Identifying Regions of Trusted Predictions

by

Nivasini Ananthakrishnan

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Mathematics

in

Computer Science

Waterloo, Ontario, Canada, 2021

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners. I understand that my thesis may be made electronically available to the public.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis is based on joint work with Shai Ben-David, Tosca Lechner, and Ruth Urner.

Abstract

Quantifying the probability of a label prediction being correct on a given test point or a given sub-population 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 generation process and show how these allow learning justifiably trusted predictors.

Acknowledgements

First and foremost, I would like to thank my advisor Prof Shai Ben-David. Working with Shai was very inspiring and I am grateful he generously shared his time and immense expertise with me. I would also like to thank Prof. Csaba Szepesvári and Prof. Lin Yang for their guidance. I am grateful to them for including me in their project and introducing me to reinforcement learning.

I thank my thesis readers – Prof. Gautam Kamath and Prof. Yaoliang Yu for their time and their valuable feedback. I thank all the other people I got to work with and learn from - Sushant Agarwal, Tosca Lechner, Prof. Ruth Urner, and Sharan Vaswani. Thanks to Sushant and Tosca for being wonderful friends as well. Thanks to Tosca for the research walks, which is where many ideas in this thesis were formed or refined.

I thank all my friends for the happy memories in Waterloo. Special thanks to Frieda and Joyce. I would like to thank Aditya for his help and support and for pushing me while also keeping me grounded. Finally, I would like to thank my family. I could not have asked for a better support system. I thank my parents, sister, and grandmother for always believing in me and giving me the confidence to pursue my interests.

Dedication

To my grandmother – Ganthimathi Tamilarasan.

Table of Contents

Chapter 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 guarantees 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 ([43, 29, 27, 6]) 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 the PAC setup, such knowledge is often expressed as restrictions on the data generating process. 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* ([43]) 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* [7, 51, 15, 21, 23], 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 point-wise validity and non-triviality.

Finally, we provide a few (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. Additionally, we show how to identify regions for successful label-coverage set predictors under an assumption low approximation error by some hypothesis class. We end by discussing how the sample complexities of CLF-coverage set learners and label-coverage set learners compare to each other. We also compare these learning problems with other common learning problems - distribution learning and approximating the Bayes classifier in terms of sample complexities.

1.1 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 (training data, test point, stochasticity in label-generation given a point), various types of regions where trusted predictions may be required (domain-wide, region-wise, point-wise). Our definitions further make the trade-off between validity and non-triviality explicit. 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 an 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.

- **Comparing sample complexities** We also analyze the (information theoretic) difficulties of various related learning problems. We show that the ordering of sample complexities of CLF-coverage set learning and label-coverage set learning depends on the family of distributions of the learning problem. We show that sample complexities of binary classification, coverage-set learning, and marginal distribution learning are in strictly increasing order.

Chapter 2

Formal Framework

We use a standard learning-theoretic setup. We let $\mathcal{X}$ denote some domain or feature space and $\mathcal{Y} = \{0, 1\}$ be 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*

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

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 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* $\alpha > 0$ *be a confidence parameter.*

- *We say the coverage set hypothesis satisfies* $(1-\alpha)$-domain-validity *(are* $(1-\alpha)$*-domain-valid) with respect to* P *if we have*

$$\mathbb{P}_{(X,Y)\sim P}[Y \in \mathrm{c}(X)] \geq 1 - \alpha \quad \textit{and}$$
$$\mathbb{P}_{X \sim P_{\mathcal{X}}}[l_P(X) \in \mathrm{r}(X)] \geq 1 - \alpha$$

  respectively.

- *For a subset $B \subseteq \mathcal{X}$ of the domain, we say that they satisfy* $(1-\alpha, B)$-region-conditional validity in B *with respect to P if we have*

$$\mathbb{P}_{(X,Y)\sim P}[Y \in \mathrm{c}(X) | X \in B] \geq 1-\alpha \quad \textit{and}$$
$$\mathbb{P}_{X\sim P}[l_P(X) \in \mathrm{r}(X) | X \in B] \geq 1-\alpha$$

  respectively.

- *We say that the label coverage hypothesis* c *satisfies* $(1-\alpha, \{x\})$-point-wise validity *at point $x \in \mathcal{X}$ with respect to P, if we have*

$$\mathbb{P}_{(X,Y)\sim P}[Y \in \mathrm{c}(X) | X = x] \geq 1-\alpha$$

  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-conditional validity for* $\mathcal{B}$ if the above condition holds for all regions $B \in \mathcal{B}$. Note that domain validity is a special case of region-conditional validity when the collection $\mathcal{B} = \{\mathcal{X}\}$. 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 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. Let $\mathcal{B} \subseteq 2^{\mathcal{X}}$ be a collection of subsets of the domain. We say that* c *satisfies* $(1-\beta, \mathcal{B})$-region-conditional non-triviality *with respect to P, if for every $B \in \mathcal{B}$,*

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

and that r *has* $(1-\beta, \gamma, \mathcal{B})$-region-conditional non-triviality *if for every $B \in \mathcal{B}$,*

$$\mathbb{P}_{X\sim P_{\mathcal{X}}}[\mu(\mathrm{r}(X)) \leq \gamma | X \in B] \geq 1-\beta.$$

When $\mathcal{B} = \{\mathcal{X}\}$, we say that they satisfy $(1-\beta)$- (or $(1-\beta, \gamma)$)-domain non-triviality *with respect to P.*

These quality criteria for coverage set hypotheses give rise to a notion of success for a coverage set learner that calls for validity and non-triviality to hold simultaneously. This success notion is relative to what can be achieved with full knowledge of the data generating distribution. When we know the CLF at every point, it is possible to achieve a CLF coverage set hypothesis with width zero (coverage set is a singleton) at all points. This is true for all types and levels of validity.

However, achieving validity and non-triviality simultaneously might not be possible for label coverage, even with full knowledge of the data generating distribution. This is often the case when the CLFs are close to a half (the labelling is far from deterministic) at many points. For example, consider a data generating distribution where the CLF is $\frac{1}{2}$ at all points. Then, if we require $(1-\alpha)$-domain validity, for $\alpha < \frac{1}{2}$, even with complete distribution knowledge, the maximum level of domain non-triviality we can achieve is α. We now define a baseline non-triviality that captures the best non-triviality we can hope to achieve for CLF and label, coverage given a target validity requirement. We have a baseline corresponding to each type of non-triviality (domain/region-conditional)

Definition 5 (Baseline non-triviality for label coverage set hypothesis)**.** *Given a validity requirement, a collection of subsets of the domain $\mathcal{B} \subseteq 2^{\mathcal{X}}$, the baseline $\mathcal{B}$-region-conditional non-triviality for label coverage relative to the distribution P is*

$$\bar{\beta}_{label}(\mathcal{B}; P) = \sup_{\substack{c:\mathcal{X}\to\{\{0\},\{1\},\{0,1\}\} \\ c \text{ is valid}}} \inf_{\beta>0} c \text{ has } (1-\beta, \mathcal{B})\text{-region conditional non triviality w.r.t. } P.$$

The special cases of $\mathcal{B} = \{\mathcal{X}\}$ and $\mathcal{B} = \{\{x\} : x \in \mathcal{X}\}$ define the domain-wide and region-conditional non-triviality baselines.

Definition 6 ((α, β, δ)- and $(\alpha, \beta, \gamma, \delta)$-successful coverage set learning)**.** *Let $\mathcal{P}$ be a class of distributions and $\mathcal{B} \subseteq 2^{\mathcal{X}}$. For a triplet of parameters $(\alpha, \beta, \delta) \in (0, 1]^3$, a label coverage set learner $\mathcal{A}$ is* $(\alpha, \beta, \delta, \mathcal{B})$-region-conditional successful *for $\mathcal{P}$ if, 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 $(1-\alpha, \mathcal{B})$-region-conditional valid and $(\bar{\beta}_{label}(\mathcal{B}; P) - \beta, \mathcal{B})$-region-conditional-non-trivial is greater than $1-\delta$. Note that $\bar{\beta}_{label}(\mathcal{B}; P)$ is the baseline non-triviality for label coverage defined in Definition 5.*

Two special cases of region-conditional learning success are domain-wide learning success and pointwise learning success which are the cases of $\mathcal{B} = \{\mathcal{X}\}$ and $\mathcal{B} = \{\{x\} : x \in \mathcal{X}\}$

respectively. 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) when there is a sample size such that, with probability at least* $1 - \delta$*, the learned CLF coverage set hypothesis satisies* $(1 - \alpha)$*-validity and* $(1 - \beta, \gamma)$*-non-triviality.*

We say that a label coverage set is successful if it is (α, β, δ)*-successful for all triplets of parameters* $(\alpha, \beta, \delta) \in (0, 1]^3$*.*

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.

Chapter 3

Related Work

Quantifications of confidence in predictions are often derived in Bayesian learning setups, and then rely on on the usual conditions on the quality of priors in a Bayesian reasoning framework [5]. 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 are papers developing algorithmic methods for confidence scores in a non-Bayesian framework. Jiang et al. [22] propose a method that assigns the confidence of a prediction for a point to be the ratio of the distance to the nearest, high density sample point with the same prediction and the distance to the nearest, high density sample point with a different prediction. Another approach is providing confidence scores by applying a scaling transformation to the real-valued predictions of learning algorithms such as support vector machines, neural networks, and boosting [38, 39, 20]. A common scaling transformation applied in these mathods is the Platt scaling introduced by Platt [40] to obtain well-calibrated posterior probabilities from the output of support vector machines. Confidence scores are well-calibrated if the confidence score corresponds to the average accuracy of points with that confidence score. The validity of the confidence scores from these algorithms rely on several technical assumptions on the data-generating process. The papers studying these algorithms empirically evaluate the calibration of the confidence scores on labelled data.

In this work, we take a step back and aim to develop 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. We now provide a brief survey of various lines of work that are most relevant to our problem.

- **Conformal prediction** One early approach to providing confidence estimates to

prediction is through the notion of *conformal prediction* [43]. Conformal prediction outputs regions in the label space for each point in the domain. The goal is for the conformal prediction of each point to contain labels that occur with significant probability according to the CLF of that point. In most previous work, methods are proposed to meet this goal with high probability over both the training data used for the conformal prediction and the test point. This type of guarantee is referred to as marginal coverage guarantee. We refer the reader to the textbook by Vovk et al. [49] for a detailed survey of this topic.

A disadvantage of the marginal guarantee is that it only holds with high probability over test points drawn i.i.d. from the training distribution. There is some work on conformal prediction aiming to overcome this limitation. These works explore guarantees that hold for all points or with high probability conditioned on membership in predefined subsets of the domain [29, 30, 48, 6]. These works aim to obtain results that hold for all distributions. Most results of these papers describe limits of distribution-free, conditional conformal prediction. Lei and Wasserman [29, 30] show that it is impossible to give point-wise-guarantees in the distribution-free, regression setting. Vovk et al. [48] extend this result to a general prediction setting that includes classification. Barber et al. [6] show that it is also impossible to give conformal predictions that have distribution-free, region-conditional validity that hold for all regions with mass greater than some minimum weight parameter δ. However, distribution-free, region-conditional validity is possible for a pre-defined collection of regions with finite VC dimension. Barber et al. provide an algorithm to achieve this type of validity. The algorithm is a modification of a common algorithm for conformal prediction - the *split conformal algorithm*. We discuss the split conformal algorithm and the modified version in more detail in the appendix in Section A.1. While this algorithm has the desired quality of providing distribution-free validity guarantees, under certain distributional assumptions, this algorithm could provide coverage sets with sub-optimal non-triviality. We show that this is the case under the distributional assumptions we consider in this work in Section 5.1.

- **Selective Classification** Another line of work that is related to our paper is selective classification / classification with abstention [7, 51, 15, 21, 23]. In this setting, a classifier is given the option to abstain from making a prediction. The goals of selective classification are to minimize incorrect predictions while also minimizing abstentions. The first goal is analogous to our validity requirement and the second is analogous to our non-triviality goal. A selective classifier implicitly describes a set of low certainty. This is the set of all points for which the classifier abstains. The level of uncertainty of this set is determined by the classifier's trade-off between

the two goals of minimizing errors and minimizing abstentions. Many works in this line provide accuracy guarantees that hold with high probability over the domain [7, 51, 15, 21, 23].

