + \Lambda(\lambda)$.*

Proof Let $\tau, r, \lambda > 0$. Consider $x \in \mathcal{X}$ such that (1) $|\eta(x) - 1/2| > \tau$, and (2) $\forall^\mu x' \in \mathcal{X}, |\eta(x) - \eta(x')| \le 1/\lambda\,\|x - x'\|$.

We then have that $\forall^\mu x' \in \mathcal{X}$ satisfying $\|x - x'\| \le r$, $|\eta(x) - \eta(x')| \le r/\lambda$, and hence $|\eta(x') - 1/2| \ge \tau - r/\lambda$. Therefore, any cluster C of T of diameter at most r containing x must have $|\eta_C - 1/2| \ge \tau - r/\lambda$. Thus, $\forall\tau, r > 0$, we have $\Gamma(\tau, r) \le \inf_{\lambda>0}\Delta(\tau + r/\lambda) + \Lambda(\lambda)$. ■

Theorem 11 (Label complexity under low-noise and P-Lipschitzness conditions) *Suppose the noise in Y satisfies, for any $\tau, \lambda > 0$,*
*(*i) $\Delta(\tau) \leq c\tau^{\beta}$ for some $c, \beta > 0$, and (ii) $\Lambda(\lambda) \leq \lambda^{\alpha}$ for some $\alpha > 0$.

There exists $C, C' > 0$ such that the folowing holds. Let $0 < \epsilon, \delta < 1/2$ and suppose $n \geq C \cdot (1/\epsilon^2) \cdot (V_{T,\delta} \log(1/\epsilon) + \log(1/\delta))$. Then, with probability at least $1 - 8\delta$, the number of labels requested by Algorithm 1 is at most

$$C' \left(2^{\kappa\alpha/(\kappa+\alpha)} \cdot \epsilon^{\alpha/(\kappa+\alpha)} \cdot \ln(1/\epsilon) + e^{-\epsilon\beta/162}\right) \cdot n.$$

Proof Pick $0 < \lambda \leq 2$. By assumption on Δ and Λ, and by Lemma 10, for any level $l \geq \log(2/\lambda)$,

$$\Gamma(\tau_\epsilon, 2^{-l}) \leq c\left(1/2 - \epsilon/162 + 2^{-l}/\lambda\right)^{\beta} + \lambda^{\alpha} \leq c(1 - \epsilon/162)^{\beta} + \lambda^{\alpha} \leq ce^{-\epsilon\beta/162} + \lambda^{\alpha}.$$

Therefore fix such an $l = \lceil \log(2/\lambda) \rceil$. First pick C such that the assumption of the theorem on n also satisfies that of Theorem 7. Using the fact that for $a, b \geq 0$, $a + \sqrt{ab} + b \leq 2(a + b)$, we have by Theorem 7 that, with probability at least $1 - 8\delta$, the labeling requirement is at most (for some universal $C_1, C_2 > 0$)

$$\begin{aligned} &C_1(2/\lambda)^{\kappa}(2\kappa \log(2/\lambda)) \cdot n \cdot \epsilon + 2cn \cdot e^{-\epsilon\beta/162} + 2n \cdot \lambda^{\alpha} + 2n \cdot \alpha_n \\ &\leq C_2 \left((2/\lambda)^{\kappa} \cdot \kappa \log(2/\lambda) \cdot n \cdot \epsilon + n \cdot \lambda^{\alpha} + n \cdot e^{-\epsilon\beta/162}\right), \end{aligned}$$

provided C is large enough so that $\alpha_n \leq \epsilon$. To finish, set $\lambda = 2^{\kappa/(\kappa+\alpha)} \cdot \epsilon^{1/(\kappa+\alpha)}$. ■

The first term of the above bound (Theorem 11) on label request depends just on α and recovers the bounds of Urner et al. (2013) in a deterministic setting. This term dominates for sufficient low-noise level w.r.t. ϵ, i.e., for $\beta \geq \Omega(1/\epsilon)$. This corresponds to situations where the Bayes classifier $\mathbb{1}_{\{\eta(x) > 1/2\}}$ has error less than ϵ on much of $\mathcal{X}$, i.e., often achieves margin of order $1/2 - \epsilon$.

