POINTWISE CONFIDENCE SCORES WITH PROVABLE GUARANTEES

Nivasini Ananthakrishnan,[1] **Shai Ben-David,** [1] **Tosca Lechner** [1]

November 29, 2020

ABSTRACT

Quantifying the probability of a machine learning prediction being correct on a given test point enables users to better decide how to use those predictions. Confidence scores are pointwise estimates of such probabilities. Ideally, these would be the label probabilities (a.k.a. the *labeling rule*) of the data generating distribution. However, in learning scenarios the learner does not have access to the labeling rule. The learner aims to figure out a best approximation of that rule based on training data and some prior knowledge about the task at hand, both for predicting a label and for possibly providing a confidence score. We formulate two goals for confidence scores, motivated by the use of these scores for safety-critical applications. First, they must not under-estimate the probability of error for any test point. Second, they must not be trivially low for most points. We consider several common types of learner's prior knowledge and provide tools for obtaining pointwise confidence scores based on such prior knowledge and the relation between the test point and the training data. We prove that under the corresponding prior knowledge assumptions, our proposed tools meet these desired goals.

1 Introduction

The reliability of machine learnt programs is of course a major concern and has been the focus of much research. Theory offers quite a selection of tools for evaluating reliability, from generalization bounds to experimental result of test sets. However, most of those guarantees are statistical, in the sense that they only hold with high probability (over the generation of the training data and of the test points) and they provide no information about the correctness of prediction on any specific instance. In cases where an error on a specific instance may incur a very high cost, like in safety-critical applications, the common statistical guarantees do not suffice. We would also wish to be able to identify predictions with low confidence so that one could apply some safety procedures (such as a review by a human expert). Ideally, no low confidence prediction should go undetected, At the same time, since expert intervention could be expensive, one also wishes to minimize the occurrence of false positives in the predictions flagged as low confidence.

Can one do better than the overall statistical estimates when it comes to evaluating reliability on a given test case?

Arguably, the most common reason for an statistically reliable machine learning program to fail on a test point is that that point is an 'outlier', in the sense of not being well represented by the sample the program was trained on. This research aims to quantify this 'outlierness'. We propose theoretically founded confidence bounds that take into account the relation between the training sample and the specific test point in question (on top of the commonly used parameters of the learning algorithm, the size of training sample and assumptions about the processes generating both the training data and the test instance).

Clearly, the confidence of any prediction of an unknown label (or any piece of information) hinges upon some prior knowledge or assumptions. In this work we consider several forms of prior knowledge that are commonly employed in machine learning theory, and develop and analyse confidence score for prediction of individual test points under such assumptions.

We consider the following types of prior knowledge:

Known hypothesis class with low approximation error We discuss two cases - the realizable setting (i.e., when that approximation error is zero) and the agnostic setup (both in Section 4).

- In the realizable case, we show that there are indeed hypothesis classes for which it is possible to define a confidence score that does not overestimate confidences for any points, while providing high confidences to many points. However, there are also hypotheses classes, that do not allow non-trivial confidence scores fulfilling such a guarantee.
- For the agnostic setup, assuming the learner has knowledge of a hypothesis class with low (but not necessarily 0) approximation error, we show that in this case it is not possible to give any non-trivial confidence score that does not overestimate confidence for some instances.

The data generating distribution is Lipschitz We provide a an algorithm that calculates confidence scores under such an a Lipschitzness assumption. We show that with high probability over samples, the resulting confidence score of every point is an underestimate of its true confidence while the confidence score we obtain is non-trivial. We provide bounds on the probability (over points and samples) of assigning a low confidence score to a point with high true confidence that converge to zero as a function of the training sizes. For more details, see Section 5.

The Bayes classifier is adversarially robust In Section 6, we analyze confidence scores for distributions for which the Bayes classifier is adversarially robust (with respect to a general notion of domain neighbourhood sets).

We show that under this assumption, for a neighbour set with high enough probability weight, all points in it will be assigned the same label by the Bayes classifier. Based on unlabelled samples, we can conclude that a neighbour set with many sample points has high probability weight. If additionally, we had a good estimate of the expected label of a high weight set, we can determine the Bayes label of all points in the set. We could get this estimate from labelled samples. In the scenario where we don't have enough labelled samples, we can still estimate the average label of the set by making other assumptions. These assumptions are discussed in Sections 6.1.1, 6.1.2

Having access to a learner with good generalization In Section 6.1.1, we assume that the underlying labeling function is deterministic as well as adversarially robust. We introduce a semi-supervised confidence score (based on both labeled and unlabeled samples) for learners that output a low-risk classifier with high probability (over training samples). In this setting, we show that with high probability over the sample generation our confidence score does not overestimate, on any instance, the probability of its label under the data generating distribution. .

Having a uniform convergence bound for a class In section 6.1.2, we consider the assumption of having a a hypothesis class with a low approximation error and a uniform convergence bound (namely, for some $m : (0,1) \times (0,1) \to \mathbb{N}$ for all ϵ, δ, if $m \geq m(\epsilon, \delta)$ then $Prob_{S \sim P^m}\left[|L_S(h) - L_P(h)| > \epsilon\right] \leq \delta$, where P is the data generating distribution). This assumption allows us to estimate the majority label of the Bayes classifier on a subset of the domain. By using this assumption along with the adversarially robust Bayes assumption, we can estimate the Bayes classifer's label for some points.

Next, in Section 2, we give an overview over related work.

2 Related work

The closest previous work to ours is Jiang et al [14]. They consider a very similar problem to the one we address here - the problem of determining when can a classifier's prediction on a given point be trusted. For the sake of saving space, we refer the reader to that paper for a more extensive review of previous work on this topic. Their theoretical results differ from our work in two essential aspects. First, they consider only one setup - a data generating distribution that satisfies several technical assumptions. In particular they rely on the following strong condition: "for any point $x \in X$, if the ratio of the distance to one class's high-density-region to that of another is smaller by some margin γ, then it is more likely that x's label corresponds to the former class." We analyse our notion of confidence under several different incomparable assumptions, arguably, none of which is as strong as that. The second significant difference is that the main result on trust of labels there (theorem 4 of [14]) states that if a certain inequality holds then the predicted label agrees with that of the Bayes optimal predictor, and if another inequality holds, there is disagreement between them. However, those inequalities are not complementary. It may very well be that in many cases every domain point (or high probability of instances) fails both inequalities. For such points, that main result tells nothing at all. That paper does not offer any discussion of the conditions under which their main result is not vacuous in that respect. Additionally their result holds with high probability over the domain and is not a point-wise guarantee.

Selective Classification/ Classification with Abstension: One line of work that is related to our paper is learning with abstention. Similar to our setting, the classification problem does not only consist of the goal of classifying correctly, but to also allows the classifier to abstain from making a prediction, if the confidence of a prediction is too low. Many works in this line provide accuracy guarantees that hold with high probability over the domain ([15],[16], [17], [18], [19]). This is different from our goal of point-wise guarantees.

Point-wise guarantees are provided in [20] and [21]. El-Yaniv et al [20] gave a theoretical analysis of the selective classification setup in which a classification function and a selective function are learned simultaneously. The risk of a classification is then only accessed on the set of instances that was selected for classification. The selective function is evaluated by their coverage - how many instances in expectation are selected for classification. They analyse the trade-off between risk and coverage, and introduce the notion of "perfect classification" which requires risk 0 with certainty. This is similar to our requirements on a confidence score in the deterministic setting, where we require 0 risk with high probability. Their notion of coverage is similar to our notion of non-redundancy - in fact non-redundancy corresponds to worst-case coverage over a family of distributions. They provide an optimal perfect learning strategy in the realizable setting and show that there are hypothesis classes with a coverage that converges to 1 and hypothesis classes for which coverage is always 0 for some distributions. We use their results in our Section 4. In contrast to their paper our setup also considers probabilistic labeling functions and our analysis also considers other assumptions on the family of probability distributions, besides generalization guarantees for some fixed hypothesis space.

In the non-realizable case, Wiener et al [21] provide guarantees of agreeing point-wise with the best classifier in hindsight from a function class. We are interested in agreement with the Bayes classifier and not the best classifier from the class. We use additional assumptions about the error of the best classifier from the function class to provide guarantees about agreement with the Bayes classifier. They introduce a quantity called disbelief index of points. We use an extension of this quantity for our results. Rather than a disbelief index for a point, we use disbelief index for a set. We show how to combine the disbelief index with accuracy guarantees of the function class to provide guarantees on agreement with Bayes classifier.

3 Problem definition

Let the domain of instances be X and the label set be $\{0,1\}$. A learning task is determined by a probability distribution P over $X \times \{0,1\}$. We denote the marginal distribution over the domain by P_X and the conditional labeling rule by ℓ_P (namely, for $x \in X$, $\ell_P(x) = \Pr_{(x',y')\sim P}[y' = 1|x' = x]$).

The Bayes classifier,

$$h_P^B(x) = 1 \text{ iff } \Pr_{(x',y')\sim P}[y' = 1|x' = x] \geq 0.5,$$

is the pointwise minimizer of the zero-one prediction loss. We sometime refer to its prediction $h_P^B(x)$ as the *majority label of a point* or the *Bayes label of a point*.

We are interested in point-wise confidence. For a point $x \in X$, the **confidence of a label** $y \in \{0,1\}$ is

$$C_P(x,y) = \Pr_{(x',y')\sim P}[y' = y|x' = x].$$

Note that the label assigned by he Bayes predictor maximizes this confidence for every domain point x.

We have two problems of interest. Both these problems are about inferring information about the labelling probability of points based on a training sample. The goal is to provide point-wise guarantees about the correctness of the information inferred.

The first information we try to infer is the actual labelling probability of a point. A **Confidence score** of the label confidence is an empirical estimate (based on some training sample S) of the true label confidence. Inevitably, the reliability of a confidence score is dependent on some assumptions on the data generating distribution (or, in other words, prior knowledge about the task at hand). Given a family of data generating distributions $\mathcal{P}$ (fulfilling some niceness assumption that reflect the learners prior knowledge or beliefs about the task), a training sample S, and a parameter δ, the *empirical confidence estimate for a point x and label y* is a function $C(x,y,S,\delta)$. We want the following property to hold: For every probability distribution $P \in \mathcal{P}$, with probability of at least $1-\delta$ over an i.i.d. generation of S by P, we have

$$\Pr_{y'\sim \text{Bernoulli}(\ell_P(x))}[y' = y] \geq C(x,y,S,\delta).$$

That is, with high probability, we do not overestimate the probability of y being the correct label for x. Ideally, this should hold for every point x in the domain. Of course, there is a trivial solution for this - just let

$C(\ ,\ ,\ ,\ )$ be the constant 0 function. The goal therefore is to get a confidence score that fulfils the condition above, while still being as high as possible for 'many' $x's$. That is, we aim for a confidence score, such that $\mathbb{E}_{x\sim P, S\sim P^m}[\max\{C(x,1,S,\delta), C(x,0,S,\delta)\}]$ is high. As mentioned above, given a data generating distribution P and a data representation available to the learner, the highest confidence on every instance $x \in X$ is achieved by the Bayes predictor $h^B_P(x)$ and it is easy to see that it is $\max\{\ell_P(x), (1-\ell_P(x))\}$.

