Identifying Regions of Trusted Predictions

Nivasini Ananthakrishnan[1] **Shai Ben-David**[1, 2] **Tosca Lechner**[1] **Ruth Urner**[3]

[1]David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, N2L 3G1, Canada
[2]Vector Institute, Toronto, ON, M5G 1M1, Canada
[3]Lassonde School of Engineering, EECS Department, York University, Toronto, ON, M3J 1P3, Canada

Abstract

Quantifying the probability of a label prediction being correct on a given test point or a given subpopulation enables users to better decide how to use and when to trust machine learning derived predictors. In this work, combining aspects of prior work on conformal predictions and selective classification, we provide a unifying framework for confidence requirements that allows for distinguishing between various sources of uncertainty in the learning process as well as various region specifications. We then consider a set of common prior assumptions on the data generating process and show how these allow learning justifiably trusted predictors.

1 INTRODUCTION

Quantifying the certainty in the output of a predictor is important for instilling (and justifying) trust in decision making that is based on machine learning. Standard (statistical) techniques for ensuring and measuring the quality of a learned predictor fall short of providing reliable and easily interpretable notions of confidence for specific predictions. Bayesian statistical tools often come with confidence scores on predictions. However, these rely on having chosen a good prior and are easily misinterpreted by users that are not well-versed in Bayesian decision making. On the other end of the spectrum, PAC-type learning theoretic frameworks are designed to provide general, ideally assumption-free guarantees. They ensure low mistake probability over the data-generating process. However, arguably, such a promise can be void when called to provide confidence in the predictions on specific instances or specific sub-regions of the space. In this work, we provide a (non-Bayesian, PAC-inspired) framework for learning predictors that come with instance or region-wise guarantees. While much of the earlier work in the PAC-inspired setup (Shafer and Vovk [2008], Lei and Wasserman [2014], Lei [2014], Foygel Barber et al. [2020]) took a distribution-free approach, the insight that drives our investigations is that confidence in any prediction of unknown information inherently relies on prior domain knowledge. In a PAC-type setup, such knowledge can be expressed as a restriction on the data generating process and a suitable choice of hypothesis class. We examine the problem of confidence in predictions under several common types of such assumptions.

Our setup can be viewed as inspired by two lines of research of similar aim: as in the framework of *conformal predictions* (Shafer and Vovk [2008]) our confidence-instilling predictors provide *coverage sets* (subsets of the output space) for the possible labeling of instances. And as the framework of *selective classification* or *learning with abstentions* [Bartlett and Wegkamp, 2008, Yuan and Wegkamp, 2010, Freund et al., 2004, Herbei and Wegkamp, 2006, Kalai et al., 2012], we distill out a trade-off between the *validity* of the provided prediction (in the case of coverage sets, a prediction is valid, if the coverage set includes the true target) and the *non-triviality* of such a coverage-set-predictor (validity can be trivially achieved by outputting the full set of possible targets; a coverage set therefore should only be considered useful if on many instances the coverage set is a singleton or at least sufficiently small).

In this work we consider binary classification tasks, and provide a unifying framework for confidence requirements that allows for distinguishing between various sources of uncertainty as well as various region specifications. Sources of uncertainty in statistical learning include the randomness of the chosen training sample, the randomness in the choice of a test-point, as well as the stochasticity in the label generation at some instance. To account specifically for the latter, we introduce coverage set learning not only for the labels in the classification task, but also for the conditional labeling function (CLF). A user may require confidence in predictions over the whole domain, only for a specific sub-region, a collection of such regions or specific points. We model this by defining notions of domain-wide, region-wide or

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

point-wise validity and non-triviality.

Finally, we provide a variety of (standard) scenarios where the success requirements of our framework can be realized. We present successful CLF-coverage set learners under assumption of the CLF satisfying a Lipschitz condition for user specified regions. Under some mixture model scenarios we identify scenarios of domain-wide successful CLF-learning. Additionally, we show how to identify regions for successful label-coverage set predictors under an assumption low approximation error by some hypothesis class.

1.1 RELATED WORK

Quantification of confidence in predictions are often derived in Bayesian learning setups. Such quantification inevitably rely on the quality of and confidence in the priors applied in the Bayesian reasoning framework [Barber, 2012]. In this work, we take a non-Bayesian perspective and therefore focus on discussing prior work that also developed notions of confidence in statistical learning theoretic setups. There is one recent paper that developed confidence scores in a non-Bayesian framework [Jiang et al., 2018]. The theoretical results in that study differ from our work. It suggests one algorithmic approach to generating confidence scores. The validity of these confidence scores relies on several technical assumptions on the data-generating process. In this work, we take a step back and aim at developing a general framework for the meaning and validity of confidence in learned prediction and then provide several concrete scenarios where such confident predictions can be derived. The two lines of prior work that are most relevant to our setup are learning *conformal predictors* [Lei and Wasserman, 2014, Vovk, 2013, Foygel Barber et al., 2020] and the PAC-type framework of *learning with abstention* or *selective classification* [Bartlett and Wegkamp, 2008, Yuan and Wegkamp, 2010, Freund et al., 2004, Herbei and Wegkamp, 2006, Kalai et al., 2012]. We here briefly outline how our work differs from existing literature in these setups. An extended discussion can be found in Section B in the appendix.