For any fixed ϵ, and $n = \widetilde{O}(1/\epsilon^2)$, the label complexity ranges in the continuum between $\widetilde{O}(2^{\kappa}/\epsilon)$ to $\widetilde{O}(1/\epsilon^2)$ over different distributions as Lipschitzness decreases ($\alpha \to 0$) and noise level increases ($\beta \to 0$). The best rate of $\widetilde{O}(2^{\kappa}/\epsilon)$ is attained for distributions with large α and β, in which case the best possible dependence on tree-size is achieved (2^{κ} captures the size of the first level in a tree). Finally, recall that if the Bayes classifier is not in $\mathcal{H}$, even under deterministic labels (or $\beta \to \infty$), passive learning requires $\Omega(1/\epsilon^2)$ labels to achieve arbitrary small excess error (Ben-David and Urner, 2014). As evidenced by Lemma 10, the conditions of Theorem 11, namely low-noise and Lipschitzness, are stronger than the clusterability condition captured by Γ. Therefore, as per Theorem 7, we may expect better label complexity in practice than illustrated by Theorem 11.

5.2. Practical Clustering procedures

So far we have been a bit informal about how our requirements on the clustering procedure T are satisfied by state-of-the-art partitioning procedures. In other words, what can be said about the cell complexity $V_{T,\delta}$ and the tree-growth rate κ in general? In this section we tie these remaining loose-ends. We will argue that both κ and $V_{T,\delta}$ can be expected to be small when the data space $\mathcal{X} \subset \mathbb{R}^D$ has low intrinsic dimension $d \ll D$.

Definition 12 (Intrinsic dimension) *$\mathcal{X}$ has* **doubling dimension** *d if all balls $B(x, r)$, $x \in \mathcal{X}, r > 0$, can be covered by $(r/2)^d$ balls of radius $r/2$.*

RP-TREES AND OTHER RANDOMIZED k-d TREES

Suppose $\mathcal{X} \subset \mathbb{R}^D$ has doubling dimension $d \ll D$. Consider RPtrees and randomly oriented k-d trees. These are binary partitioning procedures with 2^i cells at level i. It is known that, for some $\kappa = O(d \log d)$, the diameter of the data ($\text{diam}(C \cap X_{1:n})$) in each cell C at level $i = l \cdot \kappa$ is at most 2^{-l}. This is shown, w.h.p., for RPtree in Theorem 3 of Dasgupta and Freund (2008), and for randomized k-d tree in Theorem 2 of Vempala et al. (2012)).

However, these trees do not directly satisfy our assumptions in that they create cells C whose diameter $\text{diam}(C)$ might never decrease. However, we can trivially extend them to trees where the cell diameters $\text{diam}(C)$ decrease by intersecting their cells with balls centered on points $x \in \mathcal{X}$. In other words, let T_0 be the original tree procedure such as RPtree of randomized k-d tree. We convert $T_0(X_{1:n})$ to $T(X_{1:n})$ as follows. To build level l of T, replace every cell C at level $i = l \cdot \kappa$ with $C \cap B$ where B is the smallest enclosing ball of the data. $C \cap B$ has diameter at most 2^{-l}, and the new tree has tree-growth rate $\kappa = O(d \log d)$.

Next we have to ensure that the VC complexity of cells of T has not increased too much. The possible cells of T are obtained as $\mathcal{C}_T(\rho) \triangleq \{C \cap B(x, r) : C \in \mathcal{C}_{T_0}(\rho), x \in \mathcal{X}, r > 0\}$. Thus, let $\mathcal{B}$ denote the set of balls centered on $\mathcal{X}$, and let $V_{\mathcal{B}}$ denote its VC dimension. It is well known that $V_{T,\delta} \leq 2(V_{\mathcal{B}} + V_{T_0,\delta}) \log(V_{\mathcal{B}} + V_{T_0,\delta})$. For low-dimensional $\mathcal{X}$, we can expect $V_{\mathcal{B}} = O(d)$. Last, the complexity of cells of RPtree is shown in (Kpotufe and Dasgupta, 2012) to be at most $O(d)$. This also yields a bound on the complexity of cells of randomized k-d trees since RP-trees are simply a more complex, randomized version of k-d trees.