The second information we try to infer is if the labelling probability of a point is over or under 0.5. Given a family of a data generating distributions $\mathcal{P}$, a training sample S, and a parameter δ, the *empirical Bayes label estimate for a point* x is a function $C_\ell(x, S, \delta) \in \{0, 1, \text{abstain}\}$. We want the following property to hold: For every probability distribution $P \in \mathcal{P}$, with probability at least $1-\delta$ over an i.i.d. generation of S by P, we have $C_\ell(x, S, \delta) = \text{abstain}$ or $C_\ell(x, S, \delta) = h^B_P(x)$. That is, with high probability, we do not predict the Bayes label of the point incorrectly. We allow abstaining instead of predicting to allow this property to be possible. We can achieve this properly trivially by abstaining on all points. An additional property that we want is that we abstain on few points. That is, we aim for an empirical Bayes label estimate such that $\Pr_{x\sim P, S\sim P^m}[C_\ell(x, S, \delta) = \text{abstain}]$ is low.

In contrast with the common notion of a PAC style error bound is that confidence scores may vary over individual instances, capturing the heterogeneity of the domain and the specific training sample the label prediction relies on. To demonstrate this point, consider the following example:

Example 1. *Let X_1 be the 0.1 grid over $[0,1]^d$, let X_0 the 0.01 grid over $[0, 0.1]^d$ and let our domain be $X = X_0 \cup X_1$ (for some large d). Consider the family $\mathcal{P}$ of all probability distributions over X that have a deterministic labeling rule satisfying the 10- Lipschitz condition (so points of distance 0.1 or more have no effect on each other). Assume further that all the distributions in $\mathcal{P}$ have half of their mass uniformly distributed over X_1 grid points and the other half of the mass uniformly distributed over X_0.*

Since outside the $[0, 0.1]^d$ cube, every labeling is possible, for every learner there is a distribution $P \in \mathcal{P}$ w.r.t. which it errs on every domain point in $X_1 \setminus S_X$ (where $S_X = \{x : \exists y \in \{0,1\}$ s.t. $(x, y) \in S\}$). On the other hand, due to the Lipschitz condition, all the points in the $[0, 0.1]^d$ grid, X_0, must get the same label. Therefore, given a sample S that includes a point in X_0, a learner that labels all the points in X_0 by the label of the sample points in it induces confidence 1 for all these points.

We conclude that, for sample sizes between 2 and, say, $10^d/2$, for most of the samples a learner can achieve confidence 1 for points in X_0 and no learner can achieve confidence above 0 for even a half of the domain points in X_1. Note also that the No Free Lunch theorem (as formulated in, e.g., [22]) implies that for sample sizes smaller than $10^d/2$, for every learner there exists some probability distribution $P \in \mathcal{P}$ for which its expected error is at least $1/8$ ($1/4$ over a subspace X_1 that has probability weight $1/2$).

4 Confidence scores for hypothesis classes

In the following we will analyze the point-wise confidence when all the prior knowledge available about the data generating distribution P is a bound on the approximation error of a class of predictors. We will distinguish two cases here,

1. The family $\mathcal{P}_{\mathcal{H},0}$ of distributions P which are realizable w.r.t. $\mathcal{H}$, i.e. $\inf_{h\in\mathcal{H}} L_P(h) = 0$ and
2. The family $\mathcal{P}_{\mathcal{H},\epsilon}$ of distributions P for which the approximation error of class $\mathcal{H}$ is low but not guaranteed to be zero, i.e. $\inf_{h\in\mathcal{H}} L_P(h) \leq \epsilon$, for some $\epsilon > 0$.

Note that, given a class of predictors, $\mathcal{H}$, the second family of possible data generating distributions is a superset of the first. Consequently, the pointwise error guarantees one can give in that non-necessarily-realizable case are weaker[1].

Definition 1 (Confidence Score, fulfilling the no-overestimation guarantee for all instances)**.** *We say a function C, that takes as input a sample S, a point x, a hypothesis h and a parameter δ and outputs a value in $[0,1]$. We say such a function C is a* confidence score fulfilling the no-overestimation guarantee for all instances *for a family of probability functions $\mathcal{P}$ if for every $P \in \mathcal{P}$ the probability over $S \sim P^m$ that there exists $x \in X$*

$$Pr_{y\sim Bernoulli(\ell_P(x))}[h(x) = y] < C(x, y, S, \delta)$$

is less than δ.

[1]To not have to deal with the ambiguity of the labeling function for points with mass 0, we will restrict this discussion to the family of distributions which have positive mass on all points. This implies that in the realizable setting all labeling function ℓ_P we consider are part of $\mathcal{H}$.

We say a function $C(x, y, S, \delta)$, is a *confidence score fulfilling the fulfilling the no-overestimation guarantee for positive mass instances* if the above guarantee holds not for all x, but for all x with $P(\{x\}) > 0$.

Definition 2 (Non-redundancy)**.** *Given a family of probability distributions $\mathcal{P}$ and a confidence score C for $\mathcal{P}$, we define the non-redundancy $nr_{\mathcal{P}}(C)$ for a given sample size m and parameter δ, to be*

$$nr_{\mathcal{P}}(C, m, \delta) =$$

$$\inf_{P \in \mathcal{P}} \mathbb{E}_{x \sim P, S \sim P^m}[\max\{C(x, 0, S, \delta), C(x, 1, S, \delta)\}]$$

Next we consider a specific confidence score that takes into account whether a hypothesis class is undecided on a point x given a sample S.

$$C_{\mathcal{H}}(x, y, S, \delta) =$$

$$\begin{cases} 0 & \text{if there is } h \in \mathcal{H} \text{ such that } L_S(h) = 0 \text{ and } h(x) \neq y \\ 1 & \text{otherwise} \end{cases}$$

Observation 1. *For the family of distributions P that are realizable with respect to $\mathcal{H}$, $C_{\mathcal{H}}$ is indeed a confidence score fulfilling the no-overestimation guarantee for all instances.*

This observation was made in a different setup by El-Yaniv et al [20]. We note that our confidence score $C_{\mathcal{H}}$ is equivalent to their notion of consistent selective strategy. Using our terminology, they show that if the realizability assumption holds, if an instance (x, y) is classified as 1 by $C_{\mathcal{H}}$ then x is guaranteed to have true label y (with probability 1). Furthermore, their Theorems 11 and Theorem 14 as well as their Corollory 28 give rise to the following observation about confidence scores.

Observation 2. *It turns out that $nr_{\mathcal{P}}(C, m, \delta)$ for this confidence scoring rule $C_{\mathcal{H}}$ under the realizability assumption displays different behaviours for different classes (even when they have similar VC dimension):*

- *For some hypothesis classes, e.g. the class of thresholds on the real line $\mathcal{H}_{thres}$ or the class of linear separators in $\mathbb{R}^d$, $nr_{\mathcal{P}}(C, m, \delta)$ converges to 1 for every $\delta > 0$ as sample sizes go to infinity.*
- *On the other hand, for some hypothesis classes with finite VC-dimension $nr_{\mathcal{P}}(C, m, \delta) = 0$ for every sample size m and every $\delta < 1$. This phenomenon occurs for example for $\mathcal{H}$ being the class of singletons.*

For a more detailed analysis of which hypothesis classes have high non-redundancy, we refer the reader to [20], noting that our notion of non-redundancy is equivalent to worst-case coverage for a family of distributions.

We now look at the second case we wanted to address in this section: The family of probability distributions such that the approximation error of a class $\mathcal{H}$ is bounded by some ϵ. We fix a hypothesis class $\mathcal{H}$ and let $\mathcal{P}_{\mathcal{H}}$ be the family of all probability distributions P w.r.t. which there exists a classifier $h \in \mathcal{H}$ such that $L_P(h) \leq \epsilon$. We show that for for any (non-trivial) hypothesis class $\mathcal{H}$, it is not possible to find any satisfying confidence score for such a family.

Observation 3. *Let $\mathcal{H}$ be a hypothesis class for which there are two hypotheses $h_1, h_2 \in \mathcal{H}$ disagreeing on an infinite number of points, i.e. there exists a subset $X' \subset X$ with $|X'| = \infty$, such that for every $x \in X'$ we have $h_1(x) \neq h_2(x)$. Then there is no confidence score C fulfilling the no-overestimation guarantee for all positive-mass instances w.r.t. $\mathcal{P}_{\mathcal{H},\epsilon}$ that with high probability over the sample generation the confidence for no instance $x \in X$ is overestimated, which has non-redundancy $nr_{\mathcal{P}_{\mathcal{H},\epsilon}}(C, m, \delta) > 0$ for any $\delta \in (0, 1)$, $m \in \mathbb{N}$.*

This shows us that restricting ourselves to a family of probability distributions that allow for good generalization is not sufficient for allowing satisfying confidence scores. In the following section we will make stronger, more local assumptions and show that under these assumptions more satisfying confidence scores can be found.

5 Confidence scores under Lipschitz assumption

Lipschitz Assumption : We say that the probability distribution P over $X \times \{0, 1\}$ satisfies λ-Lipschitzness w.r.t. a metric $d(.,.)$ over X, if for every $x, x' \in X$, $|\ell_P(x) - \ell_P(x')| \leq \lambda d(x, x')$. When the domain is a subset of a Euclidean space, we will assume that d is the Euclidean distance unless we specify otherwise.

We provide an algorithm (1) to estimate the labelling probability of points using labelled samples. The algorithm partitions the space into cells. The algorithm outputs the same answer for points in the same cell. The input parameter r dictates the size of the cells. The algorithm estimates the average labelling probability for each cell. A confidence interval for this estimate is calculated based on the number of sample points in the cell. The interval is narrow when there are more sample points in the cell.