As we do in this work, *conformal mappings* also provide coverage sets (instead singleton predictions), that is, regions in the label space that are guaranteed, with high probability, to contain the true label value. In brief, the conformal mappings literature differs from our work in several key aspects: In most setups the probability there is the joint probability over the training data and the probability over a newly arriving test-point. Most guarantees there are distribution-free, requiring only that the data is exchangable (the common i.i.d. assumption is a special case of exchangeablility). Furthermore, most of the conformal prediction literature considers online-settings and focuses on multiclass classification or regression. In our setup, we distinguish the randomness that comes from the sampling of the training set and the randomness that comes from sampling a new instance. Furthermore, we consider a binary classification setting. In addition to analysing point-wise and region-specific guarantees for label coverage sets we also propose the use of coverage sets for the conditional labelling function (CLF) instead of the label itself. Some prior work on conformal prediction also explores guarantees conditioned on subsets or elements of the domain [Lei and Wasserman, 2014, Vovk, 2013, Foygel Barber et al., 2020]. The probabilities here, however, are still aggregated over the generation of the training set and on the randomness of the instance to be classified. Several studies have provided impossibility results for distribution free point-wise or general region-wise guarantees [Vovk, 2013, Lei and Wasserman, 2014, Foygel Barber et al., 2020]. The latter study poses the question of whether the impossibility of non-trivial point-wise guarantees or region-specific guarantees for greater collections of subsets can be overcome by additional distributional assumptions. We explicitly address this question here.

In *selective classification*, a classifier is allowed to abstain from making a prediction. Many works in this line provide accuracy guarantees that hold with high probability over the domain [Bartlett and Wegkamp, 2008, Yuan and Wegkamp, 2010, Freund et al., 2004, Herbei and Wegkamp, 2006, Kalai et al., 2012]. However, in contrast to our work, that line of work generally does not aim at point-wise or region-wise guarantees, does not consider learning of the conditional labeling rule and does not distinguish uncertainty from training data or in the labeling rule. Some point-wise guarantees are provided in earlier work [El-Yaniv and Wiener, 2010, Wiener and El-Yaniv, 2015]. The former study gave a theoretical analysis of the selective classification setup in which a classification function and a selective function are learned simultaneously [El-Yaniv and Wiener, 2010]. They also study the trade-off between (in our terminology) validity and non-triviality and develop an optimal learning strategy for learning classifiers that are perfectly valid. However, those results are derived under the rather restrictive realizability assumption and thus do not allow for stochasticity in the labeling rule as our setup and analysis does.

We instantiate our notions under three different types of standard assumptions on the data-generating process: Access to a hypothesis class that has low approximation error, Lipschitzness of the CLF and a generative (for example Gaussian) mixture mode. Low approximation error is a standard assumption in statistical learning theory (e.g., Shalev-Shwartz and Ben-David [2014]). Smooth behaviour of the CLF (such as Lipschitzness and related notions) is commonly assumed in non-parametric learning setups, for example nearest neighbor type learning (Shakhnarovich et al. [2008]). Learning of mixtures, Gaussian mixtures specifically, has been extensively studied in terms of parameter estimation [Kwon and Caramanis, 2020, Moitra and Valiant, 2010], classification [Li et al., 2017] and density estimation [Ashtiani et al., 2020], see also appendix Section B.

1.2 SUMMARY OF CONTRIBUTIONS

In this work, we study a variety of distributional assumptions under which we can identify regions (these could be the full space, a set of sub-regions of the space or a collection of points) where valid and non-trivial coverage set predictions can be learned. We demonstrate how these scenarios allow for identification of regions for trusted predictions. Additionally, in several of these setups we also show how unlabeled data can be employed to improve the non-triviality of our learned predictors. Our contributions can be summarized as follows:

- **Formal framework for coverage set learnability** We adapt notions of coverage set learning for classification and learning of CLF-functions and introduce a PAC-like framework of learning success. Our definitions allow for distinguishing between the various sources of uncertainty (the training data, its relation to the test point and stochasticity in label-generation given a point), and allow for various types of regions where trusted predictions may be required (domain-wide, region-wise, point-wise). Our definitions further make explicit the trade-off between validity and non-triviality. We then instantiate our notions under three different scenarios for the data-generating process.
- **Lipschitzness of the CLF** The first scenario that we consider is that the CLF-function of the data-generating process satisfies a Lipschitz condition. This is a standard assumption in non-parametric learning settings. We present a successful CLF-coverage learning algorithm that achieves point-wise validity and domain-wide non-triviality (with coverage intervals that decrease with the size of the input sample).
- **Low approximation error by a hypothesis class** H Under prior knowledge of a learnable hypothesis class of low approximation error, given a collection of regions, we show how to construct coverage sets satisfying region-conditional validity with respect to that collection. We show how to identify regions that allow for non-trivial validity. More specifically, we show that identifying regions that have sufficient probability mass or are areas of *high decisiveness* (a novel notion that we introduce) of the class H suffices for region-wise validity and non-triviality guarantees. Further, we demonstrate that these can be identified with the use of unlabeled data.
- **Mixture models** We show how the problem of valid and non-trivial CLF-coverage set learning can be reduced to CLF-learning in L_1-distance and also to the notion of p-concept learning. We then show that the problem of constructing coverage sets with domain validity and domain non-triviality can be reduced to the problem of learning the positive and negative components of a mixture model in total variation distance.
- **Reductions** We also systematically analyze the (information theoretic) difficulties of various related learning problems. We show that sample complexities of binary classification, coverage-set learning, and marginal distribution learning are in strictly increasing order.

2 SETUP

We use a standard learning theoretic setup. We let $\mathcal{X}$ denote some domain or feature space and $\mathcal{Y} = \{0, 1\}$ a binary label space. We assume that data is generated by a probability distribution P over $\mathcal{X} \times \mathcal{Y}$, let $l_P(x) = \mathbb{P}_{(X,Y)\sim P}[Y = 1|X = x]$ denote the corresponding *conditional labeling function (CLF)* (a real valued function) and $P_{\mathcal{X}}$ denote the corresponding marginal distribution over the domain $\mathcal{X}$. A *hypothesis* or *classifier* is a function $h : \mathcal{X} \to \mathcal{Y}$ and a *hypothesis class* H is a set of hypotheses. In a standard learning setting, a *learner* $\mathcal{A}$ takes in a sequence $S = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$ of labeled domain points and outputs a hypothesis $h = \mathcal{A}(S)$. The quality of prediction of a hypothesis h on sample (x, y) is measured by a *loss function* ℓ. For classification tasks we typically use the *binary loss*

$$\ell^{0/1}(h, x, y) = \mathbb{1}\left[h(x) \neq y\right].$$

The goal for the learner is to output a hypothesis h of low *expected loss* $\mathcal{L}_P^{0/1}(h) = \mathbb{E}_{(X,Y)\sim P}\left[\ell^{0/1}(h, X, Y)\right]$ over the data-generating distribution. We let $\mathcal{L}_S^{0/1}(h)$ denote the *empirical loss* with respect to data S (that is, the expected loss with respect to the uniform distribution over S).