AXIS PARALLEL TREES SUCH AS DYADIC TREES

The dyadic tree is a binary partitioning procedure with 2^i cells at level i. It is generally not data dependent, but partitions the original space $\mathcal{X}$. In this case, for some $\kappa_0 = O(D)$, at level $i = l \cdot \kappa_0$, the diameters of the cells are at most 2^{-i}. Thus a dyadic tree T_0 is trivially converted to a tree T satisfying our conditions by building the level l of $T(X_{1:n})$ with the cells of level $i = l \cdot \kappa_0$ of T_0. Thus the tree-growth rate κ of T is at most $\kappa_0 = O(D)$. In fact it can be much better for large trees: if $\mathcal{X}$ has low doubling dimension d, then most cells of a large tree T_0 at level $i = i(l)$ are empty. Formally, as shown in Theorem 24 of Kpotufe and Dasgupta (2012), for a dyadic number $0 < r < 1$ at most some r_0, the number of non-empty cells of T_0 of radius r is at most Cr^{-d} (where C depends on D). It follows that for all $0 < r < 1$ the number of nonempty cells is at most $C'r^{-d}$ (where C' depends on C and r_0. Thus the tree-growth rate κ of large trees T is likely closer to $O(d)$.

6. Analysis of the label complexity bound in Theorem 7

The essential result of this section is Lemma 16 where we show that *nice* clusters, i.e. clusters where the minority label has mass a most $O(\epsilon)$, remain nice under a random sample $X_{1:n}, Y_{1:n}$, provided they contain enough samples from $X_{1:n}$. Label savings are contigent on detecting such nice clusters, hence the importance of the lemma. Our analysis makes use of the following result in various ways:

Lemma 13 (Relative VC bounds Vapnik and Chervonenkis (1971)) *Consider a collection $\mathcal{A}$ of measurable subsets of some domain $\mathcal{D}$, and let $V_{\mathcal{A}}$ be its VC dimension. Let $0 < \delta_0 < 1$. Suppose a sample of size m is drawn i.i.d. from a distribution over $\mathcal{D}$. For $A \in \mathcal{A}$, let ν_A denote the mass of A under the distribution, and let $\nu_{m,A}$ denote its empirical mass. Define*

$\alpha_{m,\mathcal{A}} = (V_{\mathcal{A}} \ln 2m + \ln(8/\delta_0))/m$. *Then with probability at least* $1-\delta_0$ *over the sampling, all* $A \in \mathcal{A}$ *satisfy* $\nu_A \leq \nu_{m,A} + \sqrt{\nu_{m,A} \cdot \alpha_{m,\mathcal{A}}} + \alpha_{m,\mathcal{A}}$, *and* $\nu_{m,A} \leq \nu_A + \sqrt{\nu_A \cdot \alpha_{m,\mathcal{A}}} + \alpha_{m,\mathcal{A}}$.

This simple Corollary to the relative VC-bound (Lemma 13) is used in various proofs:

Corollary 14 (Corollary to Lemma 13) *Let* $0 < \delta_0 < 1$ *and* $0 < \epsilon_0 < 1/2$. *Consider a Bernouilli*(ν). *Let* ν_m *be the empirical estimate of* ν *from an i.i.d. sample of size* $m \geq \ln(8/\delta_0)/\epsilon_0$. *We have with probability at least* $1-\delta_0$ *over the sampling* $\nu \leq \nu_m + \sqrt{\nu_m \cdot \epsilon_0} + \epsilon_0$, *and* $\nu_m \leq \nu + \sqrt{\nu \cdot \epsilon_0} + \epsilon_0$.

