University of Alberta

PAC-LEARNING WITH LABEL NOISE

by

Shahin Jabbari Arfaee

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

Spring 2011
Edmonton, Alberta

Examining Committee

Robert C. Holte, Computing Science

Sandra Zilles, [Computer Science, University of Regina], Adjunct Computing Science

Mark Lewis, Mathematical and Statistical Sciences

Russell Greiner, Computing Science

Abstract

One of the main criticisms of previously studied label noise models in the PAC-learning framework is the inability of such models to represent the noise in real world data. In this thesis, we study this problem by introducing a framework for modeling label noise and suggesting four new label noise models. We prove positive learnability results for these noise models in learning simple concept classes and discuss the difficulty of the problem of learning other interesting concept classes under these new models. In addition, we study the previous general learning algorithm, called the minimum pn-disagreement strategy, that is used to prove learnability results in the PAC-learning framework both in the absence and presence of noise. Because of limitations of the minimum pn-disagreement strategy, we propose a new general learning algorithm called the minimum nn-disagreement strategy. Finally, for both minimum pn-disagreement strategy and minimum nn-disagreement strategy, we investigate some properties of label noise models that provide sufficient conditions for the learnability of specific concept classes.

Acknowledgements

I am very thankful to my supervisors, Robert C. Holte and Sandra Zilles, for all the encouragement, guidance and support throughout this work.

Contents

List of Tables

List of Figures

List of Symbols

symbol	interpretation	symbol	interpretation
c	concept	PAC	probability approximately correct
c^*	target concept	$\mathrm{PAC}_{\underline{\epsilon},\underline{\delta}}$	PAC under noise
$\mathcal{C}$	concept class	Pr	probability
$\mathcal{C}^1_{\mathrm{thr}}$	one-dimensional threshold functions	$\mathcal{S}$	sample
$\mathcal{C}^2_{\mathrm{thr}}$	two-dimensional axis-parallel halfspaces	supp	support of the distribution
cpcn	constant-partition classification noise	$\mathbb{R}$	the set of real numbers
d	VC-dimension of a concept class	rcn	random classification noise
D	distribution	$\mathrm{VC\text{-}dimension}$	Vapnik-Chervonenkis dimension
$\hat{D}$	estimated distribution	WB	weight ball
$\mathcal{D}_{\mathcal{X}}$	the class of all possible distributions	wmball	weight malicious classification noise
DB	distance ball	wrball	weight random classification noise
dist	metric	$\mathcal{X}$	input space
dmball	distance malicious classification noise	x	instance
drball	distance random classification noise	y	label
E	expectation	δ	probability of failure
EX	oracle	$\underline{\delta}$	probability of failure under noise
EX_{Φ}	noisy oracle	Δ^{nn}	nn-difference
$\mathcal{F}_{\mathrm{nn}}$	nn-difference	Δ^{pn}	pn-difference
$\hat{\mathcal{F}}_{\mathrm{nn}}$	nn-disagreement	Δ^{pp}	pp-difference
$\mathcal{F}_{\mathrm{pn}}$	pn-disagreement	ϵ	error
L	learning algorithm in other models	$\underline{\epsilon}$	error under noise
$\mathcal{L}$	learning algorithm in PAC model	η	noise parameter in classic noise models
mcn	malicious classification noise	ρ	radius of the distance ball
$\mathbb{N}$	the set of natural numbers	Φ	label noise
nr	noise rate	ω	supremum mass of the weight ball

Chapter 1

Introduction

1.1 Problem definition

The classification task is an important problem studied in the field of machine learning. The goal is to learn the behavior of a function by just observing some examples of the form (input, output) of the function. This function is a mapping from a set called the input space to another finite set. A specific version of this task is when the latter set is $\{0, 1\}$. This problem is called binary classification.