For a distribution P over $\mathcal{X} \times \{0, 1\}$, we let h_P^* denote the *Bayes classifier*, that is the classifier with minimal expected binary loss with respect to P. We have $h_P^*(x) = 1$ if $l_P(x) \geq 1/2$ and $h_P^*(x) = 0$ otherwise. For a hypothesis class H, we let $\mathrm{opt}_P(H) = \inf_{h\in H}\mathcal{L}_P^{0/1}(h)$ denote the *approximation error* of the class H.

In our setting, we would like to learn functions that output sets of labels (that are aimed to contain the true labels), rather than single values. A label-coverage-hypothesis is a function $\mathrm{c} : \mathcal{X} \to \{\{0\}, \{1\}, \{0, 1\}\}$.

Definition 1 (Label Coverage Set Learner)**.** *A* label coverage set learner $\mathcal{A}$ *takes as input a labelled training set* $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ *and outputs a label-coverage-hypothesis.*

We are also interested in learning functions that can provide coverage guarantees for the the conditional labeling function. A CLF-coverage-hypothesis is a function $\mathrm{r} : \mathcal{X} \to \{[a, b] : a \leq b \in [0, 1]\}$

Definition 2 (CLF-Coverage Set Learner)**.** *A* CLF-coverage set learner $\mathcal{A}$ *takes as input a labelled training set* $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ *and outputs a CLF-coverage-hypothesis.*

We use *coverage set*, *coverage hypothesis* and *coverage set learner* as umbrella terms for the label and CLF-coverage set learning settings. Success for a coverage set hypothesis is a combination of two competing requirements. Firstly, we would like the output set for a domain point x to be a *valid coverage*, in the sense that it contains the true/observed label (or the true conditional label probability in the case of CLF learning). This requirement however can be trivially met by a coverage set hypothesis that always outputs the full set of options (all of $\mathcal{Y}$ in the case of label coverage or the full interval $[0, 1]$ in the case of CLF-coverage). Such a hypothesis would be valid everywhere, however at the same time pretty useless. To provide meaningful information, we need to additionally require that the coverage hypothesis, on a substantial portion of the space, outputs a small set of of options. For label coverage, we will require a coverage set to be a singleton to be considered meaningful, while for CLF-coverage we will require the output to be a short interval. Below, we formalize these notions of *validity* and *non-triviality* requirements.

For validity requirements, we will distinguish three levels: We may require that the output coverage sets are valid over the full domain (with high probability), conditioned on being in a region or point-wise.

Definition 3 (Validity)**.** *Let* c *and* r *denote a label and a CLF-coverage set hypotheses, respectively. Let* P *be a distribution over* $\mathcal{X} \times \mathcal{Y}$ *and* $\alpha > 0$ *a confidence parameter.*

- *We say the coverage set hypothesis satisfies* α-domain-validity *(are* α*-domain-valid) with respect to* P *if we have*

$$\mathbb{P}_{(X,Y)\sim P}[Y \in \mathrm{c}(X)] \geq \alpha \quad \textit{and}$$
$$\mathbb{P}_{X\sim P_{\mathcal{X}}}[l_P(X) \in \mathrm{r}(X)] \geq \alpha$$

  respectively.

- *For a subset* $B \subseteq \mathcal{X}$ *of the domain, we say that they satisfy* α-region-conditional validity in B *with respect to* P *if we have*

$$\mathbb{P}_{(X,Y)\sim P}[Y \in \mathrm{c}(X)|X \in B] \geq \alpha \quad \textit{and}$$
$$\mathbb{P}_{X\sim P}[l_P(X) \in \mathrm{r}(X)|X \in B] \geq \alpha$$

  respectively.

- *We say that the label coverage hypothesis* c *satisfies* α-point-wise validity at point $x \in \mathcal{X}$ *with respect to* P*, if we have*

$$\mathbb{P}_{(Y\sim P(Y|x))}[Y \in \mathrm{c}(x)] \geq \alpha$$

  and we say that CLF-coverage hypothesis r *satisfies* point-wise validity at $x \in \mathcal{X}$ *if* $l_P(x) \in \mathrm{r}(x)$.

For a collection $\mathcal{B} \subseteq 2^{\mathcal{X}}$, we also speak of *region-wise validity for* $\mathcal{B}$ if the above condition holds for all regions $B \in \mathcal{B}$. Similarly, we simply refer to *point-wise validity* if the above condition holds for (almost) all $x \in \mathcal{X}$.

Similarly, non-triviality can be required (with high probability) over the full domain or conditioned on sub-regions of interest. For the output interval $[a, b] = \mathrm{r}(x) \subseteq [0, 1]$ of a CLF-coverage function, we let $\mu([a, b]) = |b - a|$ denote the length of the output interval. While a label coverage output would be considered non-trivial if contains a unique label, this is too strong a requirement for CLF-coverage function. For the latter, we introduce an additional parameter γ corresponding to a to bound on the length of an interval that would be considered a non-trivial prediction.