Point-wise guarantees for selective classification that all predictions made are correct predictions are provided in earlier work by El-Yaniv and Wiener [13]. Their results for such point-wise guarantees are developed under an distributional assumption of realizability by a hypothesis class. Point-wise guarantees in the agnostic case are also studied by El-Yaniv and Wiener [50]. The guarantee they provide, called the point-wise competitive guarantee, is different from a guarantee on the accuracy. The guarantee is that, given a function class, on each point that the classifier does not abstain, the classifier's label is the same as the label assigned by the classifier in the function class with least true error.

- **Covariate shift** Our goal of getting guarantees beyond guarantees that hold on average for test points drawn from the training distribution is shared by the problem of learning with covariate shift. Here the goal is to learn a classifier based on samples from a training distribution to achieve low error with respect to a different test distribution. The training and test distributions have different marginal distributions, but share the same conditional labelling functions for all points. There are many papers on this topic, see e.g., the book by Quionero-Candela et al. [41] for a survey of this topic. For a collection of regions, it is possible to obtain non-trivial coverage sets with region-conditional validity if it is possible to learn with non-vacuous error bounds for appropriately chosen test distributions (that depend on the collection of regions). However, existing bounds on errors on distribution with covariate shift fail to enable non-trivial region-conditional or point-wise valid coverage sets. A main issue with many existing bounds such as bounds based on $H\Delta H$-divergence and total variational distance that are studied by Ben-David et al. [8] is that they view the training and test distributions symmetrically. Our results can be interpreted as giving better generalization bounds for covariate shift under some assumptions on the training and test distributions. The bounds our results imply do not treat the training and test distributions symmetrically.

- **PQ learning** A learning framework - *PQ learning*, introduced by Goldwasser et. al. [19], combines selective classification and learning with covariate shift. This is a form of selective classification with accuracy requirements for test points drawn from a distribution (Q) different from the distribution training points are drawn from (P). The goal is to learn a classifier with minimal incorrect predictions on points drawn from the test distribution Q (minimal Q-error) and minimal abstentions on points

drawn from the training distribution P (minimal P-error). The goals of our coverage set learning problem can be interpreted as achieving both low error rate and low abstention rate on the test distribution Q, without any performance requirements on the training distribution. Goldwasser et al. note that for P and Q with low total variation distance, low abstention rate for P implies a low abstention rate for Q. In such a case, successful PQ learning also implies the goals of our problem. Our results on label coverage set learning can be interpreted as describing additional properties of the training and test distributions, other than proximity in total variation distance, for which it is possible to have low error rate and low abstention rate on the test distribution.

- **Out of distribution detection** Another common approach to finding points having uncertain predictions is to find points that are generated by the underlying distribution with low likelihood. One common approach is to assume that the underlying distribution can be approximated by generative models such as Gaussian mixture models and to then approximate the underlying distribution. Points that occur with low likelihood in the approximate distribution are identified as out of distribution points [9], [34]. Another approach is to perform statistical tests to determine the difficulty of distinguishing test points from sample points. The approaches we propose in this work go beyond in or out distribution identification to assign certainty. In some cases, our approaches are able to identify predictions of high certainty even among points that are out of distribution.

- **Probabilistic concepts** A problem setting that is closely related to the problem of constructing CLF coverage sets is *probabilistic concepts learning* of a family of probability distributions. In this setting, we are given pairs of points and labels that are drawn from a distribution belonging to the distribution family. The goal is to estimate the CLF of most points accurately. Probabilistic concepts learning can be used to construct CLF coverage sets. Likewise, we can learn probabilistic concepts using CLF coverage sets. We discuss this connection between CLF-coverage learning and probabilistic-concepts learning in Section A.2 in the appendix. Efficient learning algorithms for learning probabilistic concepts for several families of probability distributions are given by Kearns and Schapire [24]. 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 labelling is deterministic but some of the relevant variables are not observable to the learner. Alon et al. [1] take a purely statistical approach and provide a characterization of the learnability of families of distributions in terms of combinatorial dimensions that are known as *fat shattering*

dimensions.

We study the different coverage set learning problems under two different types of standard assumptions on the data-generating process: Access to a hypothesis class that has low approximation error and Lipschitzness of the CLF . Low approximation error is a standard assumption in statistical learning theory (e.g., [45]). 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 ([44]).

Chapter 4

Lipschitzness of Labelling Function Assumption

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

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 set learner (Algorithm 1) for which we show the strongest validity and non-triviality guarantees - 1-point-wise validity and domain non-triviality (see definitions 3 and 4). We also identify conditions on points that lead to more narrow CLF-coverage sets.

The CLF-coverage set 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 (c), the algorithm then calculates the average label of samples in the cell - $\hat{l}[c]$. This is an estimate of the expected label conditioned upon membership in the cell. The algorithm calculates a confidence interval (of width $w[c]$) 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 set. The CLF-coverage set for a point x contained in a cell t_x is an interval centered at $\hat{l}[t_x]$ and having width $w[t_x] + r\lambda\sqrt{d}$.

We will now analyze the validity and the non-triviality of the CLF-coverage-hypothesis provided by Algorithm 1. We start with the analysis of the validity. Theorem 1 states the point-wise validity guarantee of the CLF-coverage-hypothesis 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 $\mathrm{t_x}$ containing the test point x.
$\hat{p}[\mathrm{t_x}] :=$ fraction of samples in $\mathrm{t_x}$.
$w_p(m, \delta/2r^d) = \sqrt{\frac{1}{2m} \ln \frac{4r^d}{\delta}}$.
$w_\ell(m, \delta/2r^d, \hat{p}[\mathrm{t_x}]) = \frac{2w_p(m,\delta/2r^d)}{\hat{p}[\mathrm{t_x}] - w_p(m,\delta/2r^d)}$
$\hat{\ell}[\mathrm{t_x}] :=$ fraction of samples in the cell $\mathrm{t_x}$ with label 1.
$w[\mathrm{t_x}] := 1$
if $\hat{p}[\mathrm{t_x}] - w_p(m, \delta/2) > 0$ **then**
  $w[\mathrm{t_x}] = w_\ell(m, \delta, \hat{p}[\mathrm{t_x}])$
end if

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

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

Namely, the CLF-coverage set for x is centered at $\hat{\ell}[t_x]$ (as an estimate of $\ell_P(x)$), and has width $2(w[t_x] + r\lambda\sqrt{d})$.

Theorem 1. *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-hypothesis having point-wise validity (see definition 3).*

Proof. 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. Let l_c be the average CLF of the cell and let $\hat{l}_c$ be the estimate of l_c that is calculated as the fraction of sample points in the cell c with label 1.

The crux of the proof is bounding how far the estimate $\hat{l}_c$ can be from l_c. This involves bounding how far $\hat{p}_c$ can be from p_c. Once we show that $\hat{l}_c$ is a good estimate of l_c, we bound how far from l_c the CLF of any point in the cell c can be. This bound is in terms of the size of the cell and the Lipschitz constant. The smaller the cell size and the Lipschitz constant, the closer the CLF of any point in c must be to the average CLF - l_c. We use two technical lemmas - Lemma 7 and Lemma 8 to bound $|\hat{p}_c - p_c|$ and $|\hat{l}_c - l_c|$ respectively. These lemmas are stated and proved in the appendix. These lemmas are proved primarily by applying the Hoeffding inequality (stated in the appendix as Lemma 5).

From Lemma 7 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 $w_p(c) = w_p(m, \delta/2r^d) = \sqrt{\frac{1}{2m} \ln \frac{4r^d}{\delta}}$ (as defined in Lemma 7).

From Lemma 8, 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 $w_\ell(c) = w_\ell(m, \delta/2r^d, \hat{p}_c) = \frac{2w_p(m,\delta/2r^d)}{\hat{p}[\mathrm{t_x}] - w_p(m,\delta/2r^d)}$ (as defined in Lemma 8).

The maximum distance between any two points in any cell is $r\sqrt{d}$. By the λ-Lipschitzness property, any point in the cell has labelling probability within $\lambda r\sqrt{d}$ 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:

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

This is the interval returned by the algorithm. Therefore, for every point, the CLF lies in the coverage set returned by the algorithm, with probability at least $1 - \delta$. This proves that the algorithm returns point-wise valid CLF-coverage sets with probability at least $1 - \delta$ over the training samples.

□

Next, we analyse the non-triviality of the CLF-coverage-hypothesis of Algorithm 1 in Theorem 2. This theorem states that for a large enough sample size and an appropriately chosen input parameter r, Algorithm 1 returns a CLF-coverage-hypothesis with large domain non-triviality.

Theorem 2. *For every λ-Lipschitz distribution, for every $\beta, \gamma, \delta > 0$, there is a sample size $m(\beta, \gamma, \delta, \lambda)$ such that with probability at least $1-\delta$ over samples S of size $m \geq m(\beta, \gamma, \delta, \lambda)$, Algorithm 1 with input parameters $S, r = 1/m^{\frac{1}{8d}}, \delta, \lambda$ yields a CLF-coverage-hypothesis having $(1-\beta, \gamma)$-domain-wide non-triviality (see Definition 4).*

Proof. We know that cells with large probability weights have narrow confidence intervals for their CLF estimates. Additionally, the probability of drawing a point from a cell with low probability weight is low. Using these ideas, we can show that with high probability over domain points, we can get CLF estimates with narrow confidence intervals.

We show that the lengths of the CLF confidence intervals for cells with probability weights greater than a parameter μ is less than γ. We then show that the probability weight of all cells with weight less than μ is less than β. This suffices to show the required $(1-\beta, \gamma)$-domain-wide non-triviality stated in the theorem. The rest of the proof describes how to choose the weight parameter μ and the sample size.

When the sample size is m, with probability at least $1 - \frac{\delta}{2}$, for all cells with probability weight greater than $\mu = \frac{1}{m^{1/4}}$, the probability weight estimate, which is the fraction of samples that lie in the cell, is at least $\mu - \sqrt{\frac{1}{2m} \ln \frac{4r^d}{\delta}}$ (by Lemma 7). Therefore, when Algorithm 1 is provided with the input parameter describing the size of cells equalling $r = \frac{1}{m^{1/8d}}$, the length of the confidence interval of the labelling probability of cells with weight at least μ is less than the following quantity (see Algorithm 1):

$$\frac{\sqrt{\frac{2}{2m} \ln \frac{4m^{1/8}}{\delta}}}{\frac{1}{m^{1/4}} - 2\sqrt{\frac{1}{2m} \ln \frac{4m^{1/8}}{\delta}}} + \frac{2\lambda d\sqrt{d}}{m^{1/8d}}$$
$$= \frac{2}{\frac{m^{1/4}}{\sqrt{\ln \frac{4m^{1/8}}{\delta}}} - 2} + \frac{2\lambda d\sqrt{d}}{m^{1/8d}}.$$

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

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

Choosing a sample size $M(\beta, \gamma, \delta, \lambda)$ greater than $M_1(\gamma, \delta, \lambda)$ and $M_2(\beta)$, we get that

$$
\begin{aligned}
&\mathbb{P}_{X \sim P_{\mathcal{X}}}[\text{Algorithm 1 returns CLF-coverage set of width greater than } \gamma \text{ for } X] \\
&\leq \mathbb{P}_{X \sim P_{\mathcal{X}}}\left[X \in \text{cell with weight at most } \mu = 1/M(\beta, \gamma, \delta, \lambda)^{\frac{1}{4}}\right] \\
&\leq \mu \cdot \frac{1}{r^d} \\
&= \frac{1}{M(\beta, \gamma, \delta, \lambda)^{1/8}} \\
&< \frac{1}{M_2(\beta)^{1/8}} \\
&< \beta.
\end{aligned}
$$

This proves that with probability at least $1 - \delta$ over training data, the CLF-coverage-hypothesis returned by Algorithm 1 is $(1 - \beta, \gamma)$-domain-wide non-trivial. □

Chapter 5

Function Class With Low Approximation Error Assumption

In this section, we study the assumption that the function class H has approximation error $\text{opt}_P(H)$ less than ϵ_{approx}. That is, we assume we know that $\min_{h \in H} \mathcal{L}_P^{0/1}(h) \leq \epsilon_{\text{approx}}$. Under this assumption, we show how to construct a label-coverage-hypothesis having region-conditional validity with respect to a predefined collection of regions - $\mathcal{B}$. We construct the label-coverage-hypothesis using both a labelled sample set S_l and an unlabelled sample set S_u. We also study non-triviality of the label coverage sets and identify sufficient conditions of regions that result in non-trivial coverage-sets

First we introduce some notation that we use to state the results in this section.