There are many learning models proposed for the binary classification task. This thesis is based on Valiant's probably approximately correct (PAC) model of learning [37]. In this model, the goal is to find an algorithm, called the learning algorithm, that finds an arbitrarily accurate estimate of the underlying binary valued function, called the target concept, by observing a number of examples of it. It is assumed that the input part of each example is drawn according to some unknown probability distribution over the input space. In the PAC-learning model, the input part of any example is called the instance and the output part of it is called the label. It is assumed that the learning algorithm is provided with a set of binary valued functions from which to choose its estimate of the target concept. Any function in this set is called a concept and the set itself is called the concept class. The learning algorithm is required to return highly accurate estimates of the target concept no matter what the underlying target concept and distribution are, but since the process of drawing examples from the target concept is stochastic, the learning algorithm is allowed to fail with some probability. Finally, the learning algorithm is only allowed to use a number of examples that is polynomial in the inverse of its estimation error and the inverse of its probability of failure.

In the PAC-learning model it is assumed that the learning algorithm is only provided with noise-free examples. However, in real world applications, the source from which the examples are drawn may be erroneous. For example, consider the task of classifying a set of MRI brain images into two groups, the group of MRI images showing a tumor in the brain and the group without a tumor. The human who classifies the instances may make mistakes on some instances due to, for example, tiredness. Several different kinds of scenarios can be considered about the origin of noise in real world applications. However as mentioned above it is usually far from reality to assume that there is no noise in the examples.

Noise can be considered as any process that distorts the examples. Different kinds of noise model have been considered in the PAC-learning framework. We divide these noise models into two categories in this thesis. In the first category, called label noise, only the labels of examples are exposed to noise. In the second type, either merely the instance or both instance and label are noisy. The focus of this thesis is on the former.

One of our main criticisms of previously proposed noise models in the PAC-learning literature is that such models cannot describe the real world scenarios of why noise happens in practice. Although the mathematical analysis of these models is interesting and insightful, we feel that there is a notable gap between real world noise models and noise models that can be perfectly analyzed

using mathematics. This is the first problem we address in this thesis.

Also, to our knowledge, there is no attempt in unifying different types of label noise in the PAC model and study interesting properties of them. This is the second problem we address in this thesis.

Finally, although the approach that is mainly used to deal with noise in the previously studied noise model is universal for some types of noise (and the noise-free case), it is no longer universal for more general types of noise. In this approach, which we call the minimum pn-disagreement strategy, the learning algorithm, after observing a number of examples, returns a concept from the concept class that has the smallest number of wrong label predictions on the instances of these examples. Generalizing the minimum pn-disagreement strategy is the last problem we address in this thesis.

1.2 Approach to the problem

We try to solve the first problem regarding realistic noise models by introducing four new label noise models. We call these models locally variable noise models due to the fact that an instance can only get a noisy label if it has instances with different labels in its proximity. We define the proximity based on two different measures. One measure can be simply defined using any arbitrary metric and the other one can be defined using the amount of the probability mass around an instance. We then show some positive and negative learnability results with respect to our new noise models for simple concept classes.

We address the second problem regarding unifying the noise models by introducing a new label noise framework. In this new framework, label noise is defined as a function over the input space, the concept class, and the class of all possible distributions over the input space. We then show that this new framework can describe many of the previously proposed label noise models in the PAC-learning framework.

To address the last problem regarding the generality of the minimum pn-disagreement strategy, we propose a new learning algorithm for learning in the presence of noise called the minimum nn-disagreement strategy. In this new method, the learning algorithm first tries to find out how concepts may change in the presence of noise. Then the learning algorithm returns a concept that after being changed by the noise, labels the instances of the examples similar to the labels of examples that the learning algorithm receives. Finally, we investigate general properties of noise models that make some concept classes learnable with respect to such noise models.

1.3 Contributions

The contributions of this work can be listed as follows.

1. First, a new framework for label noise is presented in this thesis. We show that almost all the previously studied label noise models in the PAC setting can be modeled in this new framework.