Definition 4 (Non-triviality)**.** *Let* c *and* r *be label- and CLF-coverage set hypotheses,* P *be a distribution over* $\mathcal{X} \times \mathcal{Y}$*,* $\beta > 0$ *a confidence parameter and* $\gamma > 0$ *a length-tolerance parameter. We say that* c *satisfies* β-domain-non-triviality *with respect to* P*, if*

$$\mathbb{P}_{X\sim P_{\mathcal{X}}}[\mathrm{c}(X) \neq \{0, 1\}] \geq \beta.$$

and that r *has* (β, γ)-domain-non-triviality *if*

$$\mathbb{P}_{X\sim P_{\mathcal{X}}}[\mu(\mathrm{r}(X)) \leq \gamma] \geq \beta.$$

Analogously to validity, non-trivality can also be defined for a specified region $B \subseteq \mathcal{X}$ *by using the appropriate conditional probabilities.*

These quality criteria for coverage set hypotheses give rise to the following notion of success for a coverage set learner:

Definition 5 ((α, β, δ)- and $(\alpha, \beta, \gamma, \delta)$-successful coverage set learning)**.** *Let* $\mathcal{P}$ *be a class of distributions. A label coverage set learner* $\mathcal{A}$ *is* domain wide (α, β, δ)-successful *for* $\mathcal{P}$ *if for all triples of parameters* $(\alpha, \beta, \delta) \in (0, 1]^3$*, there exists an* $m(\alpha, \beta, \delta)$ *such that for all* $m \geq m(\alpha, \beta, \delta)$ *and all* $P \in \mathcal{P}$ *the probability over the generation of an i.i.d.* S *of size* m *that* $\mathcal{A}(S)$ *is* α*-domain-valid and* β*-domain-non-trivial is greater than 1-* δ*.*

Analogously, we can define region successful *(with respect to a collection of regions* $\mathcal{B} \subseteq 2^{\mathcal{X}}$*) and* point-wise successful (α, β, δ) label coverage set learners*. Additionally, we can analogously phrase the requirements for CLF-coverage set learner to be* $(\alpha, \beta, \gamma, \delta)$*-successful (*$(\beta, \gamma, \delta)$*-successful in the case of point-wise CLF-coverage learning) by adding an the additional tolerance parameter* γ*.*

The *sample complexity* of domain wide/region/point wise coverage set learning is the (point-wise) smallest function for which there exists a learner $\mathcal{A}$ satisfying the above definition.

We note that the standard PAC-learning setup can be viewed as an extreme case of (α, β, δ)-successful learning. Here, a standard hypothesis that outputs just one label for every point is required to be correct with confidence $(1-\delta)$ except

for an error allowance of ϵ. Thus the output is everywhere non-trivial, and our notion of validity corresponds to the usual notion of accuracy. The trade-off between domain-validity and non-triviality has been discussed earlier for a PAC-type setting of "selective classification" (classification with a reject option) [El-Yaniv and Wiener, 2010]. While the original PAC framework incorporated requirements on computational complexity of the learning algorithms [Valiant, 1984], in this work we focus on its component of statistical (sample) complexity, as is common [Shalev-Shwartz and Ben-David, 2014].

We also note that label coverage and CLF coverage are related. Given a label coverage hypothesis we can construct a CLF coverage hypothesis and vice versa. The label coverage hypothesis constructed from the CLF coverage hypothesis is close to optimal in terms of validity/non-triviality if the CLF coverage hypothesis fulfills point-wise validity and has good levels of non-triviality. For a more detailed discussion we refer the reader to the appendix.

3 LIPSCHITZNESS

In this section, we assume that the generating distribution satisfies Lipschitzness, which we define below. We also assume that the domain $\mathcal{X}$ is $[0,1]^d$

Definition 6. *A distribution P over $\mathcal{X} \times \{0,1\}$ satisfies λ-Lipschitzness for $\lambda > 0$, with respect to a metric $d(.,.)$ over $\mathcal{X}$ if for every $x, x' \in \mathcal{X}$, $|l_P(x) - l_P(x')| \leq \lambda d(x, x')$.*

Under the assumption that the generating distribution is Lipschitz and that an upper bound on the Lipschitz constant λ is known, we provide a CLF-coverage learner (Algorithm 1) for which we show the strongest validity and non-triviality guarantees – point-wise validity and domain non-triviality. We also identify conditions on points that lead to more narrow CLF-coverage sets.

The CLF-coverage learner is defined as Algorithm 1. This algorithm partitions the domain into cells. The input parameter r to the algorithm determines the size of the cells. For each cell t, the algorithm then calculates the average label $\hat{\ell}[\mathrm{t}]$ of samples in the cell. This is an estimate of the expected label conditioned upon membership in the cell. The algorithm calculates a confidence interval (of width $w[\mathrm{t}]$) for this estimate, based on the number of samples in the cell. The confidence interval is more narrow for cells containing many samples. The algorithm assigns all points in a cell the same CLF-coverage. The CLF coverage for a point x contained in a cell t_x is an interval centered at $\hat{\ell}[\mathrm{t}_x]$ and having width $w[\mathrm{t}_x] + r\lambda\sqrt{2}$.

Theorem 1 now states the point-wise non-triviality guarantee of the CLF-coverage sets provided by Algorithm 1.

Algorithm 1 Lipschitz CLF-coverage learner

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 t_x containing the test point x.

$\hat{p}[\mathrm{t}_x]$:= fraction of samples in t_x.

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

$w_\ell(m,\delta,\hat{p}[\mathrm{t}_x]) := \frac{2w_p(m,\delta/2)}{\hat{p}[\mathrm{t_x}]-w_p(m,\delta/2)}$

$\hat{\ell}[\mathrm{t}_x]$:= fraction of samples in the cell t_x with label 1.