- **Empirical risk minimizer:** $h_H(S_l)$ denotes an empirical risk minimizer, from the class H, for the sample S_l. That is, $h_H(S_l) \in \text{argmin}_{h \in H} \mathcal{L}_{S_l}^{0/1}(h)$.
- **Error of a region**: The error of a classifier h for a region $B \subseteq \mathcal{X}$ relative to P is defined as
$$\mathcal{L}_{P,B}^{0/1}(h) = \mathbb{P}_{(X,Y)\sim P}[h(X) \neq Y, X \in B].$$
- **Conditional error of a region**: The conditional error of a classifier h for the region $B \subseteq \mathcal{X}$, relative to P, is defined as
$$\mathcal{L}_{P|B}^{0/1}(h) = \mathbb{P}_{(X,Y)\sim P}[h(X) \neq Y | X \in B].$$

  Note that the conditional error of a region is the error of that region divided by the

probability weight of the region. That is,

$$\mathcal{L}^{0/1}_{P|B}(h) = \frac{\mathcal{L}^{0/1}_{P,B}(h)}{P(B)}.$$

- **Empirical error of a region**: An estimate for the error of a classifier h on a region B based on a labelled sample S_l is

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

Before we describe the construction of the coverage sets, here is a brief overview of the approach we propose in this section. As the first step, we show how to obtain an upper bound on the conditional error of the empirical risk minimizer - $\mathcal{L}^{0/1}_{P|B}(h_H(S_l))$, for a given $B \subseteq \mathcal{X}$. We propose methods for two upper bounds that are both calculated using the labelled and unlabelled samples. We then show how to use the conditional error bound to find a valid label-coverage-hypothesis for the collection $\mathcal{B}$. This is done by assigning trivial coverage sets for points in regions with high conditional error bounds (higher than the validity parameter α) and assigning non-trivial coverage sets having the label of $h_H(S_l)$ for all other points.

The first conditional error bound we propose does not require knowing a bound on the approximation error $\text{opt}_P(H)$. This bound pessimistically assumes that all the generalization error - difference between the true error and the sample error, occurs in the region of interest. Although we do not need to know a bound on $\text{opt}_P(H)$ to calculate this conditional error bound, this bound is non-vacuous only when $\text{opt}_P(H)$ is small. Another requirement for this bound to be non-vacuous is for the region of interest to contain many sample points. Therefore, small conditional error bounds from this method are only possible for regions with high probability weights. We now state this conditional error bound as Theorem 3.

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

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

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*

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

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

Proof. Recall that the quantity we want to bound - the conditional error of a region is the ratio of the error of the region and the probability weight of the region. This is, $\mathcal{L}^{0/1}_{P|B}(h_{H(S_l)}) = \frac{\mathcal{L}^{0/1}_{P,B}(h_H(S_l))}{P(B)}$. In the proof, we will show how to estimate an upper bound for the numerator - $\mathcal{L}^{0/1}_{P,B}(h_H(S_l))$ using S_l and a lower bound for the denominator - $P(B)$ using S_u. Together these bounds will give us the upper bound on $\mathcal{L}^{0/1}_{P|B}(h_H(S_l))$ stated in the theorem.

Upper bound for $\mathcal{L}^{0/1}_{P,B}(h_H(S_l))$: By uniform convergence, with probability $1 - \frac{\delta}{2}$, $\mathcal{L}^{0/1}_{P,B}(h_H(S_l)) < \mathcal{L}^{0/1}_{S_l,B}(h_H(S_l)) + \epsilon_{UC}(|S_l|, \delta/2)$.

Lower bound for $P(B)$: By Lemma 7, the estimate of $P(B)$ based on unlabelled samples - $\frac{|S_u \cap B|}{|S_u|}$ satisfies, with probability $1 - \frac{\delta}{2}$, $P(B) > \frac{|S_u \cap B|}{|S_u|} - \sqrt{\frac{1}{2|S_u|} \ln \frac{4}{\delta}}$.

Therefore, with probability $1-\delta$, both the upper bound on the numerator and the lower bound on the denominator are valid. This implies that the upper bound for $\mathcal{L}^{0/1}_{P|B}(h_{\mathcal{H}}(S_l))$ stated in the theorem holds true.

□

In the second conditional error bound, we aim to avoid pessimistically assigning all the generalization error to the region of interest. We introduce a sample dependent property of regions that identifies regions with low generalization error in a more refined way. We call this property of a region the *decisiveness* of the function class on that region. We say that the function class H is decisive on a region $B \subseteq \mathcal{X}$, 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. Theorem 4 provides a conditional error bound in terms of the decisiveness of the region of interest. For a region with probability weight too low to get non-vacuous conditional generalization bounds by using Theorem 3, we can still get non-vacuous bounds when the set has high decisiveness using Theorem 4. In the end of this chapter, we provide an example where this is the case. The example is a distribution that is approximated well by the class of threshold classifiers.

Regions far from the threshold of the ERM classifier have high decisiveness. We show an example of a region far from the threshold that has vacuous conditional error bound by Theorem 3 due to low weight but non-vacuous conditional error bound by Theorem 4 due to high decisiveness. Note that the bound in the following theorem, unlike the bound in the previous theorem, requires the knowledge of an upper bound on the approximation error of the function class H.

Before we state the decisiveness-based conditional error bound, we formally define the decisiveness property of regions.

Definition 8 (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 unlabelled samples 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 9 (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., $H_\gamma(S_l) = \{h \in H : \mathcal{L}^{0/1}_{S_l}(h) \leq \mathcal{L}^{0/1}_{S_l}(h_H(S_l)) + \gamma\}$. The γ-decisiveness of H in a set $B \subseteq \mathcal{X}$ is*

$$DC_{B,H}(S_l, S_u, \gamma) = 1 - \sup_{h_1, h_2 \in H_\gamma(S_l)} \Delta_{S_u}(h_1, h_2, B).$$

The following theorem provides conditional error bounds for regions in terms of their probability weights and decisiveness. When a region 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 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*

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

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.

Proof. Let us denote the classifier in the class with the least error by $h^* = \text{argmin}_{h \in H} \mathcal{L}_P^{0/1}(h)$. The main idea of the proof is that with high probability, h^* belongs to the the class H_γ (see definition 9) for $\gamma = 2\epsilon_{UC}(|S_l|, \delta/4)$. By the definition of decisiveness, all classifiers in H_γ have similar behavior on a region B with high decisiveness. In particular, $h^*, h_H(S_l) \in H_\gamma$ have similar behavior in B. Since $h_H(S_l)$ is similar to h^* and since h^* has low overall error, we can conclude that $h_H(S_l)$ has low error on the region B. We also find a lower bound on the probability weight of the region B. Combining an upper bound (using decisiveness) on the error of $h_H(S_l)$ and a lower bound on $P(B)$, we obtain the upper bound on the conditional error of $h_H(S_l)$ that is stated in the theorem.

By uniform convergence, we get that with probability at least $1 - \frac{\delta}{4}$, $\mathcal{L}_{S_l}^{0/1}(h^*) \leq \mathcal{L}_{S_l}^{0/1}(h_H(S_l)) + 2\epsilon_{UC}(|S_l|, \delta/4)$. This implies that $h^* \in H_{2\epsilon_{UC}(|S_l|, \delta/4)}$.

Also by uniform convergence, with probability at least $1 - \frac{\delta}{2}$, for any $h \in H_{2\epsilon_{UC}(|S_l|, \delta/4)}$,

$$\begin{aligned}\Delta_P(h, h^*, B) &\leq \Delta_{S_u}(h, h^*, B) + \epsilon_{UC}(|S_u|, \delta/4) \\ &\leq 1 - DC_{B,H}\left(S_l, S_u, 2\epsilon_{UC}(|S_l|, \delta/4)\right) + \epsilon_{UC}(|S_u|, \delta/4).\end{aligned}$$

In particular, the above inequality holds for $h_H(S_l)$.

$$\begin{aligned}\mathcal{L}_{P,B}^{0/1}(h_H(S_l)) &\leq \mathcal{L}_{P,B}^{0/1}(h^*) + \Delta_P(h_H(S_l), h^*, B) \\ &\leq \epsilon_{\text{approx}} + 1 - DC_{B,H}\left(S_l, S_u, 2\epsilon_{UC}(|S_l|, \delta/4)\right) \\ &\quad + \epsilon_{UC}(|S_u|, \delta/4)\end{aligned}$$

This concludes bounding the error of $h_H(S_l)$ on the region B. Next we use S_u to obtain a lower bound on $P(B)$ with Lemma 7. We get that with probability at least $1 - \frac{\delta}{8}$, $P(B) > \frac{|S_u \cap B|}{|S_u|} - \sqrt{\frac{1}{2|S_u|} \ln \frac{8}{\delta}}$.

Combining the upper bound on we obtain the upper bound on $\mathcal{L}_{P,B}^{0/1}(h_H(S_l))$ and the lower bound on $P(B)$, we get the upper bound on $\mathcal{L}_{P|B}^{0/1}(h_H(S_l)) = \frac{\mathcal{L}_{P,B}^{0/1}(h_H(S_l))}{P(B)}$ provided by the theorem.

□

Having provided two ways of bounding conditional errors, we now describe how to use these bounds to construct label-coverage-hypotheses having $(1-\alpha)$-region-conditional validity for a collection of subsets $\mathcal{B}$, with probability at least $1-\delta$ over sample generation.

1. For each $B \in \mathcal{B}$, calculate the upper bound on $L_{P|B}(h_H(S_l))$ using either Theorem 3 or Theorem 4. For this calculation, set the probability of failure of samples parameter to be $\frac{\delta}{|\mathcal{B}|}$.

2. For each $B \in \mathcal{B}$, if the upper bound on $L_{P|B}(h_H(S_l))$ is greater than α, then assign 'trivial' status to the region B. Otherwise assign 'non-trivial' status to B.

3. For any point $x \in \mathcal{X}$, if $x \in B$ for some $B \in \mathcal{B}$ with 'trivial' status, assign the trivial label-coverage set - $\{0, 1\}$ to x. Otherwise, assign the non-trivial coverage set $\{h_H(S_l)(x)\}$ to x.

To see that the above conditional error based coverage set satisfies $(1 - \alpha, \mathcal{B})$-region-conditional validity, note that for any $B \in \mathcal{B}$ with $L_{P|B}(h_H(S_l)) < \alpha$, if all points in B are assigned non-trivial coverage sets with the label of the empirical risk minimizer - $h_H(S_l)$, we have $(1 - \alpha)$-region conditional validity relative to B. And now if any point in such a region B is assigned the trivial coverage set, the $(1 - \alpha)$-region-conditional validity still holds.

5.1 Comparison of methods

We have seen a few approaches for constructing label-coverage sets with region-conditional validity so far. In this chapter, we proposed two methods that stem from the two conditional error bounds provided by Theorem 3 and Theorem 4. We will refer to these methods as 'conditional error bound methods' (abbreviated as CEB methods). We will refer to the CEB method based on Theorem 3 as the 'baseline CEB method' and the CEB method based on Theorem 4 as the 'decisiveness-based CEB method'. Another method for coverage sets with region-conditional validity is the modified split conformal algorithm proposed by Barber et al. [6] (see Algorithm 2 in Section A.1 in Appendix A for a description of this algorithm). In this section, we will discuss some differences among these approaches. We will focus on differences in the case of the data generating distribution satisfying the assumption we have been studying in this chapter - low approximation error by a function class H.