Proof Apply Lemma 13 with $\mathcal{D} = \{0,1\}$, and $\mathcal{A} = \{\{1\}\}$ with $V_{\mathcal{A}} = 0$. The rest is algebra to verify that α_m in the Lemma is at most ϵ_0. ■

The following is used in the proof of Lemma 16:

Lemma 15 (Implied by Theorem 13 of Boucheron et al. (2004)) *Consider independent r.v.s* $Y_1, \ldots, Y_m$, $0 \leq Y_i \leq 1$, *and let* ν_m *denote* $\frac{1}{m}\sum_i Y_i$, *and* $\nu = \mathbb{E}\,\nu_m$. *Let* $0 < \delta_0 < 1$, *and let* $\alpha_m = 2\ln(1/\delta_0)/m$. *With probability at least* $1-\delta_0$ *over the randomness in* $Y_1, \ldots, Y_m$, *we have*

$$\nu_m \leq \nu + \sqrt{\nu \cdot \alpha_m} + \alpha_m.$$

The main technicality in establishing Lemma 16 stems from the fact that the actual clusters considered by the procedure depend on the data. We proceed by first arguing that enough *good* points fall in each cell C: these are points x in C for which $\eta(x)$ is far from $1/2$. This event depends only on $X_{1:n}$ and not on $Y_{1:n}$ allowing a first decoupling of dependencies. We then argue that the labels Y generated for those points are typical. This event now depends just on $Y_{1:n}|X_{1:n}$.

Lemma 16 (Concentration of minority labels of clusters) *Let* $0 < \epsilon, \delta < 1$. *Let* $Y_{1:n} \doteq \{Y_i\}_1^n$ *be the underlying labeling for* $X_{1:n}$ *as defined in Definition 5. Suppose the hierarchical-clustering procedure* T *has VC-cluster dimension* $V_{T,\delta}$. *Assume* $n \geq 81\,(16 V_{T,\delta} \ln 2n + \ln(8/\delta))/\epsilon^2$. *The following holds with probability at least* $1-6\delta$ *over the sampling of* $X_{1:n}, Y_{1:n}$.

Let C *be any cluster at some level in* $T(\rho, X_{1:n})$ *such that* $\mu_n(C) \geq \epsilon$. *Let* η_C *and* $\eta_{n,C}$ *denote the probability of* 1 *under* $P_{X,Y}$ *and its empirical counterpart over* $X_{1:n}, Y_{1:n}$. *If* $\eta_C \leq \epsilon/162$, *then* $\eta_{n,C} \leq \epsilon/9$, *or if* $(1-\eta_C) \leq \epsilon/162$, *we have* $(1-\eta_{n,C}) \leq \epsilon/9$.

Proof Fix the randomness ρ of the tree-procedure T throughout the proof. Consider the collection $\mathcal{C}_T = \mathcal{C}_T(\rho)$ of possible clusters (see Definition 3). By definition, with probability at least $1-\delta$, $\mathcal{C}_T$ has VC complexity at most $V_{T,\delta}$. We only consider bounding $\eta_{n,C}$, the argument for $1-\eta_{n,C}$ being the same.

For every $C \in \mathcal{C}_T$, define $n_C \triangleq |C \cap X_{1:n}|$ and

$$\tilde{\eta}_{n,C} \triangleq \frac{1}{n_C} \sum_{x \in X_{1:n} \cap C} \eta(x) = \mathbb{E}_{Y_{1:n}} \left\{ \frac{1}{n_C} \sum_{x \in X_{1:n} \cap C} Y(x) \right\} = \mathbb{E}_{Y_{1:n}} \{\eta_{n,C}\}.$$

We will show in turn that for all C, $\tilde{\eta}_{n,C}$ is close to η_C, and then that $\tilde{\eta}_{n,C}$ is close to $\eta_{n,C}$ for clusters C with at least $O(1/\epsilon)$ data points. The first part depends only on the randomness in $X_{1:n}$,

while the second part depends only on $Y_{1:n}$ conditioned on $X_{1:n}$. This allows us to circumvent the dependency of algorithmic choices on the random sample at hand.