$w[\mathrm{t_x}] := 1$

if $\hat{p}[\mathrm{t}_x] - w_p(m,\delta/2) > 0$ **then**

$w[\mathrm{t}_x] := w_\ell(m,\delta,\hat{p}[\mathrm{t}_x])$

end if

$$I_{S,r,\lambda}(x) := \Big(\max(0, \hat{\ell}[\mathrm{t}_x] - w[\mathrm{t}_x] - r\lambda\sqrt{2}), \min(1, \hat{\ell}[\mathrm{t}_x] + w[\mathrm{t}_x] + r\lambda\sqrt{2})) \Big)$$

Return $I_{S,r,\lambda}(x)$

Namely, the CLF-coverage set for x is the interval $I_{S,r,\lambda}(x)$ centered at $\hat{\ell}[\mathrm{t}_x]$ with width $2(w[\mathrm{t}_x] + r\lambda\sqrt{2})$.

Theorem 1. *Let the domain be $[0,1]^d$. Suppose the data generating distribution P satisfies λ-Lipschitzness. For any $r > 0, \delta > 0$, with probability at least $1-\delta$ over the generation of the sample S, Algorithm 1 with input parameters S, r, δ, λ yields a CLF-coverage set having point-wise validity (see definition 3).*

We now show in Theorem 2 that as sample size increases, for an appropriately chosen input parameter r, Algorithm 1 returns a CLF-coverage hypothesis with large domain non-triviality. Theorem 2 shows that for large enough sample sizes, most domain points have CLF-coverage sets with small widths.

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 at least $1-\delta$ over samples S of size $m(\epsilon_x, \epsilon_c, \delta)$, Algorithm 1 with input parameters $S, r = 1/m^{\frac{1}{8d}}, \delta, \lambda$ yields a CLF-coverage set having $(1-\epsilon_c, 1-\epsilon_x)$-domain non-triviality (see definition 4).*

Further, we note that Algorithm 1 runs in time polynomial in the sample complexity. With standard arguments (replacing the grid type partition in Algorithm 1 with a more efficient

coverage with cells of small diameter), it can be seen that the sample complexity's dependence on the dimension d can be replaced with the intrinsic dimension of data generating distribution's support.

4 FUNCTION CLASS WITH LOW APPROXIMATION ERROR

We assume that the function class H has approximation error – $\text{opt}_P(H)$ less than ϵ_{approx}. That is, we know that $\min_{h\in H} \mathcal{L}_P^{0/1}(h) \leq \epsilon_{\text{approx}}$. Under this assumption, given a collection $\mathcal{B}$ of subsets of the domain and given a target α parameter of validity, we construct a semi-supervised label coverage set learner based on labelled samples S_l and unlabelled samples S_u, drawn i.i.d. from the underlying distribution and the underlying marginal distribution respectively. We show that, with high probability over the sample generation, the learner yields a label coverage hypothsis with α-region-conditional validity relative to $\mathcal{B}$. We identify conditions on regions that yield non-trivial coverage-sets.

Let $h_H(S_l)$ denote an empirical risk minimizer, from the class H, for the sample S_l. That is, $h_H(S_l) \in \text{argmin}_{h\in H}\mathcal{L}_{S_l}^{0/1}(h)$. The error of a classifier h w.r.t. the distribution P, conditioned upon membership in a set $B \subseteq \mathcal{X}$ is defined as

$$\mathcal{L}_{P|B}^{0/1}(h) = \mathbb{P}_{(X,Y)\sim P}[h(X) \neq Y | X \in B].$$

We start by showing how to obtain an upper bound on $\mathcal{L}_{P|B}^{0/1}(h_H(S_l))$, for a given $B \subseteq \mathcal{X}$. We later show how to use the region-conditional generalization bound to find valid label coverage sets for the collection $\mathcal{B}$. Regions with low conditional generalization errors (lower than the validity parameter $1-\alpha$) are given non-trivial coverage sets. We identify niceness conditions for regions that allow for low conditional generalization bounds and hence non-trivial coverage sets. We also show how to test if these conditions are satisfied using S_l and S_u.

The first condition that allows for low conditional-generalization bounds is high probability weight of the region. For any $B \subseteq \mathcal{X}$, we can use the unlabelled data (S_u) to estimate its probability weight. The following theorem shows how to obtain a region-conditional generalization bound for a region B based on the fraction of samples that lie in B. Here, a larger fraction of points in the region leads to a smaller generalization bound.

Theorem 3. *For every $B \subseteq \mathcal{X}$, for any classifier $h : \mathcal{X} \to \{0,1\}$, let*

$$\mathcal{L}_{S_l,B}^{0/1}(h) = \frac{|(x,y) \in S_l : x \in B, h(x) \neq y|}{|S_l|}.$$

For any $\delta > 0$, with probability at least $1-\delta$ over the generation of S_l and S_u, if $\frac{|S_u \cap B|}{|S_u|} > \sqrt{\frac{1}{2|S_u|}\ln\frac{4}{\delta}}$, then

$$\mathcal{L}_{P|B}^{0/1}(h_H(S_l)) \leq \frac{\mathcal{L}_{S_l,B}^{0/1}(h_H(S_l)) + \epsilon_{UC}(|S_l|,\delta/2)}{\frac{|S_u\cap B|}{|S_u|} - \sqrt{\frac{1}{2|S_u|}\ln\frac{4}{\delta}}}.$$

Here $\epsilon_{UC}(|S_l|,\delta/2) = C\sqrt{\frac{VCdim(H)+\log(2/\delta)}{|S_l|}}$ for a universal constant C.

We now define another sample-dependent property of regions that results in low region-conditional generalization error bounds. We call this the *decisiveness* of the function class on the subset. We say that the function class H is decisive on a set $B \subseteq \mathcal{X}$, based on S_u and S_l, if all classifiers in H with low empirical error on S_l, label the points in $S_u \cap B$ similarly. For a set with probability weight too low to get non-trivial conditional generalization bounds by using Theorem 3, we can still get non-trivial bounds when the set has high decisiveness.