The modified split conformal algorithm and the baseline CEB both have the advantage of providing distribution-free validity. The baseline CEB method uses the whole labelled training set to both train an ERM classifier and evaluate that classifier. The split conformal algorithm on the other hand partitions the labelled training set into two parts and uses one part for training a model and the other part for evaluating that model. Due to

this, the classifier used for coverage sets construction in the baseline CEB method is likely to have lower error (by a constant factor) than the classifier in the split conformal method. This could result in the baseline CEB method's coverage sets having higher non-triviality compared to the coverage sets from the split conformal algorithm. However, the split conformal method allows for more general training algorithms and is therefore likely to adapt better even when the probability distribution is not approximated well by the function class.

Compared to the split conformal algorithm and the baseline CEB method, the decisiveness-based CEB method has the disadvantage of requiring knowledge of an upper bound on the approximation error of the function class in order to construct coverage sets. However, the decisiveness-based CEB method makes better use of the distributional assumption to provide coverage sets with higher non-triviality in some cases. The distributional assumption allows the decisiveness-based CEB method to better utilize unlabelled data. Recall that both conditional error bounds are obtained by estimating the error on the region $\left(\mathcal{L}^{0/1}_{P,B}\right)$ and the probability weight of the region $(P(B))$. Unlabelled data is used to estimate the probability weight by both the baseline and the decisiveness-based CEBs. The decisiveness-based CEB also uses the unlabelled data to estimate region's error whereas the baseline CEB uses only labelled data for this. The rest of this section describes an example where decisiveness-based CEB method provides better non-triviality compared to the baseline CEB method and the split conformal method.

We use the following notation for the example: The domain $\mathcal{X}$ is the unit real interval - $[0, 1]$. The class of threshold classifiers over this domain is denoted by $H_{\text{thresholds}} = \{h_a : a \in [0, 1]\}$. The threshold classifier denoted by h_a for $a \in [0, 1]$ is such that $h(x) = 0$ for every $x \leq a$ and $h(x) = 1$ for every $x > a$. For $\epsilon > 0$, $\mathcal{P}_{\text{thresholds},\epsilon}$ denotes the class of probability distributions that are approximated by the class H with approximation error - $\text{opt}_P(H)$ at most ϵ. That is, a probability distribution P belongs to the class $\mathcal{P}_{\text{thresholds},\epsilon}$ if and only if $\min_{h\in H} \mathcal{L}^{0/1}_h \leq \epsilon$.

Example 1. *Let the domain $\mathcal{X}$ be the unit interval $[0, 1] \subseteq \mathbb{R}$. Let the marginal distribution $P_{\mathcal{X}}$ be the uniform distribution. Let the conditional distribution be:*

$$P(y = 1|x) = \begin{cases} 1 - 0.001, & \text{if } x \geq \frac{1}{2} \\ 0.001, & \text{if } x < \frac{1}{2}. \end{cases}$$

We want to construct label-coverage sets based on samples drawn from P using prior knowledge that P has approximation error by the threshold class $opt_P(\mathcal{H}_{thresholds}) = 0.001$. We have access to 100 labelled samples drawn i.i.d from P and 10^7 unlabelled samples drawn

i.i.d from $P_{\mathcal{X}}$. The goal is to provide $(0.85, 0.85, \{B\})$-region-conditional coverage sets for $B = [0, 0.01]$. The following hold:

- *With probability more than $\frac{1}{2}$ over the samples drawn, the split conformal method assigns trivial label-coverage sets for all points in B.*
- *With probability more than $\frac{1}{2}$ over the samples drawn, the baseline CEB method assigns trivial label-coverage sets for all points in B.*
- *With probability more than $\frac{1}{2}$, the decisiveness-based CEB method assigns non-trivial label-coverage sets for all points in B.*

Note that in this example, the claim we make about the split conformal algorithm is for the algorithm with parameter $(1 - \alpha)$ equalling 0.85. Barber et al. [6] show that with this parameter, the coverage sets satisfy a notion called 0.85-restricted conditional coverage. This is a weaker validity guarantee that implies $(0.85, 0.85)$-region conditional validity. Now we prove the correctness of the above example. We outline a sketch of this proof here and defer the full details of the proof to the appendix.

Proof. There are three parts to this proof:

1. Showing that the baseline CEB method returns trivial coverage sets: We show that the bound provided by Theorem 3 for B with sample failure parameter $\delta = 0.15$ is vacuous (greater than 1.0). This implies trivial coverage sets. Note that $\frac{\epsilon_{UC}(|S_l|, 0.15/2)}{\frac{|S_u \cap B|}{|S_u|}}$ is a lower bound on the baseline CEB given by Theorem 3. The numerator of this lower bound is a constant value that we can calculate. The denominator is close to $P(B) = 0.01$ with high probability. The denominator is less than the numerator with high probability and hence the baseline CEB is vacuous.
2. Showing that the split conformal algorithm returns trivial coverage sets: First we show that with probability, there are only few validation samples in B. Then we show that this implies that the split conformal algorithm returns trivial coverage sets.
3. Showing that the decisiveness-based CEB method returns non-trivial coverage sets: We first show that with high probability over the samples, the decisiveness of the region B is the highest value - one. This will imply that the error of the region B is low. The unlabelled samples provide an estimate of the probability weight of B that

is larger than the bound on the region's error. This results in a small conditional error bound due to the decisiveness-based CEB given by Theorem 4.

To show that decisiveness is one with high probability, we show that any classifier in $H_{\text{thresholds}}$ that labels any point in $S_u \cap B$ zero, has high sample error with high probability. This shows that all classifiers with low empirical error have the same behaviour on $S_u \cap B$ i.e., label zero for all points in $S_u \cap B$. Therefore the decisiveness is one.

□

Chapter 6

Comparisons of sample complexities

In this work, we introduced two learning problems for a given family of data generating distributions - label-coverage set learning and CLF-coverage set learning (see Definition 6). A natural question to ask is how the two problems compare in terms of their sample complexities. In this chapter we investigate this question. We show that the ordering of the sample complexities depends on the family of distributions of the learning problem. We also consider two other commonly studied learning problems - *distribution learning in total variation distance* and *Bayes classifier learning*. We also compare the sample complexities of these learning problems to the sample complexity of CLF-coverage set learning. We find that the problems - distribution learning, CLF-coverage set learning and Bayes classifier learning have strictly decreasing sample complexities.

We start by comparing the label-coverage set learning problem and the CLF-coverage set learning problem. We construct an example where label-coverage set learning has a higher sample complexity than CLF-coverage set learning (Example 2). Then we construct an example where the sample complexity order is reversed i.e., CLF-coverage set learning has higher sample complexity than label-coverage set learning (Example 3). For these examples, the domain is a singleton $\mathcal{X} = \{x_0\}$. Since the domain only has one element, all types of guarantees - point-wise, region-conditional, and domain-wide are equivalent. Any distribution over the domain can be described by the CLF of the point x_0. We denote the distribution with CLF value p at x_0 by $\text{Bern}_{x_0}(p)$.

Example 2 (Label-coverage set learning harder than CLF-coverage set learning)**.** *Let the domain be $\mathcal{X} = \{x_0\}$. For any coverage learning parameters $0 < \alpha < \frac{1}{2}, \beta, \gamma, \delta > 0$, consider the distribution family $\mathcal{P}_{\alpha,\gamma} = \{Bern_{x_0}(1 - \alpha - \gamma/2), Bern_{x_0}(1 - \alpha + \gamma/2)\}$. $(\alpha, \beta, \gamma, \delta)$-successful CLF coverage learning requires no samples. However, (α, β, δ)-successful label*

coverage learning requires samples.

Proof of validity of Example 2. A CLF-coverage set learner that always outputs $[1-\alpha-\gamma/2, 1-\alpha+\gamma/2]$ is valid for both distributions in $\mathcal{P}_{\alpha,\gamma}$. The CLF-coverage set has width γ and therefore satisfies $(1,\gamma)$-non-triviality (and therefore also $(1-\beta,\gamma)$-non-triviality). Therefore, such a learner is successful and does not require any input samples drawn from the generating distribution.

An (α,β,δ)-successful label-coverage set learner has to output the trivial coverage set for x_0 when the distribution is $\{\text{Bern}_{x_0}(1-\alpha-\gamma/2)$. Otherwise, the validity requirement is violated. It has to output the non-trivial coverage set - $\{0\}$ when the distribution is $\text{Bern}_{x_0}(1-\alpha+\gamma/2)$. Otherwise, the non-triviality requirement is violated. Therefore, from the output of any successful label-coverage set learner, we can distinguish between the distributions $\text{Bern}_{x_0}(1-\alpha-\gamma/2)$ and $\text{Bern}_{x_0}(1-\alpha+\gamma/2)$ with probability at least $1-\delta$. This means that any successful label-coverage set learner would need to use samples drawn from the generating distribution.

□

Example 3 (CLF-coverage set learning harder than label-coverage set learning)**.** *Let the domain be $\mathcal{X}=\{x_0\}$. For any coverage set learning parameters $0<\alpha<\frac{1}{2}, \beta, \gamma<\alpha/4, 0<\delta<\frac{1}{2}-\alpha$, consider the distribution family $\mathcal{P}_{\alpha,\gamma}=\{Bern_{x_0}(\alpha-\gamma), Bern_{x_0}(\alpha-4\gamma)\}$. (α,β,δ)-successful label-coverage set learning requires no samples. However, $(\alpha,\beta,\gamma,\delta)$-successful CLF-coverage set learning requires samples.*

Proof of validity of Example 3. A label-coverage learner that always outputs the non-trivial label-coverage set $\{0\}$ is a successful learner. The label-coverage set $\{0\}$ is valid for both distributions in $\mathcal{P}_{\alpha,\gamma}$ since the CLF of x_0 is smaller than α in both distributions. This coverage-set is also non-trivial. Therefore, this learner is successful and does not require any input samples drawn from the generating distribution.

On the other hand, an $(\alpha,\beta,\gamma,\delta)$-successful CLF-coverage set learner requires input samples drawn from the generating distribution. This is because such a learner can be used to distinguish between the two distributions in $\mathcal{P}_{\alpha,\gamma}$ with probability of success at least $1-\alpha-\delta>0$. Due to the validity requirement, with probability at least $1-\alpha-\delta$, the true CLF of the generating distribution lies in the output CLF-coverage set and the width of the CLF-coverage set is at most γ. Since the width is at most gamma and since the CLFs of the two distributions are more than 2γ distance away from each other, the CLF-coverage set contains only the CLF of the generating distribution and does not contain the CLF of the other distribution in the family. Therefore a distinguishing algorithm for the

family $\mathcal{P}_{\alpha,\gamma}$ that runs the successful CLF-coverage learner and identifies the distribution as the one with CLF contained in the output of the learner, has probability of success at least $1-\alpha-\delta$. Since this distinguishing algorithm requires samples, the successful CLF-coverage set learner must also require input samples drawn from the generating distribution.

□

In the rest of the section, we focus on analyzing how the sample complexities of distribution learning, CLF-coverage set learning, and Bayes classifier learning compare. We find that these three problems are in strictly decreasing order of sample complexity. In most of this section, we use another problem called *probabilistic concepts (p-concepts) learning* as a proxy for the CLF-coverage set learning. This problem setting, which was introduced by Kearns et al. [24], is closely related to the CLF-coverage set learning problem. Here is a formal definition of p-concepts learning:

Definition 10 (p-concepts error)**.** *Given a distribution P over $\mathcal{X}\times\{0,1\}$, and a conditional labelling function $l : \mathcal{X} \to [0,1]$, we define the p-concepts error of l relative to P as*

$$\mathcal{L}_P^{pc}(l) = \mathbb{E}_{X\sim P_{\mathcal{X}}}[|l(X) - l_P(X)|].$$

Definition 11 (p-concepts learning)**.** *A p-concepts 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 p-concepts 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*

$$\mathbb{P}_{S\sim P^m}[\mathcal{L}_P^{pc} \leq \epsilon] \geq 1-\delta.$$

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).