The following lemmas show how to estimate probability weights and average labelling probabilities of subsets of the domain:

Lemma 1. *Let P be a distribution over domain X. Let X' be a subset of X. Let S be an i.i.d. sample of size m drawn from the distribution P. Let $\hat{p}(X', S)$ be the fraction of the m samples that are in X'. For any $\delta > 0$, with probability $1 - \delta$ over the generation of the samples S,*

$$|P(X') - \hat{p}(X', S)| \leq w_p(m, \delta)$$

where

$$w_p(m, \delta) = \sqrt{\frac{1}{2m} \ln \frac{2}{\delta}}.$$

Lemma 2. *Let D be distribution over $X \times \{0, 1\}$. Let X' be a subset of X. Let S be an i.i.d. sample of size m drawn from D. Let $\hat{\ell}(X', S)$ be the fraction of the m labelled samples with label 1 in $S \cap X'$. For any $\delta > 0$, with probability $1 - \delta$ over the generation of the samples S, if $\hat{p}(X', S) - w_p(m, \delta/2) > 0$, then*

$$|\bar{\ell}_P(X') - \hat{\ell}(X', S)| < w_\ell(m, \delta, \hat{p}(X', S))$$

$$w_\ell(m, \delta, \hat{p}(X', S)) = \frac{1}{\hat{p}(X', S) - w_p(m, \delta/2)} \cdot \left(w_p(m, \delta/2) + \sqrt{\frac{1}{2m} \ln \frac{4}{\delta}} \right),$$

where $\hat{p}(X', S)$ is the fraction of the samples from S in X' that have label 1, $w_p(m, \delta/2))$ is as defined in Lemma 1.

Algorithm 1 Lipschitz labelling probability estimate

Input: Test point x, Labelled samples $S = (x_i, y_i)_{i=1}^m$,
Radius r, Estimation parameter δ,
Lipschitz constant λ
Output: Labelling probability estimate, confidence width of estimate
Split the domain $X = [0, 1]^d$ into a grid of $(1/r)^d$ hypercube cells each of side length r.
Find the cell $\mathrm{t_x}$ containing the test point x.
$\hat{p}[\mathrm{t_x}]$:= fraction of samples in $\mathrm{t_x}$.
$\hat{\ell}[\mathrm{t_x}]$:= fraction of samples in the cell $\mathrm{t_x}$ with label 1.
$w[\mathrm{t_x}] := 1$
if $\hat{p}[\mathrm{t_x}] - w_p(m, \delta/2)$ **then**
  $w[\mathrm{t_x}] = w_\ell(m, \delta, \hat{p}[\mathrm{t_x}])$
end if
Return $\hat{\ell}[\mathrm{t_x}], \min(1, w[\mathrm{t_x}] + r\lambda\sqrt{2})$

The following theorem shows that the true labelling probability of a point lies within the estimate interval provided by Algorithm 1, with high probability over the sample used to find the estimate.

Theorem 1. *Let the domain be $[0, 1]^d$. Suppose the data generating distribution P satisfies λ-Lipschitzness. For any $r > 0, \delta > 0$, $m \in \mathbb{N}$, for any $x \in [0, 1]^d$, define the confidence score based on Algorithm 1 as*

$$\hat{C}^{r,\lambda}_{\textit{Lipschitz}}(x, y; S, \delta) = \begin{cases} 1 - \hat{\ell}_S(x) - w_S(x) & \text{if } y = 1 \\ \hat{\ell}_S(x) - w_S(x) & \text{if } y = 0 \end{cases}$$

where $(\hat{\ell}_S(x), w_S(x))$ is the output of the Algorithm with input r, δ, λ. Then with probability $1 - \delta$ over samples S of size m,

$$\hat{C}^{r,\lambda}_{\textit{Lipschitz}}(x, y; S, \delta) \leq C_P(x, y)$$

We now show that as sample size increases, for an appropriately chosen input parameter r, Algorithm 1 returns narrow estimate intervals for the labelling probabilities for most points. This implies that for most points, the confidence score is not much lower than the true confidence $(2|\ell_P(x) - \frac{1}{2}|)$

Theorem 2. *For every λ-Lipschitz distribution, for every $\epsilon_x, \epsilon_c, \delta > 0$, there is a sample size $m(\epsilon_x, \epsilon_c, \delta)$ such that with probability $1 - \delta$ over samples S of size $m(\epsilon_x, \epsilon_c, \delta)$,*

$$\Pr_{x \sim P}[w_S(x) > \epsilon_c] < \epsilon_x$$

where w_S is the width of labelling probability estimate obtained from Algorithm 1 with input parameter of grid size $r = 1/m^{\frac{1}{8d}}$

6 Confidence score under adversarially robust Bayes classifier assumption

Recall that the Bayes classifier $h_P^B : X \to \{0,1\}$ assigns label 1 to a point x iff $\ell_P(x) > \frac{1}{2}$. Let $h : X \to \{0,1\}$ be a classifier.

Let $\mathcal{U} : X \to 2^X$ map domain points to their neighbour sets. Montasser, Hanneke and Srebro [23] define the adversarial robustness loss of a classifier h wrt the distribution P and $\mathcal{U}$ as

$$R_{\mathcal{U},P}(h) = \Pr_{(x,y)\sim P}\left[\sup_{z\in U(x)} \mathbb{1}[h(z) \neq y]\right].$$

We assume that the neighbour set of each point is a ball centered at that point. That is for some $\eta > 0$, for every $x \in X, \mathcal{U}(x) = \{x' \in X : ||x - x'|| \leq \eta\}$. This corresponds to the l_p-norm based notions of adversarial robustness in the literature [24]. We will denote the neighbour sets corresponding to a radius η by $\mathcal{U}_\eta : X \to 2^X$.

The adversarially robust Bayes assumption: The (η, ϵ_r)-adversarially robust Bayes assumption is that the distribution P and neighbour sets $\mathcal{U}_\eta(x)$ are such that the Bayes classifier h_P^B has adversarial robust loss less than some $\epsilon_r > 0$ i.e. $R_{\mathcal{U}_\eta,P}(h_P^B) \leq \epsilon_r$.

In this section, we analyze the implication of the assumption of the Bayes classifier being adversarially robust. We will show that this assumption allows us to determine, based on samples, the label of the Bayes classifier (Bayes label) for some test points. We show that there is value in using unlabelled data to make this determination. That is, collecting more unlabelled data without collecting more labelled data may allow us to determine the Bayes label for more points.

For any subset X' of the domain, let $\mathrm{diam}(X') = \max_{x_1,x_2\in X'} ||x_1 - x_2||$ denote the diameter of the subset. The following lemma, shows that if a set has small diameter and high enough probability weight, all points in it will be assigned the same label by the Bayes classifier

Lemma 3. *Under the (η, ϵ_r)-adversarially robust Bayes assumption, for any $X' \subseteq X$ such that $\mathrm{diam}(X') < \eta$ and $P_X(X') > 4\epsilon_r$, all points in X' are assigned the same label by h_P^B.*

Even if we do not have high probability weight for a low diameter set, if we know that the average labelling probability of the set is far from $\frac{1}{2}$, we can still conclude that all points in the set have the same Bayes label.

Lemma 4. *For any $X' \subset X$ with $\mathrm{diam}(X') < \eta$, let $\bar{\ell}_P(X')$ be the average labelling probability of the set X' i.e. $\bar{\ell}_P(X') = \Pr_{(x',y')\sim P}[y' = 1 | x' \in X']$. Under the (η, ϵ_r)-adversarially robust Bayes assumption, if $P_X(X') > \frac{2\epsilon_r}{\max(\bar{\ell}_P(X'), 1 - \bar{\ell}_P(x))}$, then all points in X' are assigned the same label by h_P^B.*

In Lemmas 1 and 2 we have seen how to estimate the probability weight and the average labelling probability of sets. Note that in order to estimate $P_X(X')$ one can settle for unlabeled samples. So if we estimate that a set X' has high probability weight. Or if we estimate that the labels of X' are homogeneous ie. the average labelling probability of X' is far from $\frac{1}{2}$, then we can conclude that the labels assigned by the Bayes optimal predictor h_P^B is the same for all points in X'. Furthermore, we may able be to estimate this label that the Bayes classifier assigns to all points in X' if we have enough labelled points in X'.

Theorem 3. *Let X' be a subset of X with $\mathrm{diam}(X') \leq \eta$. Let $\hat{p}(X', S_u, S_l)$ and $\hat{\ell}(X', S_l)$ be the estimates of the probability measure and average labelling probability of X'. $\hat{p}(X', S_u, S_l)$ is the fraction of the points in $S_u \cup S_l$ in X' and $\hat{\ell}(X', S_l)$ is the fraction of $S_l \cap X'$ that has label one. Under the (η, ϵ_r) adversarially robust Bayes assumption, if one of the two conditions hold:*

1. $\hat{p}(X', S_u, S_l) - w_p(m_u + m_l, \delta) > 4\epsilon_r$
2. $$\begin{aligned}\bar{\ell}(X', S_l) &:= \max(\hat{\ell}(X', S_l), 1 - \hat{\ell}(X', S_l)) \\ &> w_\ell(m, S_l, \hat{p}(X', S_l))\end{aligned}$$

   and

   $$\hat{p}(X', S_u, S_l) - w_p(m_u + m_l, \delta) > \frac{2\epsilon_r}{\bar{\ell}(X', S_l) - w_\ell(m, S_l, \hat{p}(X', S_l))},$$

where w_ℓ, w_p are as defined in Lemmas 2 and 1. Then with probability $1-2\delta$ over the generation of S_l, S_u, all points in X' have the same label assigned by the Bayes classifier.

Furthermore if $\left(\hat{\ell}(X', S_l) - w_\ell(m_l, \delta, \hat{p}(X', S_l)), \hat{\ell}(X', S_l) + w_\ell(m_l, \delta, \hat{p}(X', S_l))\right)$ does not contain $\frac{1}{2}$, then the label assigned to all points in X' by h_P^B is 1 if $\hat{\ell}(X', S_l) > \frac{1}{2}$ and 0 otherwise.

6.1 Additional assumptions for finding the majority label of a set

Next we consider the situation in which we have enough samples (labeled or unlabeled) to conclude that all points in a set of diameter at most η have the same Bayes label (namely the Bayes classifier assigns them all the same label). But we do not have enough labelled samples to deduce what that label is based on the robustness assumption we have used so far. We discuss some additional assumptions that suffice for reliably deducing that label.