Definition 7 (Disagreement between classifiers in a region)**.** *We define the disagreement between two classifiers $h_1, h_2 : \mathcal{X} \to \mathcal{Y}$ in a set $B \subseteq X$ as*

$$\Delta_P(h_1,h_2,B) = \mathbb{P}_{X\sim P_{\mathcal{X}}}[h_1(X) \neq h_2(X), X \in B].$$

We empirically estimate the disagreement of classifiers in B, using S_u as

$$\Delta_{S_u}(h_1,h_2,B) = \frac{|\{x \in S_u \cap B : h_1(x) \neq h_2(x)\}|}{|S_u|}.$$

Definition 8 (Decisiveness of function class in a region)**.** *For any $\gamma > 0$, let H_γ denote the set of classifiers with empirical error within γ of the least empirical error of any classifier in H i.e., $H_\gamma(S_l) = \{h \in H : \mathcal{L}_{S_l}^{0/1}(h) \leq \mathcal{L}_{S_l}^{0/1}(h_H(S_l)) + \gamma\}$. The γ-decisiveness of H in a set $B \subseteq \mathcal{X}$ is*

$$DC_{B,H}(S_l,S_u,\gamma) = \sup_{h_1,h_2\in H_\gamma(S_l)} \Delta_{S_u}(h_1,h_2,B).$$

The following theorem provides conditional generalization bounds for sets in terms of their probability weights and decisiveness. When a set has high probability weight and high decisiveness, the conditional error of the empirical risk minimizer is low.

Theorem 4. *For every $B \subseteq \mathcal{X}$, for any $\delta > 0$, with probability $1-\delta$ over the generation of S_l and S_u, if $\frac{|S_u\cap B|}{|S_u|} > \sqrt{\frac{1}{2|S_u|}\ln\frac{4}{\delta}}$, then*

$$\mathcal{L}_{P|B}^{0/1}(h_H(S_l)) \leq \frac{\epsilon_{approx} + DC_{B,H}\left(S_l,S_u,2\epsilon_{UC}(|S_l|,\delta/4)\right) + \epsilon_{UC}(|S_u|,\delta/4)}{\frac{|S_u\cap B|}{|S_u|} - \sqrt{\frac{1}{2|S_u|}\ln\frac{8}{\delta}}}.$$

Here, for any $m \in \mathbb{N}$, $\epsilon_{UC}(m,\delta/4) = C\sqrt{\frac{VCdim(H)+\log(4/\delta)}{m}}$ for a universal constant C.

Now we show how to assign label coverage sets to guarantee (α)-region-conditional validity for a collection of subsets $\mathcal{B}$, with probability at least $1-\delta$ over sample generation. First we construct a collection $\bar{\mathcal{B}}$ of disjoint sets that cover all sets in $\mathcal{B}$. That is, $\bigcup_{B\in\mathcal{B}} B = \bigcup_{\bar{B}\in\bar{\mathcal{B}}} \bar{B}$ and $\bar{B}_1 \cap \bar{B}_2 = \phi$ for every $\bar{B}_1, \bar{B}_2 \in \bar{\mathcal{B}}$. For each $\bar{B} \in \bar{\mathcal{B}}$,

1. Calculate the upper bound on $L_{P|\bar{B}}(h_{\mathcal{H}}(S_l))$ provided by Theorem 4. For this calculation, set the probability of failure of samples parameter (δ) to be $\frac{\delta}{|\bar{\mathcal{B}}|}$.
2. If the upper bound on $L_{P|\bar{B}}(h_{\mathcal{H}}(S_l))$ is bigger than $1-\alpha$, then assign each point in $\bar{B}$ the trivial coverage set of $\{0,1\}$.
3. If the upper bound on $L_{P|\bar{B}}(h_{\mathcal{H}}(S_l))$ is smaller than $1-\alpha$, then assign each point in $\bar{B}$ the non-trivial coverage set of $\{h_{\mathcal{H}}(S_l)\}$. From the definition of region-conditional coverage, we can see that region-conditional coverage is satisfied for $\bar{B}$. $\bar{B}$ has maximum region non-triviality equalling one.

By construction, each $\bar{B} \in \bar{\mathcal{B}}$, has α-region-conditional validity. This implies α-region-conditional validity for the collection $\mathcal{B}$. This is due to the following lemma:

Lemma 1. *If label coverage sets satisfy α-region-conditional coverage with respect to a collection of disjoint sets $\{B_1, B_2\}$, then the coverage sets also satisfy α-region-conditional coverage with respect to $\{B_1 \cup B_2\}$.*

We note that a low approximation is a natural assumption that models the quality of match between a chosen type of predictor and the learning task at hand. Our analysis here shows that a suitable match indeed yields improved guarantees. To show this we compare the guarantee from Theorem 4 to baseline methods that do not make use of any prior knowledge (See Appendix F for a detailed discussion). In particular we compare the method yielded by Theorem 4 to the split conformal prediction algorithm which was introduced for a distribution-free setting [Vovk et al., 2005] (See Appendix E). Furthermore, we show the improvement we get from decisiveness by comparing the method yielded by Theorem 3 to the method yielded by Theorem 4.

5 GENERATIVE MODELS

The next type of prior assumption about the data generating distribution that we consider is that it belongs to some family of distributions known to the learner. We consider two ways of defining such families. The first is that the family is described by a restriction on the behavior of the induced conditional labeling rule ($l_P(x) = P[y=1|x]$). Families defined in this way restrict only the labeling rules and are *distribution free* with respect to the underlying marginal distribution. The second representation is as mixture models. Namely, for some family of probability distributions that is known to the learner, the data is generated by a mixture of homogeneously-labeled members of that family. This kind of assumption about the data generating distributions is most commonly used for the family of Gaussian distributions.