P-concepts learnability implies a form of CLF-coverage set learnability. This is the weakest form of CLF-coverage set learnability with domain-wide validity and point-wise non-triviality. This is stated as Lemma 3 in Section A.2 of the appendix. All forms of CLF-coverage set learnabilities imply p-concepts learnability. This is because even the weakest form with domain-wide validity and non-triviality implies p-concepts learnability. We state this result as Lemma 4 in Section A.2 of the appendix.

We first show that successful distribution learning of both the joint and marginal distributions implies successful p-concepts learning. Here is a formal definition of distribution learning in total variation distance:

Definition 12 (Total Variation(TV) distance)**.** *The total variation distance between two distributions over a domain $\mathcal{U}$, represented by their probability density functions (PDFs) p_1 and p_2 is defined by: $d_{TV}(p_1, p_2) = \int_{x \in \mathcal{U}} |p_1(x) - p_2(x)| dx$.*

Definition 13 (Distribution learner)**.** *A distribution-learner $\mathcal{A}$ over a domain $\mathcal{U}$ is a function that takes a sample S drawn from the distribution as input and outputs a density function $p : \mathcal{U} \to [0, 1]$.*

Definition 14 (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 $\mathcal{P}$.

Now we state the result that shows that the TV-distance learnability of the joint distribution P and the marginal distribution $P_{\mathcal{X}}$ implies p-concepts learnability.:

Theorem 5. *Let $\mathcal{P}$ be a family of distributions and let $\mathcal{P}_{\mathcal{X}}$ be the family of marginal distributions of distributions in $\mathcal{P}$. That is $\mathcal{P}_{\mathcal{X}} = \{P_{\mathcal{X}} : P \in \mathcal{P}\}$. Suppose that $\mathcal{P}$ and $\mathcal{P}_{\mathcal{X}}$ are TV distance distribution learnable with sample complexities $m_{TV,\mathcal{P}} : (0, 1)^2 \to \mathbb{N}$ and $m_{TV,\mathcal{P}_{\mathcal{X}}} : (0, 1)^2 \to \mathbb{N}$ respectively. Then $\mathcal{P}$ is is p-concepts learnable with sample complexity $m_{pc} : (0, 1)^2 \to \mathbb{N}$ with $m_{pc}(\epsilon, \delta) = \max\{m_{TV,\mathcal{P}}(\epsilon/2, \delta/2), m_{TV,\mathcal{P}_{\mathcal{X}}}(\epsilon/2, \delta/2)\}$.*

Proof. Let $\mathcal{A}$ denote the TV distribution learner of $\mathcal{P}$ with the sample complexity described in the theorem and let $\mathcal{A}_{\mathcal{X}}$ be the TV distribution learner of $\mathcal{P}_{\mathcal{X}}$. We describe how to construct a p-concepts learner using the two TV distribution learners. Let P be a distribution from the family $\mathcal{P}$. Let $p : \mathcal{X} \times \{0, 1\} \to [0, 1]$ denote the pdf of P and let $p_{\mathcal{X}} : \mathcal{X} \to [0, 1]$ denote the pdf of the marginal - $P_{\mathcal{X}}$. Recall that $l_P : \mathcal{X} \to [0, 1]$ denotes the conditional labelling function. Given a sample S drawn from P, let $\hat{p}$ denote the pdf of the probability distribution output by $\mathcal{A}$ for the sample S and let $\hat{p}_{\mathcal{X}}$ denote the pdf of $\mathcal{A}_{\mathcal{X}}(S)$. We denote the output of the p-concepts learner to be $\hat{l} : \mathcal{X} \to [0, 1]$. For each $x \in \mathcal{X}$, we define $\hat{l}(x) = \frac{p(x)}{p_{\mathcal{X}}(x)}$ for every $x \in \mathcal{X}$. We will show that such a learner that outputs $\hat{l}$ is a successful p-concepts learner.

We know that for a sample size M greater than both $m_{TV,\mathcal{P}}(\epsilon/2, \delta/2)$ and $m_{TV,\mathcal{P}_\mathcal{X}}(\epsilon/2, \delta/2)$ with probability at least $1-\delta$, $d_{TV}(\hat{p}, p) \leq \frac{\epsilon}{2}$ and $d_{TV}(\hat{p}_\mathcal{X}, p_\mathcal{X}) \leq \frac{\epsilon}{2}$.

$$
\begin{aligned}
\frac{\epsilon}{2} &\geq d_{TV}(\hat{p}, p) \\
&= \int_{x \in \mathcal{X}} \left(\left| p_\mathcal{X}(x) l_P(x) - \hat{p}_\mathcal{X}(x) \hat{l}(x) \right| + \left| p_\mathcal{X}(x)(1 - l_P(x)) - \hat{p}_\mathcal{X}(x)(1 - \hat{l}(x)) \right| \right) dx \\
&\geq \int_{x \in \mathcal{X}} \left| p_\mathcal{X}(x) l_P(x) - \hat{p}_\mathcal{X}(x) \hat{l}(x) \right| dx \\
&= \int_{x \in \mathcal{X}} \left| p_\mathcal{X}(x) l_P(x) - p_\mathcal{X}(x) \hat{l}(x) + p_\mathcal{X}(x) \hat{l}(x) - \hat{p}_\mathcal{X}(x) \hat{l}(x) \right| dx \\
&= \int_{x \in \mathcal{X}} \left| p_\mathcal{X}(x) l_P(x) - p_\mathcal{X}(x) \hat{l}(x) + p_\mathcal{X}(x) \hat{l}(x) - \hat{p}_\mathcal{X}(x) \hat{l}(x) \right| dx \\
&= \int_{x \in \mathcal{X}} \left| p_\mathcal{X}(x)(l_P(x) - \hat{l}(x)) + \hat{l}(x)(p_\mathcal{X}(x) - \hat{p}_\mathcal{X}(x)) \right| dx \\
&\geq \int_{x \in \mathcal{X}} p_\mathcal{X}(x) \left| l_P(x) - \hat{l}(x) \right| dx - \int_{x \in \mathcal{X}} \hat{l}(x) \left| p_\mathcal{X}(x) - \hat{p}_\mathcal{X}(x) \right| dx \\
&\geq \int_{x \in \mathcal{X}} p_\mathcal{X}(x) \left| l_P(x) - \hat{l}(x) \right| dx - \int_{x \in \mathcal{X}} \left| p_\mathcal{X}(x) - \hat{p}_\mathcal{X}(x) \right| dx \\
&\geq \mathbb{E}_{X \sim P_\mathcal{X}}[|\hat{l}(X) - l_P(X)|] - d_{TV}(\hat{p}_\mathcal{X}, p_\mathcal{X}) \\
&\geq \mathbb{E}_{X \sim P_\mathcal{X}}[|\hat{l}(X) - l_P(X)|] - \frac{\epsilon}{2}
\end{aligned}
$$

This shows that with probability at least $1-\delta$ over samples of size greater than $m_{pc}(\epsilon, \delta)$, $\mathbb{E}_{X \sim P_\mathcal{X}}[|\hat{l}(X) - l_P(X)|] \leq \epsilon$. Therefore, the p-concepts learner we constructed in the proof succeeds with sample complexity m_{pc}.

□

We saw that the learnability of the joint and marginal distributions implies p-concepts learnability and hence a form of CLF-coverage set learnability (with domain-wide requirements). Now we ask if CLF-coverage set learnability can be achieved with lower sample complexity compared to distribution learnability. There are cases when CLF-coverage set learning has strictly lower sample complexity than even just the marginal distribution learning problem. This is true even for the strictest form of CLF-coverage set learning that requires point-wise validity and point-wise non-triviality. We now describe an example of this case. First we introduce some notation that we use for this example.

The domain is the real line $\mathcal{X} = \mathbb{R}$. The family of distributions contain mixtures of unit variance Gaussians. For any $\mu > 0$, let $\mathcal{F}_\mu$ be the class of distributions:

$$\mathcal{F}_\mu = \left\{ \frac{1}{2}\mathcal{N}(x, 1) \times \{1\} + \frac{1}{2}\mathcal{N}(-x, 1) \times \{0\} : x \geq \mu \right\}.$$

For a large value of $\mu > 0$, $\mathcal{F}_\mu$, contains distributions with the positive and negative Gaussian components that are well-separated. When the separation increases, approximating the CLFs with low p-concepts-error becomes easier. As separation tends to infinity, the labelling function $l : \mathbb{R} \to [0, 1]$ becomes a good approximation to the CLF:

$$l(x) = \begin{cases} 0 & \text{if } x < 0; \\ 1 & \text{otherwise.} \end{cases}$$

However, the separation does not make learning the marginal distribution in TV distance easier. This is the intuition we use to construct an example where CLF-coverage set learning has higher sample complexity than marginal distribution learning.

We first show a lower bound on the sample complexity of marginal distribution learning of each distribution family $\mathcal{F}_\mu$. We later show that we can find a distribution family $\mathcal{F}_\mu$ with an arbitrarily low sample complexity. We state the lower bound on marginal distribution learning's sample complexity as the following lemma:

Lemma 1 (Lower bound for marginal distribution learning sample complexity)**.** *For every $\epsilon > 0$, $\delta = \frac{1}{3}$, for any $\mu > 0$, the sample complexity for (ϵ, δ)-TV distance learning of the marginals of family $\mathcal{F}_\mu$ is at least $m_{tv,\mathcal{F}_\mu}(\epsilon, \frac{1}{3}) \geq C\frac{1}{\epsilon^2}$ for a universal constant C.*

Proof. In this proof, we show that successful learning of the marginal implies that we can distinguish between two distributions having low KL divergence. Using a lower bound for the sample complexity of distinguishing between distributions with low KL divergence, we get the lower bound stated in the theorem for learning the marginal distribution.

We outline the steps of this proof and defer the full details of the proof to the appendix:

1. For any $m > 0$, let $\mathcal{D}_m = \frac{1}{2}\mathcal{N}(m, 1) \times \{1\} + \frac{1}{2}\mathcal{N}(-m, 1) \times \{0\}$. The magnitude of m indicates the separation of the two Gaussian components in the distribution $\mathcal{D}_m$. We pick two distributions: $\mathcal{D}_m$ and $\mathcal{D}_{m'}$ such that the TV distance between them is in the range $(\epsilon, 2\epsilon)$ and m, m' are large (both distributions have well-separated components). We describe why we can pick such distributions.

2. Since the TV distance between the distributions is greater than ϵ, successful $(\epsilon, 1/3)$-TV distance learning of the marginals implies that we can distinguish between $\mathcal{D}_m$ and $\mathcal{D}_{m'}$ with probability at least $\frac{2}{3}$.

3. We show that if we can distinguish between $\mathcal{D}_m$ and $\mathcal{D}_{m'}$ with probability at least $\frac{2}{3}$, then we can also distinguish between $\mathcal{N}(m, 1)$ and $\mathcal{N}(m', 1)$ with the same probability of success and the same sample complexity.

4. We show that the TV distance between $\mathcal{N}(m, 1)$ and $\mathcal{N}(m', 1)$ is at most 3ϵ. This is true since the TV distance between $\mathcal{D}_m$ and $\mathcal{D}_{m'}$ is less than 2ϵ and since both distributions have well-separated components.

5. Using the fact that the TV distance between $\mathcal{N}(m, 1)$ and $\mathcal{N}(m', 1)$ is small, we show that the KL divergence between these distributions is also small. Since successful marginal distribution implies successful distinguishing between $\mathcal{N}(m, 1)$ and $\mathcal{N}(m', 1)$, a lower bound on the sample complexity of distinguishing between $\mathcal{N}(m, 1)$ and $\mathcal{N}(m', 1)$ is also a lower bound on the sample complexity for learning the marginals.