6.1.1 Assumption1 : Access to a low error classifier

In this subsection we suppose to have access to a classifier h that we know to have true error less than ϵ_{err}. That is $L_P(h) < \epsilon_{\text{err}}$. For any subset $X' \subseteq X$, let us denote the by $h_P(X')$ the majority label h assigns to the set X'. That is $h_P(X') = \arg\max_{y\in\{0,1\}} P(\{x \in X' : h(x) = y\})$. We may estimate $h_P(X')$ using a sample S. We define $h_S(X') = \arg\max_{y\in\{0,1\}} |\{x \in X' \cap S : h(x) = y\}|$. We say that $h_S(X')$ is the majority label of the set X', based on the sample S. We analyse this case under the additional assumption that the labeling function is labeling function. In the following let $\mathcal{U}_{\frac{\eta}{2}}(x) = \{x' \in X : ||x - x'||_2 \leq \frac{\eta}{2}\}$

Lemma 5. *Let P be a distribution with deterministic labeling fulfilling the (η, ϵ_r)-adversarial robustness assumption (η, ϵ_r)-adversarial robustness assumption. Then every $z \in \mathcal{U}_{\frac{\eta}{2}}(x)$ in a high density region $P(\mathcal{U}_{\frac{\eta}{2}}(x)(x)) \geq 4\epsilon_r$, has the same label, $\ell_P(z) = \ell_P(x)$ for every $z \in \mathcal{U}_{\frac{\eta}{2}}(x)$.*

Lemma 6. *If for a hypothesis h and the underlying probability distribution P and a region $\mathcal{U}(x)$ around a point $x \in X$, all of the following conditions hold*

1. *The labeling function ℓ_P is deterministic*
2. *P fulfills the (η, ϵ_r)-adversarial robustness assumption.*
3. *If h has loss $L_P(h) < \epsilon_{err}$*
4. *one label in region $\mathcal{U}(x)$ is assigned at least $\epsilon = \max\{4\epsilon_r, 2\epsilon_{err}\}$ mass by h, i.e. there is $y \in \{0, 1\}$ with $P(\mathcal{U}(x) \cap h^{-1}(y)) \geq \epsilon$*

then the majority label of h on $\mathcal{U}_{\frac{\eta}{2}}(x)$ agrees with the true labeling, i.e., $h(\mathcal{U}_{\frac{\eta}{2}}(x)) = \ell_P(z)$ for every $z \in \mathcal{U}(x)$.

We now have a guarantee that holds for all instances of a high-density region if h has low loss. Now, if given access to samples from P, we can test whether requirements (3) and (4) from Lemma 6 hold. If h was learned by some algorithm A with a generalization guarantee we have a guarantee for (3) with high probability over the generation of the (labeled) input sample S_l. For (4) we can use an unlabeled sample S_u and the Hoeffding bound to give a guarantee.

Lemma 7. *For any $x \in X$ and any probability distribution Pany hypothesis h and any $\delta > 0$, we have that if $P(\mathcal{U}_{\frac{\eta}{2}}(x) \cap h^{-1}(y)) < \epsilon$*

$$Pr_{S_u \sim P_X^m}[\frac{|S_u \cap \mathcal{U}_{\frac{\eta}{2}}(x) \cap h^{-1}(y)|}{n} - \sqrt{\frac{\ln \frac{\delta}{2}}{n}} > \epsilon] < \delta$$

Thus, if given access to a hypothesis, fulfilling (3), we can define the following confidence score:

$$C_{h,\epsilon}(x, y, S_u, \delta) = \begin{cases} 1 & \text{, if } \frac{|S_u \cap \mathcal{U}_{\frac{\eta}{2}}(x) \cap h^{-1}(y)|}{n} - \sqrt{\frac{\ln \frac{\delta}{2}}{n}} > \epsilon \\ 0 & \text{otherwise} \end{cases}$$

Theorem 4. *Let $\epsilon = \max\{4\epsilon_r, 2\epsilon_{err}\}$For every $P \in \mathcal{P}_{\epsilon_r, \mathcal{U}, det}$ with $L_P(h) < \epsilon_{err}$ and every $x \in X$ we have*

$$Pr_{S_u \sim P_X^m}[C_{h,\epsilon}(x, y, S_u, \delta) > |y - \ell_P(x)|] < \delta.$$

This guarantee only holds for individual $x \in X$. In order to get a guarantee that with high probability over the sample generation we do not overestimate the confidence for any $x \in X$ we need to further tweak our confidence score. We need to ensure two define it in a way such that the union bounds works.

$$C_{h,\epsilon,\text{uniform}}(x, y, S_u, \delta) =$$

$$\begin{cases} 1 & \text{, if there is a } z \in \mathcal{U}_{\frac{\eta}{2}}(x) \cap S_u \text{ with } C_{h,\epsilon}(z, y, S_u \setminus \{z\}, \frac{\delta}{n}) = 1 \\ 0 & \text{otherwise} \end{cases}$$

Theorem 5. *For every $P \in \mathcal{P}_{\epsilon_r,\mathcal{U},det}$ with $L_P(h) < \epsilon_{err}$ we have that with probability of at least $1-\delta$ over $S_u \sim P_X^m$ for every $x \in X$*

$$C_{h,\epsilon,\textit{uniform}}(x, y, S_u, \delta) \leq |y - \ell_P(x)|$$

where $\epsilon = \max\{4\epsilon_r, 2\epsilon_{err}\}$

Now if instead of being given access to a hypothesis h with a deterministic guarantee $L_P(h) \leq \epsilon_{err}$, we have access to an algorithm A with a probabilistic error guarantee for $\mathcal{P}_{\epsilon_r,\mathcal{U},det}$, we can define a confidence score that incorporates that knowledge. That is, we have an algorithm A with guarantee $\delta(\cdot, \cdot)$ such that for any $\epsilon > 0$ and any sample size m, for every $P \in \mathcal{P}_{\epsilon_{err},\mathcal{U},det}$ we have $Pr_{S_l \sim P^m}[L_P(A(S_l) < \epsilon] \geq 1 - \delta(\epsilon, m)$

$$\begin{aligned} &C_{A,\epsilon,\text{uniform}}(x, y, S_u, S_l, \delta) \\ &\quad = C_{A(S_l),2\epsilon,\text{uniform}}(x, y, S_u, S_l, \delta - \delta(\epsilon, |S_l|)) \end{aligned}$$

Theorem 6. *Let A be an algorithm and $\delta_A(\epsilon, m)$ be a function, such that for every $\epsilon > 0$, $m \in \mathbb{N}$, and every $P \in \mathcal{P}_{\epsilon_r,\mathcal{U},det}$, we have*

$$Pr_{S_l \sim P^m}[L_P(A(S_l) < \epsilon] \geq 1 - \delta(\epsilon, m).$$

Then for any $\epsilon > 2\epsilon_r$, we have that with probability $1 - \delta$ over the generation of samples $S_u \sim P_X^{m_1}, S_l \sim P^{m_2}$ for every $x \in X$:

$$C_{A,\epsilon,\textit{uniform}}(x, y, S_u, S_l, \delta) < |y - \ell_P(x)|.$$

6.1.2 Assumption2 : Access to a proper PAC-learning learning algorithm

This section discusses the implications of knowing uniform convergence bounds of a function class and generalization bounds of a proper learning algorithm for the function class. We show that these assumptions allow us to determine the majority label the Bayes classifier assigns to some sets. Recall that the adversarially robust Bayes assumption could imply that all points in a set have the same Bayes label. Knowing the majority label for such a set means knowing the Bayes label of all points in that set.

We introduce a notion of decisiveness which captures how stable the learning algorithm is (over sampling) about the majority label it assigns to a set. For a set with lower probability weight, we can still deduce something about the majority label of the Bayes classifier on that set, when the algorithm is highly decisive about that set. In this case, we conclude that the majority label assigned by the learnt classifier is the majority label assigned by the Bayes classifier.

Definition 3 (Decisiveness of algorithm about majority of a set)**.** *We say that the decisiveness of an algorithm A, for a sample size m, about the majority label of a set $X' \subseteq X$ is*

$$DC_P^{A,m}(X') = 2\left|\Pr_{S \sim P^m}[A(S)(X') = 1] - 0.5\right|$$

Definition 4 (Correctness of algorithm about majority of a set)**.** *We say that the correctness of an algorithm A, for a sample size m, about the majority label of a set $X' \subseteq X$ is*

$$Cor_P^{A,m}(X') = \Pr_{S \sim P^m}[A(S)(X') = h_P^B(X')]$$

where h_P^B is the Bayes classifier

High decisiveness of an algorithm on a set does not normally imply high correctness of the algorithm on that set, even if the algorithm has good generalization. But this implication does hold if the probability weight of the set is sufficiently high. We will now relate the decisiveness and correctness of an algorithm on a set in terms of the probability weight of the set and the generalization guarantees of the algorithm.

Definition 5. *(Failure probability in terms of error) For a PAC learner A and a sample size m, define $\delta_P^{A,m} : (0, 1) \to (0, 1)$ such that for any $\epsilon > 0$,*

$$\delta_P^{A,m}(\epsilon) = \Pr_{S \sim P^m}[L_P(A(S)) > \epsilon]$$

Note that as ϵ increases, $\delta_P^{A,m}(\epsilon)$ decreases.

Theorem 7. *Let X' be a subset of X. If $\frac{1+DC_P^{A,m}(X')}{2} < \delta_P^{A,m}(P(X')/2)$, then $Cor_P^{A,m}(X') = \frac{1+DC_P^{A,m}(X')}{2}$.*

The above theorem shows a non-trivial lower-bound on the correctness of an algorithm for a region when the region has high probability mass ($P(X')$ being high implies $\delta^{A,m}(P(X'))$ is low) and when decisiveness of the region is high.

For a lower bound on correctness, we need lower bounds on decisiveness and probability weight of a subset. We have already seen how to get a lower bound for the probability weight using unlabelled samples (Lemma 1). Now we will see how to get a lower bound on decisiveness of a subset using samples.

Estimating decisiveness from samples:

We use the following sample based quantity to estimate the decisiveness of a proper agnostic PAC learning algorithm of a function class. We later show how to use the sample-based decisiveness along with uniform convergence bounds of the function class and generalization bounds of the proper learner to obtain a lower bound on the true decisiveness (Definition 3).