For this problem we will focus on learning CLF coverage sets. As a first step to achieve this, we look at the learning the CLF. That is, our tool for obtaining coverage sets will be to learn an L_1 approximation of the label generating function. Namely, let the expected CLF-loss of a function f w.r.t. a distribution P be $\mathcal{L}_P^{CLF}(f) = \mathbb{E}_{X\sim P_X}[|f(X) - l_P(X)|]$

Definition 9 (Supervised CLF-learning)**.** *A CLF-learner $\mathcal{A}$ is a function that takes a labeled sample S as input and outputs a function $\hat{l} : \mathcal{X} \to [0,1]$. We say a family of distributions $\mathcal{P}$ is CLF-learnable with sample complexity $m : (0,1)^2 \to \mathbb{N}$ if for any $\epsilon, \delta > 0$, any $m \geq m(\epsilon,\delta)$ and any distribution $P \in \mathcal{P}$ we have*

$$\mathbb{P}_{S\sim P^m}[\mathcal{L}_P^{CLF}(\mathcal{A}(S)) \leq \epsilon] \geq 1-\delta$$

One should note that this is different than the task of learning a regression (real-valued) function. Whereas in the common setup of regression function learning, the training consists of pairs $(x, g(x))$ labeled by the real value of the function g one wishes to approximate, here we only get binary labeled samples (where the binary label is drawn according to the real valued target function l_p). This setup is known as *Learning Probabilistic Concepts* (Kearns and Schapire [1994]).

The following observation explains how one can construct a CLF-coverage learner from a CLF learner.

Observation 1. *Let $\mathcal{A}$ be a CLF-learner for $\mathcal{P}$ with sample complexity $m_{CLF}(\epsilon,\delta)$. Let $\epsilon(m,\delta) = \min\{\epsilon' \in (0,1) : m \geq m(\epsilon',\delta)\}$. Then we can define a respective CLF-coverage set learner $\mathcal{A}'_{c,\delta}$ by $\mathcal{A}'_{c,\delta}(S)(x) = [\mathcal{A}(S)(x) - \frac{\epsilon(|S|,\delta)}{c}, \mathcal{A}(S)(x) + \frac{\epsilon(|S|,\delta)}{c}]$. Then for any constant $c \in [0,1]$ the learner $\mathcal{A}'_{c,\delta}$ is $(c, 1, 2c\epsilon, \delta)$-successful on i.i.d. samples of size $m \geq m_{CLF}(\epsilon,\delta)$.*

5.1 CLF- LEARNING UNDER VARIOUS GENERATIVE ASSUMPTIONS

The problem of CLF-learning under a restricted family of distributions can be reduced to different previously analysed learning tasks (depending on the representation of the family of generating distributions).

5.1.1 Learning CLF's as Probabilistic Concepts

As described above, when the prior assumptions can be expressed as a restriction on the family of the labeling rule functions, the CLF- learning task is equivalent to the probabilistic concepts learning. Kearns and Schapire [1994] offer

efficient learning algorithms for several families of probabilistic concepts. Among those families are the family of non-decreasing functions, the family of probabilistic decision lists and some classes motivated by the assumption that the labeling is deterministic but some of the relevant variables are not observable to the learner.

Alon et al. [1997] take a purely statistical approach (without any algorithmic and computational complexity considerations) and provide a characterization of the learnability of families of such functions in terms of combinatorial dimensions that have become known as *fat shattering dimensions*.

5.1.2 Learning CLF's under the Mixture Model Representation

We show that CLF-learning can be achieved whenever there is an unsupervised learner for the family of underlying marginal distributions. Furthermore, we bound the sample complexity of learning a mixture of homogeneously labeled distributions as a function of the (unsupervised) sample complexity of learning the family of distributions used in the mixtures with respect to the total variation distance.

Definition 10 (Total Variation (TV) distance)**.** *The total variation distance between two distributions, represented by their probability density functions (PDFs)* p_1 *and* p_2 *is defined by:* $d_{TV}(p_1, p_2) = \int |p_1(x) - p_2(x)|dx$.

Definition 11 (Distribution learner)**.** *A distribution-learner* $\mathcal{A}$ *is a function that takes an unlabeled sample* S *as input and outputs a density function* $p : \mathcal{X} \to [0, 1]$.

Definition 12 (TV distance learning of distributions)**.** *We say a family of distributions* $\mathcal{P}$ *is TV-learnable with sample complexity* $m_{TV,\mathcal{P}} : (0, 1)^2 \to \mathbb{N}$ *if there exists a distribution learner* $\mathcal{A}$ *such that for any* $\epsilon, \delta > 0$*, any* $m \geq m(\epsilon, \delta)$ *and any distribution* $P \in \mathcal{P}$ *we have*

$$\mathbb{P}_{S\sim P^m}[d_{TV}(\mathcal{A}(S), P) \leq \epsilon] \geq 1 - \delta$$

In this case we say $\mathcal{A}$ *is a TV-learner of* $\mathcal{P}$.

Theorem 5. *Let* $\mathcal{F}$ *be a family of distributions (identified by their PDFs) over* $\mathcal{X}$ *that can be learned with respect to total variation distance with sample complexity* $m_{TV,\mathcal{F}}$ *that fulfills* $m_{TV,\mathcal{F}}(\frac{\epsilon}{c}, \delta) \leq cm_{TV,\mathcal{F}}(\epsilon, \delta)$ *for all* $c \in (0, 1)$*. Then the family of distributions* $\mathcal{P} = \{aP_1 \times \{1\} + (1 - a)P_0 \times \{0\} : P_0, P_1 \in \mathcal{F}, a \in (0, 1)\}$ *can be CLF-learned with sample complexity* $m_{CLF,\mathcal{P}}(\epsilon, \delta) = \max\{m_{TV,\mathcal{F}}(\frac{\epsilon}{3}, \frac{\delta}{3}), \frac{-9\ln(\frac{\delta}{2})}{2\epsilon^2}\}$

Labeled Gaussian Mixture Models A labeled Gaussian Mixture model in Euclidean space $\mathbb{R}^d$ is defined by

- A collection of Gaussians, $G_1 = (\mu_1, \Sigma_1), \ldots G_k = (\mu_k, \Sigma_k)$, where the μ_i's are the means of the Gaussians and the Σ_i's their covariance matrices'
- A weight vector $(w_1, \ldots w_k)$ with $\sum_{i=1}^k w_i = 1$.
- A binary label vector $(y_1, \ldots y_k)$.