First, to bound $\tilde{\eta}_{n,C}$ in terms of η_C, recall that for any r.v. $Z > 0$, $\mathbb{E}\, Z = \int_0^\infty \mathbb{P}\left(Z > t\right)\, dt$. Using $Z = \eta(X)$ under both the distribution $P_{X|X\in C}$ and its empirical counterpart, we have

$$\eta_C = \frac{1}{\mu(C)} \int_0^1 \mu\left\{x \in C : \eta(x) > t\right\}\, dt, \text{ and } \tilde{\eta}_{n,C} = \frac{1}{\mu_n(C)} \int_0^1 \mu_n\left\{x \in C : \eta(x) > t\right\}\, dt.$$

Hence, consider the sets $C_t \triangleq \left\{x \in C : \eta(x) > t\right\}, t \in [0,1]$, and let $\mathcal{X}_t \triangleq \left\{x \in \mathcal{X} : \eta(x) > t\right\}$. We bound the VC complexity of the collection $\{C_t\}$ as follows. Consider the projection of $\{C_t\}$ onto $X_{1:n}$:

$$\{C_t\}_{|X_{1:n}} \triangleq \{C_t \cap X_{1:n}\} = \{\mathcal{X}_t \cap C \cap X_{1:n}\} = \left\{A \cap B : A \in \{\mathcal{X}_t\}_{|X_{1:n}} \text{ and } B \in \mathcal{C}_{T\,|X_{1:n}}\right\},$$

therefore $\left|\{C_t\}_{|X_{1:n}}\right| \leq \left|\{\mathcal{X}_t\}_{|X_{1:n}}\right| \cdot \left|\mathcal{C}_{T\,|X_{1:n}}\right| \leq en \cdot (\frac{e}{V_{T,\delta}} n)^{V_{T,\delta}}$, where we used Sauer's lemma and the fact that the collection $\{\mathcal{X}_t\}$ is ordered by inclusion along the segment $t \in [0,1]$, and hence has VC dimension 1. It follows from this shattering bound that the VC dimension V of $\{C_t\}$ is at most $16V_{T,\delta}$ since by definition of V, we then have $2^V \leq eV \cdot (\frac{e}{V_{T,\delta}} V)^{V_{T,\delta}}$ (the above argument holds for any sample size, including $n = V$).

We can therefore apply Lemma 13 over $\{C_t\}$ with $\alpha_n = \left(V \ln 2n + \ln(8/\delta)\right)/n$. With probability at least $1 - 2\delta$, for all $C_t, t \in [0,1]$, we have $\mu_n(C_t) \leq \mu(C_t) + \sqrt{\mu(C_t)\alpha_n} + \alpha_n$, hence

$$\begin{aligned}
\tilde{\eta}_{n,C} \leq & \frac{1}{\mu_n(C)} \int_0^1 \left(\mu(C_t) + \sqrt{\mu(C_t)\alpha_n} + \alpha_n\right)\, dt && (1)\\
\leq & \frac{1}{\mu_n(C)} \int_0^1 \mu(C_t)\, dt + \sqrt{\frac{\alpha_n}{\mu_n(C)}} \left(\frac{1}{\mu_n(C)} \int_0^1 \mu(C_t)\, dt\right)^{1/2} + \frac{\alpha_n}{\mu_n(C)}, && (2)
\end{aligned}$$

where we used Jensen's inequality on the $\sqrt{\cdot}$ term of (1) and rearranged. Now, with probability at least $1 - \delta$, again by Lemma 13 we have for all $C \in \mathcal{C}_T$, $\mu(C) \leq \mu_n(C) + \sqrt{\mu_n(C)\alpha_n} + \alpha_n$.

This last inequality implies that for all C satisfying $\mu_n(C) \geq \epsilon \geq 81\alpha_n/\epsilon$, we have $\mu_n(C) \geq \mu(C)/(1 + 2\sqrt{\epsilon/81})$. Also, for such C, $\alpha_n/\mu_n(C) \leq \epsilon/81$. It follows from (2) that for all such C,

$$\tilde{\eta}_{n,C} \leq \left(1 + 2\sqrt{\frac{\epsilon}{81}}\right)\eta_C + \sqrt{\frac{\alpha_n}{\mu_n(C)}}\left(\left(1 + 2\sqrt{\frac{\epsilon}{81}}\right)\eta_C\right)^{1/2} + \frac{\alpha_n}{\mu_n(C)} \leq 2\eta_C + \sqrt{\frac{\epsilon}{81} \cdot 2\eta_C} + \frac{\epsilon}{81}.$$

Thus, if $\eta_C \leq \epsilon/162$, then with probability at least $1 - 3\delta$, $\tilde{\eta}_{n,C} \leq \epsilon/27$.