For a classifier $h : X \to \{0, 1\}$, for a region X', denote the average label according to the distribution P as

$$\bar{h}_P(X') = \mathbb{E}_{(x,y)\sim P}[h(x)|x \in X']$$

Denote the average label based on sample S as

$$\bar{h}_S(X') = \frac{|\{x \in X' \cap S : h(x) = 1\}|}{|X' \cap S|}$$

Definition 6 (Sample-based decisiveness of a class on majority of a set)**.** *Let $\mathcal{H}$ be a function class, and S_l and S_u be labelled and unlabelled samples respectively. We say that the γ-sample-based decisiveness of $\mathcal{H}$ based on S_l and S_u, about the majority label of X' is*

$$\hat{DC}^{\mathcal{H}}(X', S_l, S_u, \gamma) = \min_{h^{(0)}, h^{(1)} \in \mathcal{H}} \quad \left|L_{S_l}(h^{(1)}) - L_{S_l}(h^{(2)})\right|$$
$$\text{s.t.} \quad \bar{h}^{(0)}_{S_u \cup S_l}(X') < \frac{1}{2} - \gamma$$
$$\bar{h}^{(1)}_{S_u \cup S_l}(X') > \frac{1}{2} + \gamma$$

Now we introduce notation for the uniform convergence bounds of a function class ($\mathcal{H}$) and generalization bounds of s a proper agnostic-PAC learner (A) for a function class $\mathcal{H}$. Define functions $\delta^{\mathcal{H}}_{\text{UC}} : \mathbb{N} \times [0, 1] \to [0, 1]$ and $\delta^{A}_{\text{gen}} : \mathbb{N} \times [0, 1] \to [0, 1]$ that capture the uniform convergence of the class and the generalization of the algorithm respectively.

$$\delta^{\mathcal{H}}_{\text{UC}}(m, \epsilon) = \Pr_{S\sim P^m}[\exists h \in \mathcal{H} : |L_S(h) - L_P(h)| > \epsilon]$$
$$\delta^{A}_{\text{gen}}(m, \epsilon) = \Pr_{S\sim P^m}\left[L_P(A(S)) > \min_{h\in\mathcal{H}} L_P(h) + \epsilon\right]$$

We assume that for the learner A and the class $\mathcal{H}$, we know upper bounds on these functions for every m and ϵ.

Finally, the following theorem shows how to combine the sample-based decisiveness with the uniform convergence and generalization bounds to obtain a lower bound on the true decisiveness of the algorithm.

Theorem 8. *[Lower bound on true decisiveness using sample-based decisiveness] For any $X' \subseteq X$, for any $\delta > 0$, $m_l, m_u \in \mathbb{N}$, with probability $1 - \delta$ over generation of samples S_l, S_u, there exist $\alpha > 0$ and $\gamma > 0$ that can be calculated based on the samples, such that*

$$DC_P^{A,m}(x) \geq 1 - 2\delta^{A}_{gen}(m_l, \hat{DC}^{\mathcal{H}}(X', S_l, S_u, \gamma) + \alpha).$$

Since we have a sample-based lower bound on decisiveness and we have a sample-based lower bound on probability weights of sets (Lemma **??**), we can put these together to obtain a sample-based lower bound on the correctness of the algorithm about the majority of the Bayes classifier. Theorem 7 shows how to do this. We get non-trivial lower bounds on correctness when the sample-based decisiveness and probability weight of a subset is high.

References

[1] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To trust or not to trust a classifier. In *Advances in neural information processing systems*, pages 5541–5552, 2018.

[2] Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(59):1823–1840, 2008.

[3] Ming Yuan and Marten H. Wegkamp. Classification methods with reject option based on convex risk minimization. *J. Mach. Learn. Res.*, 11:111–130, 2010.

[4] Yoav Freund, Yishay Mansour, Robert E Schapire, et al. Generalization bounds for averaged classifiers. *The annals of statistics*, 32(4):1698–1722, 2004.

[5] Radu Herbei and Marten H Wegkamp. Classification with reject option. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 709–721, 2006.

[6] Adam Tauman Kalai, Varun Kanade, and Yishay Mansour. Reliable agnostic learning. *Journal of Computer and System Sciences*, 78(5):1481–1495, 2012.

[7] Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *J. Mach. Learn. Res.*, 11:1605–1641, 2010.

[8] Yair Wiener and Ran El-Yaniv. Agnostic pointwise-competitive selective classification. *Journal of Artificial Intelligence Research*, 52:171–201, 2015.

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

[10] Omar Montasser, Steve Hanneke, and Nathan Srebro. Vc classes are adversarially robustly learnable, but only improperly. *arXiv preprint arXiv:1902.04217*, 2019.

[11] Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl. Motivating the rules of the game for adversarial example research. *arXiv preprint arXiv:1807.06732*, 2018.

[12] Roman Vershynin. High-dimensional probability, 2019.

[13] Tom M. Mitchell. Version spaces: A candidate elimination approach to rule learning. In Raj Reddy, editor, *Proceedings of the 5th International Joint Conference on Artificial Intelligence. Cambridge, MA, USA, August 22-25, 1977*, pages 305–310. William Kaufmann, 1977.

Appendix

Useful lemmas

The following lemma appears as Theorem 2.2.6 of the book of [25], where the reader can find its proof.

Lemma 8 (Hoeffding's inequality for general bounded random variables)**.** *Let $X_1, \ldots, X_N$ be independent random variables. Assume that $X_i \in [m_i, M_i]$ for every i. Then, for any $t > 0$, we have*

$$\Pr\left[\sum_{i=1}^{N}(X_i - \mathbb{E}[X_i]) \geq t\right] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{N}(M_i - m_i)^2}\right).$$

Proof of Lemma 1. Let X_i be a random variable indicating if the i^{th} sample belongs to set X'. $X_i = 1$ if the i^{th} sample belongs to X' and zero otherwise. For each i, $\mathbb{E}[X_i] = P(X')$. $\hat{p}(X', S) = \frac{\sum_{i=1}^{N} X_i}{m}$. Applying Hoeffding's inequality, we get the inequality of the theorem. □

Proof of Lemma 2. Let X_i be a random variable such that

$$X_i = \begin{cases} 1 & \text{If } i^{\text{th}} \text{ sample belongs to the set } X' \text{ and has label one.} \\ 0 & \text{otherwise.} \end{cases}$$

$\mathbb{E}[X_i] = P(X')\bar{\ell}_P(X')$, for each i. $\sum_{i=1}^{m} X_i = m\hat{p}\hat{\ell}_P(X', S)$. Note that by triangle inequality,

$$\begin{aligned}
&|P(X')\hat{\ell}_P(X', S) - P(X')\bar{\ell}_P(X')| \\
&\leq |\hat{p}\hat{\ell}_P(X', S) - P(X')\bar{\ell}_P(X')| + |\hat{p} - P(X')|\hat{\ell}_P(X', S) \\
&\leq |\hat{p}\hat{\ell}_P(X', S) - P(X')\bar{\ell}_P(X')| + w_p.
\end{aligned}$$

For any $\epsilon > 0$,

$$\begin{aligned}
&\Pr[|\bar{\ell}_P(X') - \hat{\ell}(X', S)| > \epsilon] \\
&= \Pr[P(X') \cdot |\bar{\ell}_P(X') - \hat{\ell}(X', S)| > P(X')\epsilon] \\
&\leq \Pr[|\hat{p}\hat{\ell}_P(X', S) - P(X')\bar{\ell}_P(X')| + w_p > (\hat{p} - w_p)\epsilon] \\
&= \Pr[|m\hat{p}\hat{\ell}(X', S) - mP(X')\bar{\ell}_P(X')| > m(\hat{p} - w_P)\epsilon - w_p] \\
&= \Pr[\sum_{i=1}^{m} |X_i - \mathbb{E}[X_i]| > m((\hat{p} - w_P)\epsilon - w_p)]] \\
&\leq 2\exp\left(-2m((\hat{p} - w_P)\epsilon - w_p)^2\right) \qquad \text{(By Hoeffding's inequailty).}
\end{aligned}$$

When $\hat{p} - w_p > 0$, choosing

$$w_l(m, \delta, \hat{p}) > \frac{w_p}{\hat{p} - w_p} + \frac{1}{\hat{p} - w_p}\sqrt{\frac{1}{2m}\ln\frac{4}{\delta}},$$

we get that with probability $1 - \delta$, $|\bar{\ell}_P(X') - \hat{\ell}(X', S)| < w_\ell(m, \delta, \hat{p})$. □

Confidence scores for hypothesis classes

We start by recalling the definition of confidence scores fulfilling the no-overestimation guarantee for all instances:

Definition 1 (Confidence Score, fulfilling the no-overestimation guarantee for all instances)**.** *We say a function C, that takes as input a sample S, a point x, a hypothesis h and a parameter δ and outputs a value in $[0, 1]$. We say such a function C is a* confidence score fulfilling the no-overestimation guarantee for all instances *for a family of probability functions $\mathcal{P}$ if for every $P \in \mathcal{P}$ the probability over $S \sim P^m$ that there exists $x \in X$*

$$Pr_{y\sim \text{Bernoulli}(\ell_P(x))}\left[h(x) = y\right] < C(x, y, S, \delta)$$

is less than δ.

6.2 Proof of Observation 1

Observation 1. *For the family of distributions P that are realizable with respect to $\mathcal{H}$, $C_{\mathcal{H}}$ is indeed a confidence score fulfilling the no-overestimation guarantee for all instances.*

We need to show that $C_{\mathcal{H}}$ fulfills Definition 1, that is, we need to show that for every hypothesis class $\mathcal{H}$ and every distribution P that fulfills the realizable condition w.r.t. $\mathcal{H}$, the probability over $S \sim P^m$ that there exists $x \in X$

$$Pr_{y\sim \text{Bernoulli}(\ell_P(x))}\left[h(x) = y\right] < C_{\mathcal{H}}(x, y, S, \delta)$$

is less than δ. Since $C_{\mathcal{H}}$ only assigns values 0 and 1 and the condition is trivially fulfilled for instances with confidence score 0, we will now only discuss the case where $C_{\mathcal{H}}$ assigns confidence score 1. Recall, that $C_{\mathcal{H}}$ only assigns confidence 1 if and only if there is no $h \in \mathcal{H}$ with $L_S(h) = 0$ and $h(x) \neq y$. Since we know, that realizability holds, we know that $\ell_P \in \mathcal{H}$ and since S is an i.i.d sample from P we know $L_S(\ell_P) = 0$. Now let $C_{\mathcal{H}}(x, y, S, \delta) = 1$, then by definition we know that $L_S(h) = 0$ implies $h(x) = y$. Thus $\ell_P(x) = y$. Thus, this confidence score does not overestimate the confidence of a point in any label.