□

The previous lemma provides a lower bound on the sample complexity of marginal distribution learning for all families $\mathcal{F}_\mu$. In the following theorem, we show that there is a class $\mathcal{F}_\mu$ with arbitrarily small sample complexity for learning CLF-coverage sets with point-wise validity and point-wise non-triviality. Combined with the previous lemma, this theorem shows that there is a class for which marginal distribution learning has higher sample complexity than CLF-coverage set learning.

Theorem 6. *For every $M \in \mathbb{N}$ and every $\epsilon > 0$, there exists a $\mu > 0$ such that the sample complexity for $(\epsilon, 1/3)-$p-concepts-learning $\mathcal{F}_\mu$ is at most M.*

Proof. For any $m > 0$, let $\mathcal{D}_m = \frac{1}{2}\mathcal{N}(m, 1) \times \{1\} + \frac{1}{2}\mathcal{N}(-m, 1) \times \{0\}$. The labelling probability for a point x according to this distribution is

$$\begin{aligned} f_m(x) &:= \frac{\exp(-(x-m)^2)}{\exp(-(x-m)^2) + \exp(-(x+m)^2)} \\ &= \frac{1}{1 + \exp(-2xm)} \end{aligned}$$

Consider $\mu(\epsilon, M) > \frac{1}{\sqrt{M}} \ln \frac{1-\epsilon}{\epsilon} + \frac{1}{\sqrt{M}}$. We will show that the $(\epsilon, 1/3)$-p-concepts-learning problem for $\mathcal{F}_{\mu(\epsilon,M)}$ can be reduced to the problem of finding, for each $\mathcal{D}_m \in \mathcal{F}_{\mu(\epsilon,M)}$, an

m' s.t. $|m - m'| < \frac{1}{\sqrt{M}}$ with probability $\frac{2}{3}$, based on samples from D_m. This problem has sample complexity at most $\sqrt{M}$

For any $m \geq \mu$, for any m' s.t. $|m - m'| < \Delta(M) = \frac{1}{\sqrt{M}}$, we will show that for every $x \in \mathbb{R}$, $|f_m(x) - f_{m'}| < \epsilon$.

$$\begin{aligned}
&|f_m(x) - f_{m'}(x)| \\
&= \frac{|\exp(-2xm) - \exp(-2xm')|}{(1 + \exp(-2xm))(1 + \exp(-2xm'))} \\
&\leq |\exp(-2xm) - \exp(-2xm')|
\end{aligned}$$

For $x < \bar{x} = \frac{1}{2\Delta(M)} \ln \frac{1}{1-\epsilon}$,

$$\begin{aligned}
|\exp(-2xm) - \exp(-2xm')| &\leq 1 - \exp(-2x|m - m'|) \\
&< 1 - \exp(-2\bar{x}|m - m'|) \\
&\leq \epsilon
\end{aligned}$$

For $x \geq \bar{x}$,

$$\begin{aligned}
|\exp(-2xm) - \exp(-2xm')| &< \exp\left(-2x \min\{m, m'\}\right) \\
&\leq \exp\left(-2\bar{x}(\mu(\epsilon, M) - \Delta)\right) \\
&\leq \exp\left(-2 \cdot \frac{\sqrt{M}}{2} \ln \frac{1}{1-\epsilon} \cdot \frac{1}{\sqrt{M}} \ln \frac{1-\epsilon}{\epsilon} \right) \\
&= \epsilon
\end{aligned}$$

□

We now move on to comparing CLF-coverage set learning and Bayes classifier learning. First we define the Bayes classifier learning problem. Recall that the Bayes classifier h^*_P is the classifier with the least possible error and is defined with the knowledge of P.:

Definition 15 (Bayes classifier learning of distributions)**.** *We say a family of distributions $\mathcal{P}$ is Bayes-classifier-learnable with sample complexity $m_{bayes,\mathcal{P}} : (0,1)^2 \to \mathbb{N}$ if there exists a learner $\mathcal{A}$ that takes as input a sample set S drawn from P and outputs a classifier $\mathcal{A}(S) : \mathcal{X} \to \{0, 1\}$ such that for any $\epsilon, \delta > 0$, any $m \geq m_{bayes,\mathcal{P}}(\epsilon, \delta)$ and any distribution $P \in \mathcal{P}$ we have*

$$\mathbb{P}_{S \sim P^m}[\mathcal{L}^{0/1}_P(\mathcal{A}(S)) \leq \epsilon + \mathcal{L}^{0/1}_P(h^*_P)] \geq 1 - \delta$$

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

CLF-coverage set learnability implies Bayes-classifier-learnability. This is true of even the easiest form of CLF-coverage learnability with domain-wide requirements of validity and non-triviality. We can show that when a CLF-coverage hypothesis r has good domain-wide validity and non-triviality, we can construct a classifier h, based on r, that has error close to the Bayes classifier. This results in the following observation:

Lemma 2. *Suppose r has $(1-\epsilon/3, 1-\epsilon/3, \epsilon/3)$-domain-wide-validity. Then consider the classifier $h : \mathcal{X} \to \{0,1\}$ constructed from r such that for each $x \in \mathcal{X}$,*

$$h(x) = \begin{cases} 1 & \text{if } \frac{1}{2} \notin r(x) \text{ and } r(x) \text{ contains values greater than } \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

Then $\mathcal{L}_P^{0/1}(\mathcal{A}(S)) \leq \epsilon + \mathcal{L}_P^{0/1}(h_P^*)$. That is, the excess error of h compared to h_P^* is at most ϵ.

Proof. Consider the following cases for points x in the domain. Let us consider each case and provide an upper bound on the excess error over h_P^* for points belonging to that case.

(C1) $l_P(x) \notin r(x)$: The probability of a point satisfying this condition is at most $\frac{\epsilon}{3}$ due to the validity of r. We can trivially upper bound the excess error of a point in this case by one. This results in an upper bound of $\frac{\epsilon}{3}$ for this case.

(C2) Width of $r(x)$ is greater than $\frac{\epsilon}{3}$: The probability of a point satisfying this condition is at most $\frac{\epsilon}{3}$ due to the non-triviality condition. Again, we can trivially upper bound the excess error of a point in this case by one. We again get an upper bound of $\frac{\epsilon}{3}$ for this case.

(C3) C1 and C2 do not hold and $\frac{1}{2} \notin r(x)$: When this condition is satisfied, $h(x) = h_P^*(x)$. Therefore, there is no contribution to the excess error by points satisfying this condition.

(C4) C1 and C2 do not hold and $\frac{1}{2} \in r(x)$: For points satisfying this condition, $|l_P(x) - 0.5| < \frac{\epsilon}{3}$. If $h(x) \neq h_P^*(x)$, then the excess error is at most $\frac{\epsilon}{3}$. This case contributes at most $\frac{\epsilon}{3}$ to the total excess error.

Summing up the contribution due to the different cases, we get that the excess error of h compared to h_P^* is at most ϵ.

□

Now we show that learning CLF-coverage sets can be harder than learning the Bayes classifier. We show this by constructing an example where p-concepts learning has higher sample complexity than Bayes classifier learning. Since all forms of CLF-coverage learnability imply p-concepts learnability (Lemma 4), we can conclude that in this example, all CLF-coverage set learning problems have higher sample complexity than Bayes classifier learning.

We don't need any samples to learn the Bayes classifier exactly for the classes $\mathcal{F}_\mu$ (the Bayes classifier is always the one that assigns label zero to negative domain elements and one to non-negative domain elements). However, for learning p-concepts, we will need samples. Note that given any $P \in \mathcal{F}_\mu$, the CLF of any different P' has positive p-concepts error with respect to P. Let us say that this p-concepts error is ϵ_μ. Then, for p-concepts learning of $\mathcal{F}_\mu$ with $\epsilon < \epsilon(\mu)$ and $\delta > \frac{1}{2}$, we will need samples since we will be able to distinguish between P and P' using such a learner. This proves that for every $\mathcal{F}_\mu$, there is an $\alpha(\mu), \beta(\mu), \gamma(\mu)$ such that the $(\alpha(\mu), \beta(\mu), \gamma(\mu), \delta)$-CLF-coverage learning problem (even domain-wide) has higher sample complexity compared to the $(\epsilon(\mu), \delta)$-Bayes classifier learning problem (follows from Lemma 4).

Chapter 7

Conclusion and future work

In this work, we developed a unifying framework that takes into account various desiderata for uncertainty quantification in binary classification:

- Identifying sources of uncertainty in binary classification – randomness of the training sample, randomness in the choice of a test point, and randomness in the label generation of a test point.
- Refined control on the randomness in the choice of a test point - standard PAC frameworks assume that the test point is drawn from the underlying distribution and incurs the randomness from this process. However, we are often interested in accuracy guarantees for specific test points and the standard average guarantees can be meaningless. We model this by defining region-conditional requirements. These are requirements for when test-points are drawn from the distribution, conditioned upon membership in predefined subsets that encode which test points are of interest.
- Ability to make trade-offs between validity and non-triviality - We have distinct quality measures of validity and non-triviality. In situations where the most important goal is to avoid incorrect predictions, we can set the validity parameter to be high. In situations where we have limited resources and can only deal with a limited number of regions marked as uncertain, we can set the non-triviality paramter to be high.

In this framework, we analyze the sample complexities for constructing valid and non-trivial coverage sets under a few different assumptions on the underlying distributions. We observe that unlabelled data is helpful for improving the quality of coverage sets. One

way of extending this framework is to extend to problems beyond binary classification such as multi-label classification and regression. Another extension is to also analyze the computational complexity of the coverage set learning problems. Even in the current framework, many questions remain open:

- The method proposed under the assumption that the distribution is approximated well by a function class uses only the empirical risk minimizer to construct label coverage sets. It would be interesting to see if we can use other classifiers as well. This might give rise to other properties beyond decisiveness that give rise to non-trivial coverage sets for regions.

- We analyze the implications of probabilistic concepts learning for CLF coverage set learning. Using p-concepts, we get CLF-coverage sets with only domain-wide guarantees. It would be interesting to see how to get stronger CLF-coverage sets guarantees, perhaps for certain families of distributions such as the well studied family of Gaussian mixture models. For this family, what are properties of regions for which we can get valid CLF-coverage sets with greater non-triviality?

- In the comparison of sample complexities of learning problems, the problems we compared often had different levels of difficulties. When we said that Problem 1 is easier than Problem 2, in all but one pair of problems, we were able to show that the hardest form of Problem 1 is easier than the easiest form of Problem 2. The exception to this is when we compared p-concepts learning and CLF-coverage learning. We showed that the easiest form of CLF-coverage learning implies p-concepts learning. However, we were only able to show that p-concepts learning implies the easiest form of CLF-coverage learning. It remains open in which cases (if any) p-concepts learning also implies harder forms of CLF-coverage learning (with region-conditional or point-wise requirements).

References

[1] 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.

[2] Hassan Ashtiani, Shai Ben-David, Nicholas Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Nearly tight sample complexity bounds for learning mixtures of gaussians via sample compression schemes. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

[3] Akshay Balsubramani. Learning to abstain from binary prediction. *arXiv preprint arXiv:1602.08151*, 2016.

[4] Akshay Balsubramani and Yoav Freund. Pac-bayes with minimax for confidence-rated transduction. *arXiv preprint arXiv:1501.03838*, 2015.

[5] David Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, USA, 2012.

[6] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference, 2020.

[7] 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.

[8] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

[9] Julian Bitterwolf, Alexander Meinke, and Matthias Hein. Certifiably adversarially robust detection of out-of-distribution data. *Advances in Neural Information Processing Systems*, 33, 2020.

[10] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

[11] Maxime Cauchois, Suyash Gupta, and John Duchi. Knowing what you know: valid confidence sets in multiclass and multilabel prediction. *arXiv preprint arXiv:2004.10181*, 2020.

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

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

[14] Ran El-Yaniv and Yair Wiener. Active learning via perfect selective classification. *The Journal of Machine Learning Research*, 13(1):255–279, 2012.

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

[16] Alex Gammerman, Volodya Vovk, and Vladimir Vapnik. Learning by transduction. *arXiv preprint arXiv:1301.7375*, 2013.