Next, we want to show that for large clusters C where $\mu_n(C) \geq \epsilon$, $\tilde{\eta}_{n,C}$ is close to $\eta_{n,C}$. Condition on $X_{1:n}$ fixed. There are then at most $\left|\mathcal{C}_{T|X_{1:n}}\right| \leq (4n)^{V_{T,\delta}}$ equivalent clusters to consider since the quantities of interest depend only on the sample points falling in the clusters. We can thus apply Lemma 15, for each cluster C, with $\delta_0 = \delta/(4n)^{V_{T,\delta}}$ and $\alpha_{n_C} = 2\ln(1/\delta_0)/n_C$. Notice that if $\mu_n(C) \geq \epsilon$, i.e. $n_C \geq \epsilon \cdot n$, we have $\alpha_{n_C} \leq 2\ln(1/\delta_0)/\epsilon \cdot n \leq \epsilon/27$. We therefore have that, with probability at least $1 - 4\delta$, for all C satisfying $\mu_n(C) \geq \epsilon$ and $\eta_C \leq \epsilon/162$,

$$\eta_{n,C} \leq \tilde{\eta}_{n,C} + \sqrt{\tilde{\eta}_{n,C} \cdot \alpha_{n_C}} + \alpha_{n_C} \leq \epsilon/9.$$

To finish, repeat the same argument for $1 - \eta_{n,C}$, and take into accounts the shared events to obtain a total probability of failure of 6δ. ■

6.1. Proof of Theorem 7

Fix any level $l \in \mathbb{N}$. We first bound the number of label requests up to level l (including all labels requested at level l). To start, consider the total number of label requests on the active clusters $\mathcal{C}^l$. On each cluster $C \in C^l$, we request at most $n_{l,\epsilon} \triangleq \max\{n \cdot \epsilon, n_l(\epsilon)\}$ labels. Note that $n_{l,\epsilon}$ grows with l. Now w.l.og. the labels on cluster C can be requested as follows: let $\left\{Y_{(i)}\right\}_1^m$, for some $m \geq 0$, be the sequence of labels requested for points in C at level $l-1$ (while requesting labels for points in the parent of C); simply ask for $(n_{l,\epsilon} - m)_+$ more labels to add to $\left\{Y_{(i)}\right\}_1^m$. Hence, recursively from the first level 0, the number of labels requested on the union of cells of C^l up to level l is at most $\left|C^l\right| \cdot n_{l,\epsilon}$. The same argument applies to non-active clusters, hence the total number of label requests up to level l is at most $|T_l(\rho, X_{1:n})| \cdot n_{l,\epsilon}$. We have by definition of κ that $|T_l(\rho, X_{1:n})| \leq 2^{\kappa l}$. Use this fact to also bound $n_l(\epsilon)$. Combined with the lower-bound on n, we have that the number of requests up to level l is at most $2^{\kappa l} \cdot n \cdot \epsilon \cdot (\kappa l + l + 1) \leq n \cdot \epsilon \cdot 2^{\kappa l}(2\kappa l)$.