6.3 Proof of Observation 2

Observation 2. *It turns out that $nr_{\mathcal{P}}(C, m, \delta)$ for this confidence scoring rule $C_{\mathcal{H}}$ under the realizability assumption displays different behaviours for different classes (even when they have similar VC dimension):*

- *For some hypothesis classes, e.g. the class of thresholds on the real line $\mathcal{H}_{thres}$ or the class of linear separators in $\mathbb{R}^d$, $nr_{\mathcal{P}}(C, m, \delta)$ converges to 1 for every $\delta > 0$ as sample sizes go to infinity.*

- *On the other hand, for some hypothesis classes with finite VC-dimension $nr_{\mathcal{P}}(C, m, \delta) = 0$ for every sample size m and every $\delta < 1$. This phenomenon occurs for example for $\mathcal{H}$ being the class of singletons.*

We start by noting that our definition of the confidence score $C_{\mathcal{H}}$ is equivalent to the consistent selective strategy from [20]. In order to state their definition, we will first need to introduce some other concepts. First, we will state the definition of version space from [26].

Definition 7 (Version Space)**.** *Given a hypothesis class $\mathcal{H}$ and a labeled sample S, the* version space $\mathcal{H}_S$ *the set of all hypotheses in $\mathcal{H}$ that classify S correctly.*

Now, we can define the agreement set and maximal agreement set as in [20].

Definition 8 (agreement set)**.** *Let $\mathcal{G} \subset \mathcal{H}$. A subset $X \subset X$ is an* agreement set *with respect to $\mathcal{G}$ if all hypotheses in $\mathcal{G}$ agree on every instance in X', namely for all $g_1, g_2 \in \mathcal{G}$, $x \in X$*

$$g_1(x) = g_2(x).$$

Definition 9 (maximal agreement set)**.** *Let $\mathcal{G} \subset \mathcal{H}$. The* maximal agreement set *with respect to $\mathcal{G}$ is the union of all agreement sets with respect to $\mathcal{G}$.*

We can now state the definition of consistent selective strategy. Note, that a selective strategy is defined by a pair (h, g) of a classification function h and a selective function g. For our purposes, we will only need to look at the selective function g.

Definition 10 (consistent selective strategy (CSS))**.** *Given a labeled sample S, a* consistent selective strategy (CSS) *is a selective classification strategy that takes h to by any hypothesis in $\mathcal{H}_S$ (i.e., a consistent learner), and takes a (deterministic) selection function g that equals one for all points in the maximal agreement set with respect to $\mathcal{H}_S$ and zero otherwise.*

We now see that for any $\mathcal{H}$ and any labeled sample S the selected function g assigns one to x, if for every two $h_1, h_2 \in H$ with $L_S(h_1) = L_S(h_2) = 0$ implies $h_1(x) = h_2(x)$. Thus, $g(x) = C_{\mathcal{H}}(x, h(x), S, \delta)$ for any $h(x) \in \mathcal{H}_S$. In [20] the selection function is then analysed with respect to its *coverage*, which is defined by $\phi(h, g) = \mathcal{E}_{x\sim P}[g(x)]$ for a given distribution P. Note, that our notion of non-redundancy can be seen as worst-case coverage over the family $\mathcal{P}_{\mathcal{H},0}$ of distributions realizable distributions. We can now use some of their results to show our observation.

Theorem 9 (non-achievable coverage, Theorem 14 from [20])**.** *Let m and $d > 2$ be given. There exist a distribution P, an infinite hypothesis class $\mathcal{H}$ with a finite VC-dimension d, and a target hypothesis in $\mathcal{H}$, such that $\phi(h, g) = 0$ for any selective classifier (h, g), chosen from $\mathcal{H}$ by CSS using a training sample S of size m drawn i.i.d. according to P.*

This directly implies the second part of our observation. For a more concrete example consider the class of singletons over the real line $\mathcal{H}_{singleton} = \{h_z : \mathbb{R} \to \{0, 1\} : h_z(x) = 1 \text{ if and only if} z = x\}$. Now consider any sample size m. We can construct a probability distribution P_n such that the marginal P_X is uniform over some finite set $X' \subset X$ and such that $|X'| = n$ and such that the labeling function is realizable in $\mathcal{H}_{singleton}$. Then, $\mathbb{E}_{x\sim P_{n,X}, S\sim P_n^m}[\max\{C_{\mathcal{H}_{singleton}}(x, 1, S, \delta), C_{\mathcal{H}_{singleton}}(x, 0, S, \delta)\}] \leq \frac{m}{n}$. It is clear that we can construct such a distribution P_n for any $n \in \mathbb{N}$. Therefore, since $\lim_{n\to\infty} \frac{m}{n} = 0$, we get $nr_{\mathcal{P}_{\mathcal{H},0}}(C_{\mathcal{H}}, m, \delta) = \inf_{P\in\mathcal{P}} \mathbb{E}_{x\sim P_X, S\sim P^m}[\max\{C_{\mathcal{H}_{singleton}}(x, 1, S, \delta), C_{\mathcal{H}_{singleton}}(x, 0, S, \delta)\}]$.

Let us consider the class of thresholds $\mathcal{H}_{thres} = \{h_a : \mathcal{R} \to \{0, 1\}, h_a(x) = 1 \text{ if and only if } x > a\}$. We can define the following two learning rules for thresholds:

$$A_1(S) = h_{a_1}, \text{ where } a_1 = \operatorname*{arg\,max}_{x_i\in\mathbb{R}:(x_i,0)\in S} x_i$$

$$A_2(S) = \bar{h}_{a_2}, \text{ where } \bar{h}_{a_2}(x) = 1 \text{ if and only if } x \geq a_2 := \operatorname*{arg\,min}_{x_i\in\mathbb{R}:(x_i,1)\in S} x_i.$$

Furthermore, The way $C_{\mathcal{H}_{thres}}$ is defined, for any S and $\delta \in (0, 1)$ we have $A_1(S)(x) = A_2(S)(x)$ if and only if there is $y \in \{0, 1\}$ with $C_{\mathcal{H}_{thres}}(x, y, S, \delta) = 1$. Furthermore, both of these learning rules are empirical risk minimizers for $\mathcal{H}_{thres}$ in the realizable case. Thus both of them are PAC-learners. Thus for any ϵ, δ, there is a $m(\epsilon, \delta)$, such that for any $m \geq m(, \delta)$ and any $P \in \mathcal{P}_{\mathcal{H}_{thres},0}$,

$$1 - \delta \leq Pr_{S\sim P^m, x\sim P}[A_1(S) = A_2(S)] = Pr_{S\sim P^m, x\sim P}[\max\{C_{\mathcal{H}_{thres}}(x, 1, S, \delta), C_{\mathcal{H}_{thres}}(x, 0, S, \delta)\}]$$

Thus $\lim_{m\to\infty} nr_{\mathcal{P}_{\mathcal{H}_{thres},0}}(C_{\mathcal{H}thres,m,\delta}) = 1$ for any $\delta \in (0, 1)$.

6.4 Proof of Observation 3

Observation 3. *Let $\mathcal{H}$ be a hypothesis class for which there are two hypotheses $h_1, h_2 \in \mathcal{H}$ disagreeing on an infinite number of points, i.e. there exists a subset $X' \subset X$ with $|X'| = \infty$, such that for every $x \in X'$ we have $h_1(x) \neq h_2(x)$. Then there is no confidence score C fulfilling the no-overestimation guarantee for all positive-mass instances w.r.t. $\mathcal{P}_{\mathcal{H},\epsilon}$ that with high probability over the sample generation the confidence for no instance $x \in X$ is overestimated, which has non-redundancy $nr_{\mathcal{P}_{\mathcal{H},\epsilon}}(C, m, \delta) > 0$ for any $\delta \in (0, 1)$, $m \in \mathbb{N}$.*

For every $\epsilon > 0$, every $x \in X'$ and every $n \in \mathbb{N}$ with $m > \frac{1}{\epsilon}$, we can construct a distribution $P_{x,n}$, such that $l_{P_{x,n}}(x') = h_1(x')$ for every $x' \in X \setminus \{x\}$ and $h_1(x') \neq l_{P_{x,n}}(x')$ and such that the marginal $P_{x,n,X}$ is uniform over some $X_n \subset X'$ with $|X_n| = n$. For a sample to distinguish between two such distributions $P_{x_1,n}$ and $P_{x_2,n}$ either the point x_1 or x_2 needs to be visited by the sample. Thus in order to give a point-wise guarantee for all instances with positive mass, only points in the sample can be assigned a positive confidence in this scenario. Thus any confidence score fulfilling this guarantee would have $\mathbb{E}_{S\sim P^m_{x,n}, x\sim P_{x,n,X}}[\max\{C(x, 1, S, \delta), C(x, 0, S, \delta)\}] \leq \frac{m}{n}$. Thus, $nr(C, m, \delta) = 0$ for any such score.

Confidence scores using Lipschitz assumption

Proof of Theorem 1. The algorithm partitions the space into r^d cells. Let p_c be the probability weight of a cell c and let $\hat{p}_c$ be the estimate of p_c that is calculated based on a sample to be the fraction of sample points in the cell c. From Lemma 1 and a union bound, we know that with probability $1 - \frac{\delta}{2}$, for every cell c,

$$p_c \in [\hat{p}_c - w_p(c), \hat{p}_c + w_p(c)].$$

Here $w_p(c) = w_p(m, \delta/2r^d)$ (as defined in Lemma 1).

The algorithm also estimates the average label of a cell c - ℓ_c as $\hat{\ell}_c$. This is the fraction of the sample point in the cell that have the label one. This is the same as the labelling probability estimate defined in Lemma 2. When the true probability weights of cells lie within the calculated confidence interval, by Lemma 2, we know that with probability $1 - \frac{\delta}{2}$, for every cell c,

$$\hat{\ell}_c \in [\hat{\ell}_c - w_\ell(c), \hat{\ell}_c - w_\ell(c))].$$

Here $w_\ell(c) = w_\ell(m, \delta/2r^d, \hat{p}_c)$ (as defined in Lemma 2).