[17] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in neural information processing systems*, pages 4878–4887, 2017.

[18] 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.

[19] Shafi Goldwasser, Adam Tauman Kalai, Yael Tauman Kalai, and Omar Montasser. Beyond perturbations: Learning guarantees with arbitrary adversarial test examples. *arXiv preprint arXiv:2007.05145*, 2020.

[20] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

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

[22] 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.

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

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

[25] Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, pages 3792–3803, 2019.

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

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

[28] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[29] Jing Lei and Larry Wasserman. Distribution free prediction bands, 2012.

[30] 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.

[31] Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 71–96, 2014.

[32] Tianyang Li, Xinyang Yi, Constantine Carmanis, and Pradeep Ravikumar. Minimax Gaussian Classification and Clustering. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1–9, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

[33] Wei Liu, Frank Bretz, Natchalee Srimaneekarn, Jianan Peng, and Anthony J Hayter. Confidence sets for statistical classification. *Stats*, 2(3):332–346, 2019.

[34] Alexander Meinke and Matthias Hein. Towards neural networks that provably know when they don't know. *arXiv preprint arXiv:1909.12180*, 2019.

[35] 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.

[36] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. *CoRR*, abs/1004.4223, 2010.

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

[38] Alexandru Niculescu-Mizil and Rich Caruana. Obtaining calibrated probabilities from boosting. In *UAI*, volume 5, pages 413–20, 2005.

[39] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005.

[40] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[41] Joaquin Quiñonero-Candela, Masashi Sugiyama, Neil D Lawrence, and Anton Schwaighofer. *Dataset shift in machine learning*. Mit Press, 2009.

[42] Craig Saunders, Alexander Gammerman, and Volodya Vovk. Transduction with confidence and credibility. 1999.

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

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

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

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

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

[48] Vladimir Vovk. Conditional validity of inductive conformal predictors. In Steven C. H. Hoi and Wray Buntine, editors, *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pages 475–490, Singapore Management University, Singapore, 04–06 Nov 2012. PMLR.

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

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

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

[52] Chicheng Zhang and Kamalika Chaudhuri. The extended littlestone's dimension for learning with mistakes and abstentions. In *Conference on Learning Theory*, pages 1584–1616, 2016.

[53] Shengjia Zhao, Tengyu Ma, and Stefano Ermon. Individual calibration with randomized forecasting. *arXiv preprint arXiv:2006.10288*, 2020.

APPENDICES

Appendix A

Extended related work

A.1 Split conformal algorithm

The split conformal was algorithm introduced by Vovk et al. [49]. This algorithm is shown to provide distribution-free, marginal coverage by Lei and Wasserman [30]. The split conformal method partitions the sample into two parts - the training set and the validation set. The training set is used to train a model. The validation set is used to evaluate that model. The size of the coverage sets is determined by how well the trained model fits the validation set. When the model fits the validation set well, the algorithm outputs small coverage sets. A modification of this algorithm for regression conformal prediction satisfying a more refined guarantee than the marginal coverage is provided by Barber et al. [6]. Rather than simply guaranteeing coverage with high probability over test points drawn from the domain, the refined guarantee is for coverage with high probability over test-points drawn conditioned on membership in predefined subsets of the domain.

We now state this form of the split conformal algorithm by Barber et al.[6] as Algorithm 2.

We refer the reader to the paper of Barber et al.[6] for a proof that this algorithm yields coverage sets satisfying region-conditional validity. While this algorithm has the desired quality of providing distribution-free validity guarantees, under certain distributional assumptions, this algorithm could provide coverage sets with sub-optimal non-triviality. We show that this is the case under the distributional assumptions we consider in this work.

Algorithm 2 Split conformal algorithm for restricted conditional coverage

Input: Validity parameter: α, Collection of subsets of the domain: $\mathcal{B}$,
Labelled samples: $S = (x_i, y_i)_{i=1}^m$, Binary classification learning algorithm: $\mathcal{A}$,
Test point: x.
Output: Label coverage set for x
$S_t = \{(x_1, y_1), \ldots, (x_n, y_n)\}$
$S_v = \{(x_{n+1}, y_{n+1}), \ldots, (x_m, y_m)\}$
Default coverage set for x is $\hat{C}(x) = \{\mathcal{A}(S_t)(x)\}$.
for $B \in \mathcal{B}$ such that $x \in B$ **do**
  $N_{S_v}(B) = |S_v \cap B|$
  $\mathcal{L}^{0/1}_{S_v,B}(\mathcal{A}(S_t)) = |\{(x', y') : (x', y') \in S_v \cap B \text{ and } A(S_t)(x') \neq y'\}|$
  threshold $= N_{S_v}(B) - \left\lceil \left(1 - \alpha + \frac{1}{m-n}\right)(N_{S_v}(B) + 1) \right\rceil$
  if $\mathcal{L}^{0/1}_{S_v,B}(\mathcal{A}(S_t)) \geq$ threshold **then**
    Set coverage set of x to be trivial i.e. $\hat{C}(x) = \{0, 1\}$
  end if
end for
Return $\hat{C}(x)$

A.2 Probabilistic-concepts

Probabilistic concepts learning is closely related to the problem of CLF-coverage learning. We formally defined probabilistic concepts learning in Definition 11. In this section, we discuss the connection between probabilistic concepts and CLF-coverage sets.

Learnability of probabilistic concepts implies the learnability of CLF-coverage sets having domain-wide validity and non-triviality guarantees. The following lemma makes this concrete by showing how to construct CLF-coverage sets from a conditional labelling function with low p-concepts error:

Lemma 3. *An real-valued function $l : \mathcal{X} \to \{0, 1\}$ such that $\mathbb{E}_{X \sim P_{\mathcal{X}}}[|l(X) - l_P(X)|] \leq \epsilon$ can be used to construct a CLF-coverage hypothesis r with $(1 - \nu)$-domain-wide validity and $(1, 2\nu\epsilon)$-domain-wide non-triviality, for any $\nu \in (0, 1)$. r is such that for each $x \in \mathcal{X}$, $r(x) = \left[l(x) - \frac{\epsilon}{\nu}, l(x) + \frac{\epsilon}{\nu}\right]$.*

Proof. By the Markov inequality, for any $\nu \in (0, 1)$, $\mathbb{P}_{X \sim P}[|l_P(X) - \hat{l}(X)| > \epsilon/\nu] \leq \nu$. Which means that the CLF-coverage hypothesis based on $\hat{l}$ satisfies $(1-\nu)$-domain-validity. The CLF-coverage sets of all points have width $2\frac{\epsilon}{\nu}$. Therefore, this CLF-coverage hypothesis satisfies $(1, 2\epsilon/\nu)$-non-triviality. □

All forms of CLF-coverage learnability imply p-concepts-learnability. This is because even the weakest form of CLF-coverage learnability that requires only domain-wide validity and non-triviality implies p-concepts learnability. We state this as the following lemma:

Lemma 4. *A CLF-coverage set r that has $(1-\alpha)$-domain-wide validity and $(1-\beta,\gamma)$-domain-wide non-triviality yields an approximate CLF $l : \mathcal{X} \to [0,1]$ with p-concepts error at most $(2-\alpha-\beta+\gamma)$. That is, $\mathbb{E}_{X\sim P_{\mathcal{X}}}[|l(X) - l_P(X)|] \leq (2-\alpha-\beta+\gamma)$. l is such that for each $x \in \mathcal{X}$, $l(x) \in r(x)$.*

Proof. Due to $(1-\alpha)$-domain validity and $(1-\beta,\gamma)$-domain non-triviality, the probability weight of points satisfying at least one of the following two (bad) conditions is at most $\alpha+\beta$.

1. The true CLF lies outside the CLF-coverage set r.
2. The CLF-coverage set r has length more than γ.

For points satisfying one of the above conditions, we can trivially upper bound the difference between the true CLF and the CLF estimate obtained from the CLF-coverage set - l, by one. The weight of the points not satisfying the above conditions can be trivially bounded above by one. For all points not satisfying either condition, the difference between the true CLF and the CLF estimate can be bounded by γ. Therefore, we can bound the p-concepts error of l by $2-\alpha-\beta+\gamma$. □

Appendix B

Useful lemmas

Lemma 5 (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*

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

Lemma 6 (Bretagnolle-Huber inequality)**.** *Let P and Q be probability measures on the same measurable space $(\Omega, \mathcal{F})$, and let $A \in \mathcal{F}$ be an arbitrary event. Then, $P(A)+Q(A^c) \geq \frac{1}{2}\exp(-KL(P,Q))$. Here, $KL(P,Q)$ is the KL-divergence between P and Q.*

Lemma 7. *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

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

Proof. 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. □

Lemma 8. *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 7.

Proof of Lemma 8. 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 $1 - \delta$, $|\bar{\ell}_P(X') - \hat{\ell}(X', S)| < w_\ell(m, \delta, \hat{p})$. □

Appendix C

Extended proofs

Proof of validity of Example 1. Here we expand on the proof outline from Chapter 5 and provide all the full calculations that were omitted in the outline. There are three parts to this proof:

1. Showing that the baseline CEB method returns trivial coverage sets: We show that the bound provided by Theorem 3 for B with sample failure parameter $\delta = 0.15$ is greater 1.0. This implies trivial coverage sets.

   Note that a lower bound for this bound is

   $$\frac{\epsilon_{UC}(|S_l|, 0.15/2)}{\frac{|S_u \cap B|}{|S_u|}} = \frac{\sqrt{\frac{9(1+\log(2/0.15))}{150}}}{\frac{|S_u \cap B|}{|S_u|}} > \frac{0.35 \cdot 10^7}{|S_u \cap B|}.$$

   The expected value of $|S_u \cap B|$ is $10^7 \cdot 0.01 = 10^5$. With high probability, $|S_u \cap B|$ is not much larger than 10^5. By the Hoeffding inequality,

   $$\mathbb{P}\left[|S_u \cap B| > 10^5 + \sqrt{\frac{10^7}{2} \ln 4}\right] \leq \frac{1}{10}.$$

With probability at least 0.9,

$$\begin{aligned}\frac{|S_u \cap B|}{|S_u|} &< \frac{10^5}{10^7} + 10^{-3.5}\sqrt{\frac{\ln 10}{2}} \\ &< 0.013.\end{aligned}$$

Since $0.35 > 0.013$, the bound given by Theorem 3 is bigger than 1.0.

2. Showing that the split conformal algorithm returns trivial coverage sets: First we show that with probability at least 0.52, the number of validation samples that lie in the region B is at most six. Then we show that this low number of validation samples in B implies that the split conformal method returns trivial coverage sets.

   (a) Showing that there are few validation samples in B. The expected size of $|S_v \cap B|$ is $P(B)|S_v| = 0.01 \cdot 75 = 0.75$. By applying the Hoeffding inequality, we get that

   $$\begin{aligned}\mathbb{P}\left[|S_v \cap B| \geq 6\right] &\leq \exp\left(-\frac{2(6 - P(B)|S_v|)^2}{|S_v|}\right) \\ &\leq \exp\left(-\frac{2(6 - 0.75)^2}{75}\right) \\ &< 0.48.\end{aligned}$$

   (b) Showing that split conformal algorithm returns trivial coverage sets when $|S_v \cap B| \leq 6$. Recall that the split conformal algorithm (Algorithm 2) calculates a threshold value and the number of errors in $S_v \cap B$ made by the empirical risk minimizer from $\mathcal{H}_{\text{thresholds}}$. The split conformal algorithm returns trivial coverage sets if the errors in $S_v \cap B$ is greater than the threshold value. If the threshold value is negative, then the split conformal algorithm returns trivial coverage sets regardless of the number of errors in $S_v \cap B$. We will now show that when $|S_v \cap B| \leq 6$, the threshold value is negative. The threshold is at most

   $$|S_v \cap B| - \left\lceil \left(1 - \alpha + \frac{1}{|S_v|}\right)(|S_v \cap B| + 1) \right\rceil.$$

   When $|S_v \cap B| \leq 6$, this threshold value is negative.