Next we bound the number of labels yet to request at later levels. For any $x \in \mathcal{X}$, let $\mathcal{C}^l(x)$ denote the cluster at level l to which it belongs. Define the set

$$\mathcal{X}_l \triangleq \left\{x \in \mathcal{X} : \left|\eta_{\mathcal{C}^l(x)} - \frac{1}{2}\right| \geq \frac{1}{2} - \epsilon/162\right\}.$$

By definition, for every such $x \in \mathcal{X}_l$, the minority label of $\mathcal{C}^l(x)$, has mass (conditioned on $\mathcal{C}^l(x)$) at most $\epsilon/162$ under $P_{X,Y}$. Suppose $\mu_n(\mathcal{C}^l(x)) < \epsilon$, then by definition of the algorithm, x is labeled at level l. Now suppose that $\mu_n(\mathcal{C}^l(x)) \geq \epsilon$, then by Lemma 16, with probability at least $1 - 6\delta$, the minority label of $\mathcal{C}^l(x)$ has mass at most $\epsilon/9$ under the empirical distribution on $X_{1:n}, Y_{1:n}$. Fix any such $C = \mathcal{C}^l(x)$ and w.l.o.g. let the minority label be 1, and let $\eta_{n,C}$ be the empirical probability of 1 in $Y_{1:n}$ over points in $C \cap X_{1:n}$. By Corollary 14, with probability at least $1 - \delta/\left|\mathcal{C}^l\right| 2^{l+1}$,

$$\hat{\eta}_S \leq \eta_{n,C} + \sqrt{\eta_{n,C} \cdot \epsilon/9} + \epsilon/9 \leq \epsilon/3,$$

so all of C is labeled by the procedure. Thus, with probablity at least $1 - 7\delta$, for every level l, for every $x \in \mathcal{X}_l$, the cluster $\mathcal{C}^l(x)$ is labeled by the procedure at level l. So the number of points left to label after level l is at most $|X_{1:n} \setminus \mathcal{X}_l|$. We bound this quantity next.

Let $\tau \triangleq 1/2 - \epsilon/162$, and recall that, by definition of a level, all clusters at level l have diameter at most 2^{-l}. By Definition 4, $|X_{1:n} \setminus \mathcal{X}_l|$ would be bounded in expectation by $\Gamma\left(\tau, 2^{-l}\right)$. For a high probability bound, which holds simultaneously for all levels l, we proceed by first bounding the VC complexity of typical points in $|X_{1:n} \setminus \mathcal{X}_l|$. Recall the $\mathcal{C}_{T,r}$ (Definition 4). Notice that, for τ fixed as above, the sets $\mathcal{X}^r \triangleq \{x : \exists C \in \mathcal{C}_{T,r}(x), \{\eta_C - 1/2\} < \tau\}$ are ordered by inclusion along the line $r \in \mathbb{R}^+$, i.e. $\mathcal{X}^r \subset \mathcal{X}^{r'}$ if $r < r'$. Therefore, the collection $\mathcal{A} \triangleq \{\mathcal{X}^r\}_{r>0}$ has VC dimension $V_{\mathcal{A}} = 1$. By Definition 4, $\mu(\mathcal{X}^r) \leq \Gamma(\tau, r)$, and hence by Lemma 13, with probability at least $1 - \delta$ over the choice of $X_{1:n}$, for all levels l,

$$|X_{1:n} \setminus \mathcal{X}_l| \leq n \cdot \mu_n\left(\mathcal{X}^{2^{-l}}\right) \leq n \cdot \left(\Gamma\left(\tau, 2^{-l}\right) + \sqrt{\Gamma\left(\tau, 2^{-l}\right) \cdot \alpha_n} + \alpha_n\right).$$

Since the number of labels requested is bounded by $\epsilon \cdot n \cdot 2^{\kappa l}(2\kappa l) + |X_{1:n} \setminus \mathcal{X}_l|$ for any $l \in \mathbb{N}$, the result follows. ■

References

M.-F. Balcan, S. Hanneke, and J. Wortman. The true sample complexity of active learning. *COLT*, 2008.

N. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. *ICML*, 2006.

Shai Ben-David and Ruth Urner. The sample complexity of agnostic learning under deterministic labels. In *Proceedings of The 27th Conference on Learning Theory*, pages 527–542, 2014.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Advanced Lectures on Machine Learning*, pages 208–240. Springer, 2004.

S. Dasgupta and Y. Freund. Random projection trees and low dimensional manifolds. *STOC*, 2008.

S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. *NIPS*, 2007.

Sanjoy Dasgupta. Two faces of active learning. *Theoretical computer science*, 412(19):1767–1781, 2011.

Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning*, pages 208–215. ACM, 2008.

S. Hanneke. Adaptive rates of convergence in active learning. *COLT*, 2009.

Samory Kpotufe and Sanjoy Dasgupta. A tree-based regressor that adapts to intrinsic dimension. *Journal of Computer and System Sciences*, 78(5):1496–1515, 2012.

Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Annals of Statistics*, 27(6):1808–1829, 1999.

P. Massart and E. Nedelec. Risk bounds for statististical learning. *Annals of Statistics*, 34(5):2326–2366, 2006.

Philippe Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. In *Journal of Machine Learning Research*, pages 1369–1392, 2007.

Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32(1):135–166, 2004.

Ruth Urner, Sharon Wullf, and Shai Ben-David. Plal: Cluster-based active learning. In *Proceedings of the Conference on Learning Theory (COLT)*, 2013.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their expectation. *Theory of probability and its applications*, 16:264–280, 1971.

Santosh Vempala, Deepak D'Souza, Telikepalli Kavitha, and Jaikumar Radhakrishnan. Randomly-oriented kd trees adapt to intrinsic dimension. In *FSTTCS*, pages 48–57, 2012.

Appendix A. Miscelleneous

The following is rather well known.

Lemma 17 *Let H_1, H_2 be classes of binary valued functions over the same domain set, X. Assume $VCdim(H_1) = d_1$ and $VCdim(H_2) = d_2$ are both finite. Let $H_1 \sqcap H_2$ denote $\{h_1 \cap h_2 : h_1 \in H_1$ and $h_2 \in H_2\}$. Then*

$$VCdim(H_1 \sqcap H_2) \leq 2(d_1 + d_2)\log(d_1 + d_2).$$

Proof [Proof Outline] Let A be any set shattered by $H_1 \sqcap H_2$. Then $2^{|A|} \leq |\{h \cap A : h \in H_1 \sqcap H_2\}|$. Note that $|\{h \cap A : h \in H_1 \sqcap H_2\}| \leq |\{h \cap A : h \in H_1\}| \times |\{h \cap A : h \in H_2\}|$. It follows that, for $D = VCdim(H_1 \sqcap H_2$, one must have $2^D \leq |\{h \cap A : h \in H_1\}| \times |\{h \cap A : h \in H_2\}|$. Applying Sauer's lemma, it is straightforward to see that this inequality fails for any $D > 2(d_1 + d_2)\log(d_1 + d_2)$. ■

Appendix B. Labeling Error

Proof [Proof of Theorem 6]

Fix $X_{1:n}$ and $Y_{1:n}$ (unknown by the procedure) throughout. Consider any cluster C at some level l that was labeled by the procedure. Suppose w.l.o.g. that the majority label is 0.

Since the same label is returned each time the same X is sampled from $C \cap X_{1:n}$, the sampling is equivalent to sampling with replacement from the unknown $Y_{1:n}$ over points in C. Therefore apply Corollary 14 with ν being the empirical distribution of label 1 out of $Y_{1:n}$ over points in $C \cap X_{1:n}$, and $m = |S|$, $\delta_0 = \delta / \left|\mathcal{C}^l\right| 2^{l+1}$. Correspondingly, let ν_m be the empirical distribution of 1 in the sample S. Then, with probability at least $1 - \delta / \left|\mathcal{C}^l\right| 2^{l+1}$,

$$\nu \leq \nu_m + \sqrt{\nu_m \cdot \epsilon/9} + \epsilon/9 \leq \epsilon.$$

Thus, conditioned on C, labeling all $X_{1:n} \cap C$ as 0 has error at most ϵ. Note that, while the decision to sample a given C depends on past events, the sampling itself, conditioned on C is independent of past events. Therefore, the labeling error is at most ϵ for all such C labeled by the procedure, this happening with probability at least $1 - \delta \sum_{l, C \in \mathcal{C}^l} 1 / \left|\mathcal{C}^l\right| 2^{l+1} \geq 1 - \delta$. The result follows by integrating the error (w.r.t. the empirical distribution on $X_{1:n}$) over the disjoint clusters. ■