The data is generated by the following procedure: an index $i \leq k$ is picked with probability w_i, a point x is then generated by G_i and labeled y_i. Under such an assumption the probability of a point x being labeled y is fully determined by these model parameters.

Ashtiani et al. [2020] show that for the family of (unlabeled) mixtures of Gaussians in $\mathbb{R}^d$ there is an unsupervised learning algorithm that requires $\tilde{O}(\frac{kd^2}{\epsilon^2})$ samples (where k is the number of Gaussians in the mixture and ϵ is the TV approximation guaranteed).

Corollary 1. *The family of mixtures of* k *label-homogeneous Gaussians in* $\mathbb{R}^d$ *can be CLF-learned with sample complexity* $\tilde{O}(\frac{kd^2}{\epsilon^2})$.

6 COMPARING SAMPLE COMPLEXITIES OF TASKS

In this section, we compare the CLF-coverage learning problem, the CLF-learning problem, and the problem of learning the Bayes classifier, in terms of sample complexity. We construct classes of distributions for which the following hold (see appendix for elaborations).

- The hardest CLF-coverage learning problem requiring point-wise validity has lower sample complexity than the problem of learning the CLF in TV distance.
- The easiest CLF-coverage learning problem requiring domain validity has higher sample complexity than the problem of learning the Bayes optimal classifier.

6.1 CONNECTING CLF-LEARNING TO CLASSIFICATION

We also examine the connection between CLF learning and learning a good classification rule.

We first show that CLF learning implies learning the classification problem up to excess risk.

Observation 2. *If a family of distributions* $\mathcal{P}$ *is CLF-learnable with sample complexity* $m_{CLF,\mathcal{P}}(\epsilon, \delta)$*, then* $\mathcal{P}$ *is learnable with respect to excess risk with sample complexity at most* $m(\epsilon, \delta) \leq m_{CLF,\mathcal{P}}(2\epsilon, \delta)$.

Now we show that learning CLF-coverage sets can be harder than learning the Bayes classifier. In the appendix we define classes $\mathcal{F}_\mu$ such that the Bayes classifier for any distribution in any class $\mathcal{F}_\mu$ is the classifier that thresholds at zero. We do not need any samples to learn the Bayes classifier for the classes $\mathcal{F}_\mu$. However, to provide CLF-coverage sets, we will need samples. This holds even for the easiest CLF-covering problem requiring domain validity and domain non-triviality.

7 CONCLUSION

This paper investigated the problem of confidence in predictions of a statistical learning algorithm on sub-domains or specific instances of the data. We address the problem from a theory perspective and provide formal definitions and provable results (rather than experimental evidence). It is important to realize that any non-trivial guarantees concerning the confidence in predicting some unknown label inevitably rely on prior knowledge about the learning task at hand. For formal analysis of statistical learning such prior knowledge is typically modeled in form of assumptions on the data-generating process, the type of predictors used and the quality of match between these components. We consider three common types of such assumptions, namely Lischitzness of the labeling rule, a given generative model (such as a Gaussian mixture model), or the availability of a hypothesis class (a.k.a. concept class) that is a suitable match with the data-generating process (technically, this is stated as the hypothesis class having a low approximation error). Under each of these assumptions we derive confidence guarantees that depend on the parameters of the assumptions made as well as on the relationship between the region of interest and the available training data. We hope that these results will inspire follow up work on identifying more general types of prior knowledge (assumptions) that allow for the type of confidence guarantees set out in our framework. Additionally, it would also be important to derive complementing lower bounds on the type of guarantees here. Finally, future work may explore the role that requirements for computational efficiency play for the type of coverage guarantees we analyze here.

Acknowledgements

All authors would like to thank Toronto's Vector Institute: it supported Nivasini Ananthakrishnan and Tosca Lechner through Vector research grants, Shai Ben-David through a faculty appointment and Ruth Urner through faculty affiliate membership. This work was further funded by two NSERC discovery grants (held by Shai Ben-David and Ruth Urner).

References

Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44(4):615–631, 1997.

Hassan Ashtiani, Shai Ben-David, Nicholas J. A. Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Near-optimal sample complexity bounds for robust learning of gaussian mixtures via compression schemes. *J. ACM*, 67(6):32:1–32:42, 2020.

David Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, USA, 2012. ISBN 0521518148.

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.

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians. *arXiv preprint arXiv:1810.08693*, 2018.

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

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 08 2020.

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

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

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.

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

Michael J. Kearns and Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. *J. Comput. Syst. Sci.*, 48(3):464–497, 1994.

Jeongyeol Kwon and Constantine Caramanis. The em algorithm gives sample-optimality for learning mixtures of well-separated gaussians, 2020.

Jing Lei. Classification with confidence. *Biometrika*, 101 (4):755–769, 2014.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B*, 76(1):71–96, January 2014.

Tianyang Li, Xinyang Yi, Constantine Caramanis, and Pradeep Ravikumar. Minimax gaussian classification & clustering. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS*, pages 1–9, 2017.

Loizos Michael. Partial observability and learnability. *Artif. Intell.*, 174(11):639–669, 2010.

Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS*, pages 93–102, 2010.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12): 371–421, 2008.

Gregory Shakhnarovich, Trevor Darrell, and Piotr Indyk. Nearest-neighbor methods in learning and vision. *IEEE Trans. Neural Networks*, 19(2):377, 2008.

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

Leslie G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, 1984.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Mach. Learn.*, 92(2-3):349–376, 2013.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

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

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