3. Showing that the decisiveness-based CEB method returns non-trivial coverage sets: We first show that with probability at least 0.62 over samples, the decisiveness of

the region B is the highest value - one. Like in the first part, we also show that the fraction of sample points that lie in B is at least 0.013 with probability at least 0.9. The high decisiveness combined with the lower bound on the fraction of samples that lie in B results in the conditional error bound in Theorem 4 with sample failure parameter $\delta = 0.15$ being less than 0.85. This results in non-trivial coverage sets for $(0.85, 0.85)$-region-conditional validity.

To show that decisiveness is one with high probability, we show that any classifier in $H_{\text{thresholds}}$ that labels any point in $S_u \cap B$ zero, has high sample error with high probability. This shows that all classifiers with low empirical error have the same behaviour on $S_u \cap B$ - labelling all points in $S_u \cap B$ zero. And therefore the decisiveness is one.

(a) Showing that the decisiveness of set B is one with high probability over the samples. We show this by showing that all classifiers in $\mathcal{H}_{\text{thresholds}}$ having sample error within $2\epsilon_{UC}\left(|S_l|, \frac{1}{4} \cdot 0.15\right)$ of the optimal sample error all label all points in $S_u \cap B$ zero. First we show that the sample error of the classifier $h_{\frac{1}{2}}$ is at most 0.0686 with probability at least 0.84. This implies that the sample of the empirical risk minimizing classifier is also at most 0.0686. We show this by applying the Hoeffding inequality. The expected sample error is 0.001.

$$\begin{aligned}\mathbb{P}\left[\mathcal{L}^{0/1}_{S_l,B} \geq 0.1\right] &\leq \exp\left(-\frac{2(12-0.15)^2}{150}\right) \\ &< 0.16.\end{aligned}$$

All classifiers with sample error within $2\epsilon_{UC}(100, 0.15/4)$ have sample error at most $0.0686 + 2 \cdot 0.34 = 0.77$.

$$\begin{aligned}0.686 + 2\epsilon_{UC}\left(|S_l|, \frac{1}{4} \cdot 0.15\right) &= 0.686 + 2\sqrt{\frac{9(1+\log(4/0.15))}{150}} \\ &< 0.832\end{aligned}$$

Now we show that with high probability any classifier that labels some point in B one (a classifier $h_a \in \mathcal{H}_{\text{thresholds}}$ with $a < 0.01$) has sample error larger than 0.832. We first show that there are few labelled samples in B similar to how we showed that there are few validation samples in B. With probability at least

0.77, $|S_l \cap B| \leq 12$.

$$\begin{aligned}\mathbb{P}\left[|S_l \cap B| \geq 12\right] &\leq \exp\left(-\frac{2(12 - P(B)|S_l|)^2}{|S_l|}\right) \\ &< 0.23.\end{aligned}$$

Next we show that of the labelled sample points not in B, most have labels different from the labels assigned by a classifier $h_a \in \mathcal{H}_{\text{thresholds}}$ with $a < 0.01$. A labelled sample not in B with label agreeing with a classifier $h_a \in \mathcal{H}_{\text{thresholds}}$ differs from the label assigned to it by the classifier $h_{\frac{1}{2}}$. We have already shown that most labelled samples have labels agreeing with the classifier $h_{\frac{1}{2}}$. Therefore, the number of labelled samples in $S_l \setminus B$ that are correctly labelled by a classifier h_a with $a < 0.01$ is upper bounded by the number of labelled samples on which $h_{\frac{1}{2}}$ makes an error. We have shown that this is at most 12 with probability at least 0.832. Therefore, any classifier h_a with $a < 0.01$ makes error on at least $(150 - 12 - 12)$ labelled sample points with probability at least 0.62. This concludes our proof that the decisiveness of the set B is one with probability at least 0.62.

(b) Showing a lower-bound on the number of unlabelled samples in B. By Lemma 7, with probability at least $\frac{1}{10}$, $\frac{|S_u \cap B|}{|S_u|} > 0.01 - \sqrt{\frac{1}{2|S_u|}\ln 10}$.

Therefore, with probability at least 0.52, the conditional generalization error of the empirical risk minimizer can be bounded above by 0.15, using Theorem 4. The decisiveness-based method for $(1 - \alpha) = 0.85$-region conditional validity returns non-trivial coverage sets for all points in the set B.

□

Proof of Lemma 1. In this proof, we show that successful learning of the marginal implies that we can distinguish between two distributions having low KL divergence. Using a lower bound for the sample complexity of distinguishing between distributions with low KL divergence, we get the lower bound stated in the theorem for learning the marginal distribution.

Let $m' \geq m \geq \mu$ be such that

- $\epsilon < TV(\mathcal{D}_{m'}, \mathcal{D}_m) < 2\epsilon$
- $\mathcal{N}(m', 1)((\infty, 0)) < \mathcal{N}(m, 1)((\infty, 0)) < \epsilon$.

For every m, since as m' approaches infinity, $TV(\mathcal{D}_{m'}, \mathcal{D}_m)$ approaches one, there should exist an m' such that $TV(\mathcal{D}_{m'}, \mathcal{D}_m) > \epsilon$. Since $TV(\mathcal{D}_{m'}, \mathcal{D}_m)$ is continuous in m and m', there should exist m' such that $TV(\mathcal{D}_{m'}, \mathcal{D}_m)$ is also less than 2ϵ. In particular, we can choose m large enough so that the second condition is also satisfied. Then, successful $(\epsilon, 1/3)$-TV distance learning of the marginals of family $\mathcal{F}_\mu$ implies that we can distinguish between the distributions $\mathcal{D}_m$ and $\mathcal{D}_{m'}$ with probability at least $\frac{2}{3}$. We can do this by learning the marginal and choosing the distribution out of $\mathcal{D}_m$ and $\mathcal{D}_{m'}$ which is closest in TV distance to the learnt marginal. This works because $TV(\mathcal{D}_{m'}, \mathcal{D}_m) > \epsilon$.

We now show that distinguishing between distributions $\mathcal{D}_m$ and $\mathcal{D}_{m'}$ with probability at least $\frac{2}{3}$ implies that we can also distinguish between the distributions $\mathcal{N}(m, 1)$ and $\mathcal{N}(m', 1)$ with probability at least $\frac{2}{3}$ with the same sample complexity. Let $\mathcal{A}_\mathcal{D}$ denote an algorithm that takes samples of size n drawn from one of $\mathcal{D}_m$ or $\mathcal{D}_{m'}$ and identifies the distribution the samples are drawn from with probability of success at least $\frac{2}{3}$. We now describe a distinguishing algorithm $\mathcal{A}_\mathcal{N}$ for $\mathcal{N}(m, 1)$ and $\mathcal{N}(m', 1)$ that takes as input a sample $S = (X_1, \ldots, X_n)$ where each X_i is drawn i.i.d. from the distribution $\mathcal{P}$ which is one of $\mathcal{N}(m, 1)$ and $\mathcal{N}(m', 1)$. Consider a random variable Z_i that takes value X_i with probability $\frac{1}{2}$ and value $-X_i$ with probability $\frac{1}{2}$, for each $i \in \{1, \ldots, n\}$. Each Z_i is a random variable of the distribution $\mathcal{D}_m$ if X_i is a random variable of $\mathcal{N}(m, 1)$ and is a random variable of $\mathcal{D}_{m'}$ if X_i is a random variable of $\mathcal{N}(m', 1)$. $\mathcal{A}_\mathcal{N}$ runs $\mathcal{A}_\mathcal{D}$ on the sample $S' = (Z_1, \ldots, Z_n)$. The output of $\mathcal{A}_\mathcal{N}$ is $\mathcal{N}(m, 1)$ if the output of $\mathcal{A}_\mathcal{D}$ is $\mathcal{D}(m, 1)$. And the output of $\mathcal{A}_\mathcal{N}$ is $\mathcal{N}(m, 1)$ otherwise. With sample size n, $\mathcal{A}_\mathcal{N}$ distinguishes between samples drawn from $\mathcal{N}(m, 1)$ and $\mathcal{N}(m', 1)$. In the rest of the proof we will find a lower bound on the sample complexity to distinguish between $\mathcal{N}(m, 1)$ and $\mathcal{N}(m', 1)$.

We start by finding an upper bound on $|m-m'|$ implied by the fact that $TV(\mathcal{D}_{m'}, \mathcal{D}_m) < 2\epsilon$. First note that $TV(\mathcal{D}_{m'}, \mathcal{D}_m) < 2\epsilon$ implies that $TV(\mathcal{N}(m, 1), \mathcal{N}(m', 1)) < 2\epsilon + \epsilon$ (which also means $TV(\mathcal{N}(m, 0), \mathcal{N}(m', 0)) < 2\epsilon + \epsilon$). Let $p_{0,m}$ denote the pdf of $\mathcal{N}(-m, 1)$ and let Let $p_{1,m}$ denote the pdf of $\mathcal{N}(m, 1)$.

$$
\begin{aligned}
2\epsilon &= TV(\mathcal{D}_{m'}, \mathcal{D}_m) \\
&= \int_{-\infty}^{\infty} \frac{1}{2} |p_{0,m}(x) + p_{1,m}(x) - p_{0,m'}(x) - p_{1,m'}(x)| dx \\
&= \int_{0}^{\infty} |p_{0,m}(x) + p_{1,m}(x) - p_{0,m'}(x) - p_{1,m'}(x)| dx
\end{aligned}
$$

Since $m' > m$, $p_{0,m}(x) > p_{0,m'}(x)$ for $x \geq 0$. Using this fact and the triangle inequality,

$$\begin{aligned}
&\geq \int_0^\infty |p_{1,m}(x) - p_{1,m'}(x)|dx - \int_0^\infty p_{0,m}(x) - p_{0,m'}(x)dx \\
&> \int_0^\infty |p_{1,m}(x) - p_{1,m'}(x)|dx - \int_0^\infty p_{0,m}(x)dx \\
&\geq \int_0^\infty |p_{1,m}(x) - p_{1,m'}(x)|dx - \epsilon \qquad \text{(By choice of } m\text{)} \\
\implies 3\epsilon &> TV(\mathcal{N}(m,1), \mathcal{N}(m',1)).
\end{aligned}$$

Theorem 1.3 of [12] states a lower bound on the difference in the means of two Gaussians in terms of their TV distance. From this theorem, we get that $TV(\mathcal{N}(m,1), \mathcal{N}(m',1)) < 3\epsilon$ implies $|m - m'| < C'\epsilon$ for a universal constant C.

Finally, we provide a lower bound on the sample complexity of distinguishing between two univariate Gaussians – $\mathcal{N}(m,1)$ and $\mathcal{N}(m',1)$ with probability of success at least $\frac{2}{3}$ in terms of $|m-m'|$. First note that the KL divergence between these distributions is $\frac{|m-m'|^2}{2}$. For a sample size of n, the KL divergence between $\mathcal{N}(m,1)^n$ and $\mathcal{N}(m',1)^n$ is $\frac{n|m-m'|^2}{2}$ Then we can apply the Bretagnolle-Huber inequality (stated in the appendix as Lemma 6) to show that the probability of failure for any distinguishing algorithm for $\mathcal{N}(m,1)^n$ and $\mathcal{N}(m',1)^n$ is at least:

$$\begin{aligned}
\frac{1}{2}\exp(-KL(\mathcal{N}(m,1)^n, \mathcal{N}(m',1)^n)) &\geq \frac{1}{2}\exp\left(\frac{n|m-m'|^2}{2}\right) \\
&\geq \frac{1}{2}\exp\left(\frac{C'n\epsilon^2}{2}\right)
\end{aligned}$$

In order for the probability of failure to be less than $\frac{1}{3}$,

$$\begin{aligned}
\frac{1}{3} &< \frac{1}{2}\exp\left(\frac{C'n\epsilon^2}{2}\right) \\
\implies n &> \frac{2}{C'\epsilon^2}\ln\frac{3}{2}.
\end{aligned}$$

This proves the lower bound on the sample complexity stated in the lemma.

□