The maximum distance between any two points in any cell is $r\sqrt{2}$. By the λ-Lipshitz, any point in the cell has labelling probability within $\lambda r\sqrt{2}$ of the average labelling probability of the cell. Therefore, with probability $1 - \delta$, for each cell c, for every point x in the cell c, the labelling probability of x satisfies:

$$\ell_P(x) \in [\hat{\ell}_c - w_\ell(c) - \lambda r\sqrt{2}, \hat{\ell}_c + w_\ell(c) + \lambda r\sqrt{2}].$$

This is the interval returned by the algorithm. Now we lower bound true confidence based on the confidence interval of the labelling probability. For a point x, let $c(x)$ denote the cell containing the point.

$$\begin{aligned} C_P(x, 0) &= \ell_P(x) \\ &\geq \hat{\ell}_{c(x)} - w_\ell(c(x)) - \lambda r\sqrt{2} \\ C_P(x, 1) &= 1 - \ell_P(x) \\ &\geq 1 - \hat{\ell}_{c(x)} - w_\ell(c(x)) - \lambda r\sqrt{2}. \end{aligned}$$

□

Proof of Theorem 2. We choose the input to the algorithm to be $r = \frac{1}{m^{1/8d}}$. With probability $1 - \frac{\delta}{2}$, for all cells with probability weight greater than $\gamma = \frac{1}{m^{1/4}}$, the length of the confidence interval of the labelling probability is less than

$$\begin{aligned} &\frac{\frac{1}{m^{1/2}}}{\frac{1}{m^{1/4}} + \frac{1}{m^{1/2}}} - \frac{1}{\frac{1}{m^{1/4}} - \frac{1}{m^{1/2}}}\sqrt{\frac{1}{2m}\ln\frac{4m^{1/8}}{\delta}} + \frac{\lambda\sqrt{2}}{m^{1/8}} \\ &\leq \frac{1}{m^{1/4} - 1} + \frac{1}{m^{1/4} - 1}\sqrt{\frac{1}{16}\ln\frac{4m}{\delta}} + \frac{\lambda\sqrt{2}}{m^{1/8}}. \end{aligned}$$

This quantity decreases with increase in m and converges to zero. Therefore, for every $\epsilon_c > 0$, there is $M_1(\epsilon_c, \delta)$ such that this interval is less than ϵ_c. When sample size is larger than $M_1(\epsilon_c, \delta)$, with probability $1 - \frac{\delta}{2}$, the size of confidence intervals for labelling probabilities of cells with weights greater than $\gamma = \frac{1}{m^{1/4}}$, is smaller than ϵ_c.

The points for which we can't say anything about the interval lengths are points in cells with weight at most γ. The total weight of such points is at most $\gamma \frac{1}{r^d} = \frac{1}{m^{1/8}}$. For any $\epsilon_x > 0$, let $M_2(\epsilon_x)$ be such that $\frac{1}{M_2(\epsilon_x)^{1/8}} < \epsilon_x$.

Choosing a sample size M greater than $M_1(\epsilon_c, \delta)$ and $M_2(\epsilon_x)$, we get that

$$\Pr_{S \sim P^M}[w_\ell > \epsilon_c] < \epsilon_x.$$

□

Adversarially robust Bayes assumption

Proof of Lemma 3. Suppose the Bayes classifier (h_P^B) takes both values zero and one for points in a set $X' \subset X$ with $\mathrm{diam}(X') < \eta$. Let $\bar{y} \in \{0, 1\}$ be the Bayes label of the majority of the set ie. $P_X(\{x \in X' : h_P^B(x) = \bar{y}\}) > P(\{x \in X' : h_P^B(x) \neq \bar{y}\})$. The adversarial loss in $\mathcal{U}(x)$ is at least

$$\begin{aligned}
&P(\{(x', y') : x' \in X', h_P^B(x') = (\bar{y}), y' = \bar{y}\}) \\
&\geq \frac{1}{2} P(\{(x', y') : x' \in X', h_P^B(x') = (\bar{y})\}) \\
&\geq \frac{1}{4} P(X').
\end{aligned}$$

Since the adversarial robust loss is at most ϵ_r, we get that if the Bayes label is not the same for all points in the neighbour set X', then $P(X') \leq 4\epsilon_r$. □

Proof of Lemma 4. Suppose the Bayes classifier (h_P^B) takes values both zero and one for points in a set $X' \subset X$ with $\mathrm{diam}(X') < \eta$. Then the adversarial robust loss of the Bayer classifier is at least

$$\begin{aligned}
&\max_{y \in \{0,1\}} P(\{(x', y') : x' \in X', y' = y\}) \\
&\geq \max_{y \in \{0,1\}} \Pr_{(x', y') \sim P}[y' = y | x' \in X'] \cdot P(X') \\
&\geq P(X') \cdot \max(\bar{\ell}_P(X'), 1 - \bar{\ell}_P(X')).
\end{aligned}$$

Therefore if h_P^B does not have the same label for all points in X', then the adversarial robust loss of the Bayes classifier is at least $\max(\bar{\ell}_P(X'), 1 - \bar{\ell}_P(X')$. If we know that the adversarial robust loss is at most ϵ_r, then any set X' with $\mathrm{diam}(X') < \eta$ that has weight at least $\frac{\epsilon_r}{\max(\bar{\ell}_P(\mathcal{U}(x)), 1 - \bar{\ell}_P(\mathcal{U}(x)))}$ has the same label by the Bayes classifier for all points in X'. □

Access to low error classifier assumption

6.5 Proof of Lemma 5

Lemma. *Let P be a distribution with deterministic labeling fulfilling the (η, ϵ_r)-adversarial robustness assumption (η, ϵ_r)-adversarial robustness assumption. Then every $z \in \mathcal{U}_{\frac{\eta}{2}}(x)$ in a high density region $P(\mathcal{U}_{\frac{\eta}{2}}(x)(x)) \geq 4\epsilon_r$, has the same label, $\ell_P(z) = \ell_P(x)$ for every $z \in \mathcal{U}_{\frac{\eta}{2}}(x)$.*

First we note that $diam(\mathcal{U}_{\frac{\eta}{2}}(x)) = \eta$. The lemma directly follows from Lemma 3 and the fact that in the case of a deterministic labeling the Bayes classifier h^* equals the labeling function ℓ_P.

6.6 Proof of Lemma 6

Lemma. *If for a hypothesis h and the underlying probability distribution P and a region $\mathcal{U}(x)$ around a point $x \in X$, all of the following conditions hold*

1. *The labeling function ℓ_P is deterministic*
2. *P fulfills the (η, ϵ_r)-adversarial robustness assumption*

3. *If h has loss $L_P(h) < \epsilon_{err}$*
4. *one label in region $\mathcal{U}_{\frac{\eta}{2}}(x)$ is assigned at least $\epsilon = \max\{4\epsilon_r, 2\epsilon_{err}\}$ mass by h, i.e. there is $y \in \{0,1\}$ with $P(\mathcal{U}_{\frac{\eta}{2}}(x) \cap h^{-1}(y)) \geq \epsilon$*

then the majority label of h on $\mathcal{U}_{\frac{\eta}{2}}(x)$ agrees with the true labeling, i.e., $h(\mathcal{U}_{\frac{\eta}{2}}(x)) = \ell_P(z)$ for every $z \in \mathcal{U}(x)$.

We can first use 1.) and 2.) to conclude, using Lemma 5 that for $P(\mathcal{U}_{\frac{\eta}{2}}(x)) \geq 4\epsilon_r \geq \epsilon$ and for every $z \in \mathcal{U}_{\frac{\eta}{2}}(x)$ we have $\ell_P(x) = \ell_P(z)$. Furthermore, since h has error smaller than ϵ, h needs to agree with ℓ_P on at least half the mass of $\mathcal{U}_{\frac{\eta}{2}}(x)$. Thus the majority label of h needs to agree with ℓ_P.

6.7 Proof of Lemma 7

Lemma. *For any $x \in X$ and any probability distribution P any hypothesis h and any $\delta > 0$, we have that if $P(\mathcal{U}_{\frac{\eta}{2}}(x) \cap h^{-1}(y)) < \epsilon$*

$$Pr_{S_u \sim P_X^m}[\frac{|S_u \cap \mathcal{U}_{\frac{\eta}{2}}(x) \cap h^{-1}(y)|}{n} - \sqrt{\frac{\ln \frac{\delta}{2}}{n}} > \epsilon] < \delta$$

Let $X_1, \ldots, X_n$ be i.i.d. random variables drawn from distribution P. We define the following random variable dependent on X_i:

$$Z_i = \begin{cases} 1 & \text{if } X_i \in \mathcal{U}_{\frac{\eta}{2}}(x) \cap h^{-1}(y) \\ 0 & \text{otherwise} \end{cases}$$

We note, that $Z_1, \ldots, Z_n$ are i.i.d. Bernoulli random variables drawn from $Bernoulli(P(\mathcal{U}_{\frac{\eta}{2}}(x) \cap h^{-1}(y)))$. Furthermore, we note that $\mathbb{E}_{Z_i \sim Bernoulli(p)}[Z_i] = p$. Thus the Hoeffding bound gives us

$$Pr_{S_u \sim P_X^m}[\frac{|S_u \cap \mathcal{U}_{\frac{\eta}{2}}(x) \cap h^{-1}(y)|}{n} - P_X(\mathcal{U}_{\frac{\eta}{2}}(x) \cap h^{-1}(y)) > t] < \exp(\frac{-2t^2}{n}).$$

For $\mathcal{U}_{\frac{\eta}{2}}(x)$ with $P_X(\mathcal{U}_{\frac{\eta}{2}}(x) \cap h^{-1}(y))$ this implies the statement of the lemma.

6.8 Proof of Theorem 4

Theorem. *Let $\epsilon = \max\{4\epsilon_r, 2\epsilon_{err}\}$For every $P \in \mathcal{P}_{\epsilon_r, \mathcal{U}, det}$ with $L_P(h) < \epsilon_{err}$ and every $x \in X$ we have*

$$Pr_{S_u \sim P_X^m}[C_{h,\epsilon}(x, y, S_u, \delta) > |y - \ell_P(x)|] < \delta.$$

First, recall the definition of confidence score $C_{h,\epsilon}$:

$$C_{h,\epsilon}(x, y, S_u, \delta) = \begin{cases} 1 & \text{, if } \frac{|S_u \cap \mathcal{U}_{\frac{\eta}{2}}(x) \cap h^{-1}(y)|}{n} - \sqrt{\frac{\ln \frac{\delta}{2}}{n}} > \epsilon \\ 0 & \text{otherwise} \end{cases}$$

The proof of this theorem will be split in two parts:

1. If $P_X(\mathcal{U}_{\frac{\eta}{2}}(x) \cap h^{-1}(y)) > \epsilon$, then $\ell_P(x) = y$
2. If $P_X(\mathcal{U}_{\frac{\eta}{2}}(x) \cap h^{-1}(y)) < \epsilon$, then $Pr_{S_u \sim P_X}[C_{h,\epsilon}(x, y, S_u, \delta) = 1] < \delta$.

1. is shown by Lemma 6 and 2. is shown by Lemma 7. This implies the theorem.

6.9 Proof of Theorem 5

Theorem. *For every $P \in \mathcal{P}_{\epsilon_r, \mathcal{U}, det}$ with $L_P(h) < \epsilon_{err}$ we have that with probability of at least $1 - \delta$ over $S_u \sim P_X^m$ for every $x \in X$*

$$C_{h,\epsilon,\textit{uniform}}(x, y, S_u, \delta) \leq |y - \ell_P(x)|$$

where $\epsilon = \max\{4\epsilon_r, 2\epsilon_{err}\}$

First, recall the definition of confidence score $C_{h,\epsilon,uniform}$:

$$C_{h,\epsilon,\text{uniform}}(x,y,S_u,\delta) =$$

$$\begin{cases} 1 & \text{, if there is a } z \in \mathcal{U}_{\frac{n}{2}}(x) \cap S_u \text{ with } C_{h,\epsilon}(z,y,S_u \setminus \{z\}, \frac{\delta}{n}) = 1 \\ 0 & \text{otherwise} \end{cases}$$

From Theorem 4 we know that for a given x, y the probability over the sample generation of S'_u that $C_{h,\epsilon}(x,y,S'_u,\frac{\delta}{n}) > |y - \ell_P(x)|$ is bounded by $\frac{\delta}{n}$. Thus,

$$Pr_{S_u \sim P_X^m}[\exists z \in X : C_{h,\epsilon,uniform}(z,y,S_u,\delta) > |y - \ell_P(x)|]$$

$$= Pr_{S_u \sim P_X^m}[\exists x \in S_u : C_{h,\epsilon}(x,y,S_u \setminus \{x\}, \frac{\delta}{n}) > |y - \ell_P(x)|]$$

$$\leq \sum_{x_i \in S_u}^{n} Pr_{S_u^i \sim P_X^{m-1}}[C_{h,\epsilon}(x_i,y,S_u^i,\frac{\delta}{n}) > |y - \ell_P(x)|] \leq n \cdot \frac{\delta}{n} = \delta$$

6.10 Proof of Theorem 6

Theorem. *Let A be an algorithm and $\delta_A(\epsilon, m)$ be a function, such that for every $\epsilon > 0$, $m \in \mathbb{N}$, and every $P \in \mathcal{P}_{\epsilon_r,\mathcal{U},det}$, we have*

$$Pr_{S_l \sim P^m}[L_P(A(S_l) < \epsilon] \geq 1 - \delta(\epsilon, m).$$

Then for any $\epsilon > 2\epsilon_r$, we have that with probability $1 - \delta$ over the generation of samples $S_u \sim P_X^{m_1}, S_l \sim P^{m_2}$ for every $x \in X$:

$$C_{A,\epsilon,\textit{uniform}}(x,y,S_u,S_l,\delta) < |y - \ell_P(x)|.$$

Recall the following definition

$$C_{A,\epsilon,\text{uniform}}(x,y,S_u,S_l,\delta)$$
$$= C_{A(S_l),2\epsilon,\text{uniform}}(x,y,S_u,S_l,\delta - \delta(\epsilon,|S_l|))$$

Now we can use the definition of A and Theorem 5 to get the following:

$$Pr_{S_u \sim P_X^{m_1}, S_l \sim P^{m_2}}[\exists x \in X : C_{A,\epsilon,\text{uniform}}(x,y,S_u,S_l,\delta) > |y - \ell_P(x)|]$$

$$\leq Pr_{S_u \sim P_X^{m_1}, S_l \sim P^{m_2}}[A(S_l) = h \text{ with } L_P(h) \leq \epsilon \text{ and } \exists x \in X : C_{A,\epsilon,\text{uniform}}(x,y,S_u,S_l,\delta) > |y - \ell_P(x)|]$$

$$+ Pr_{S_u \sim P_X^{m_1}, S_l \sim P^{m_2}}[[A(S_l) = h \text{ with } L_P(h) > \epsilon \text{ and } \exists x \in X : C_{A,\epsilon,\text{uniform}}(x,y,S_u,S_l,\delta) > |y - \ell_P(x)|]$$

$$\leq Pr_{S_u \sim P_X^{m_1}, S_l \sim P^{m_2}}[[A(S_l) = h \text{ with } L_P(h) \leq \epsilon \text{ and } \exists x \in X : C_{h,2\epsilon,\text{uniform}}(x,y,S_u,\delta - \delta(\epsilon,m_2)) > |y - \ell_P(x)|]$$

$$+ Pr_{S_u \sim P_X^{m_1}, S_l \sim P^{m_2}}[[A(S_l) = h \text{ with } L_P(h) > \epsilon \text{ and } \exists x \in X : C_{h,2\epsilon,\text{uniform}}(x,y,S_u,\delta - \delta(\epsilon,m_2)) > |y - \ell_P(x)|]$$

$$\leq Pr_{S_u \sim P_X^{m_1}, S_l \sim P^{m_2}}[[A(S_l) = h \text{ with } L_P(h) \leq \epsilon \text{ and } \exists x \in X : C_{h,2\epsilon,\text{uniform}}(x,y,S_u,\delta - \delta(\epsilon,m_2)) > |y - \ell_P(x)|]$$

$$+ Pr_{S_l \sim P^{m_2}}[L_P(A(S_l)) > \epsilon] \leq \delta - \delta(\epsilon,m_2) + \delta(\epsilon,m_2) = \delta$$

Access to proper PAC learning algorithm assumption

Proof of Theorem 7. Let $\bar{y} = \text{argmax}_{y \in \{0,1\}} \Pr_{S \sim P^m}[A(S)(X') = y]$ be the label that the learner is more prone to assign as the majority label of the set X'. By the definition of decisiveness, the learner assigns $\bar{y}$ as majority of X' with probability $\frac{1+DC_P^{A,m}(X')}{2}$. When $\bar{y} = h_P^B(X')$, $Cor_P^{A,m}(X') = \frac{1+DC_P^{A,m}(X')}{2}$.

When the learner assigns X' a majority label that is different from the majority label of h_P^B on the set X', the learner makes error at least $P(X')/2$. By the generalization property of the learner, the probability of the learner making error more than $P(X')/2$ is at most $\delta_P^{A,m}(P(X')/2)$.

Therefore, if $\frac{1+DC_P^{A,m}(X')}{2} < \delta_P^{A,m}(P(X')/2)$, then $\bar{y} = h_P^B(X')$ and $Cor_P^{A,m}(X') = \frac{1+DC_P^{A,m}(X')}{2}$. □

Proof of Theorem 8. Given $X' \subseteq X$, and labelled and unlabelled samples S_u, S_l, let

$$\mathcal{H}^0(X', S_u, S_l) = \{h \in \mathcal{H} : \bar{h}_{S_u \cup S_l}(X') < \frac{1}{2} - \gamma\}$$
$$\mathcal{H}^1(X', S_u, S_l) = \{h \in \mathcal{H} : \bar{h}_{S_u \cup S_l}(X') > \frac{1}{2} + \gamma\}$$
$$\mathcal{H}^0_P(X') = \{h \in \mathcal{H} : h_P(X') = 0\}$$
$$\mathcal{H}^1_P(X') = \{h \in \mathcal{H} : h_P(X') = 1\}.$$

There are $|(S_u \cap S_l) \cap X'|$ sample points in X'. With probability $1 - \frac{\delta}{3}$, the number of sample of sample points in X' is at least $m(X') = |(S_u \cap S_l) \cap X'| - (m_l + m_u) \cdot w_p(m_u + m_l, \delta/3)$. With probability $1 - \delta^{\mathcal{H}}_{\text{UC}}(m(X'), \gamma)$, for all $h \in \mathcal{H}$, $|\bar{h}_{S_u \cup S_l}(X') - \bar{h}_P(X')| < \gamma$. This implies that the majority of h on X' according to the distribution P is the same as the majority of h on sample points in X'. Therefore with probability $1 - \delta^{\mathcal{H}}_{\text{UC}}(m(X'), \gamma)$, $\mathcal{H}^0(X', S_u, S_l) = \mathcal{H}^0_P(X')$ and $\mathcal{H}^1(X', S_u, S_l) = \mathcal{H}^1_P(X')$

By the definition of $\hat{DC}^{\mathcal{H}}(X', S_l, S_u, \gamma)$, for all $h \in \mathcal{H}_P^{1-A(S_c)}(X')$,

$$|L_{S_l}(h) - L_{S_l}(A(S_l))| \geq \hat{DC}^{\mathcal{H}}(X', S_l, S_u, \gamma).$$

By UC, for any $\alpha > 0$, with probability $1 - \delta^{\mathcal{H}}_{\text{UC}}(m_l, \alpha)$ over generation of labelled samples S_l, for all $h \in \mathcal{H}_P^{1-A(S_l)}(X')$,

$$\begin{aligned} L_P(h) - L_P(A(S_l)) + 2\alpha &\geq L_{S_l}(h) - L_{S_l}(A(S_l)) \\ &\geq \hat{DC}^{\mathcal{H}}(X', S_l, S_u, \gamma) \end{aligned}$$

$$\begin{aligned} L_P(h) &\geq L_P(A(S_l)) - 2\alpha + \hat{DC}^{\mathcal{H}}(X', S_l, S_u, \gamma) \\ &\geq \min_{h' \in H} L_P(h') - 2\alpha + \hat{DC}^{\mathcal{H}}(X', S_l, S_u, \gamma). \end{aligned}$$

By the generalization property of A,

$$\begin{aligned} &\Pr_{S \sim P^{m_l}}[A(S)(x) \neq A(S_l)(x)] \\ &= \Pr_{S \sim P^m}\left[A(S) \in \mathcal{H}_P^{1-A(S_l)}(X')\right] \\ &\leq \Pr_{S \sim P^m}\Bigg[L_P(A(S)) \geq \min_{h' \in H} L_P(h') - 2\alpha \\ &\quad + \hat{DC}^{\mathcal{H}}(X', S_l, S_u, \gamma)\Bigg] \\ &\leq \delta^A_{\text{gen}}\left(m, 2\alpha - \hat{DC}^{\mathcal{H}}(X', S_l, S_u, \gamma)\right). \end{aligned}$$

$DC^{A,m_l}_P(X') \geq \Pr_{S \sim P^{m_l}}[A(S)(x) \neq A(S_l)(x)]$. Therefore, with probability $1 - \frac{\delta}{3} - \delta^{\mathcal{H}}_{\text{UC}}(m(X'), \gamma) - \delta^{\mathcal{H}}_{\text{UC}}(m_l, \alpha)$, the lower bound on true decisiveness based on sample decisiveness holds. To ensure that the failure probability is at most δ, we could choose α so that $\delta^{\mathcal{H}}_{\text{UC}}(m_l, \alpha) < \frac{\delta}{3}$ and choose γ so that $\delta^{\mathcal{H}}_{\text{UC}}(m(X'), \gamma) < \frac{\delta}{3}$. □