2. We introduce four new label noise models. These models are important as they model the label noise in a more realistic way than previously studied label noise models. We also show some positive and negative learnability results for these new noise models regarding some simple concept classes.

3. We study general characteristics of label noise models that make learning of certain concept classes possible. We also introduce a new general learning algorithm and introduce some sufficient learnability conditions when we use this specific learning algorithm.

1.4 Outline

This thesis is organized as follows. Chapter 2 introduces some preliminaries and the formal definition of our framework for PAC-learning in the presence of noise. It further demonstrates how previously studied noise models can be cast into our new framework.

In Chapter 3, our locally variable noise models are introduced and a few learnability results for simple concept classes with respect to these noise models are reported.

Chapter 4 deals with learning from noisy examples when the input space is finite. Finiteness of the input space in general eases learning.

In Chapter 5, the minimum pn-disagreement strategy for PAC-learning in the presence of noise is studied in more detail. We show the applications of this strategy in a previously studied label noise model. We also study some general characteristics of noise models that provide sufficient conditions for learning concept classes with respect to such noise models.

Our new general learning algorithm, the minimum nn-disagreement strategy, for PAC-learning under noise is presented in Chapter 6. As in Chapter 5, we also propose some general characteristics of noise models that give sufficient conditions for learning concept classes in the presence of noise using our proposed method.

Other research related to noise models in PAC-learning is reviewed in Chapter 7, which also includes a brief review of the noise models in other learning frameworks.

Chapter 8 summarizes this thesis, draws conclusions and outlines directions for future research.

Chapter 2

Formal Framework for Learning with Label Noise

In this chapter, we define the basic notation that we use throughout this thesis and introduce our generic framework for learning with label noise. More notation will be introduced in the next chapters when needed. Many of the definitions in this section and Section 2.1 are adapted from the textbook by Kearns and Vazirani [25].

By $\mathbb{N}$ we denote the set of all natural numbers, including 0. $\mathbb{R}$ denotes the set of all real numbers. If A is any arbitrary set, 2^A denotes the power set of A, *i.e.*, the set of all subsets of A, and $|A|$ denotes the cardinality of A, where $|A| = \infty$ if A is an infinite set.

We denote by $\mathcal{X}$ an arbitrary metric space called the *input space* and by dist an arbitrary metric on $\mathcal{X}$. In most cases below, $\mathcal{X}$ will be either finite or equal to $\mathbb{R}^n$ for some $n \in \mathbb{N}$.

A *concept* c over $\mathcal{X}$ is a subset of $\mathcal{X}$ or, equivalently, a binary-valued function on $\mathcal{X}$. Hence we use $c \subseteq \mathcal{X}$ and $c : \mathcal{X} \to \{0,1\}$ interchangeably to refer to a concept c. A *concept class* over $\mathcal{X}$ is a set of concepts over $\mathcal{X}$, typically denoted by $\mathcal{C}$.

A *probabilistic concept* $c : \mathcal{X} \to [0,1]$ over $\mathcal{X}$ is a real-valued function that assigns to each element of $\mathcal{X}$ a value in the closed interval $[0,1]$. A probabilistic concept can be treated like a set in which membership is probabilistic. Intuitively, for any $x \in \mathcal{X}$, $c(x) = p$ indicates that, with probability p, x belongs to the set associated with c, and with probability $1-p$, x does not belong to the set associated with c. Note that a concept is a special case of a probabilistic concept.

Let $\mathcal{D}_{\mathcal{X}}$ denote the class of all *probability distributions* over $\mathcal{X}$. If $D \in \mathcal{D}_{\mathcal{X}}$ is a probability distribution over $\mathcal{X}$ and c is a probabilistic concept over $\mathcal{X}$ then the *oracle* $\mathrm{EX}(c, D)$ is a procedure that on each call returns a pair $(x, y) \in \mathcal{X} \times \{0,1\}$, called an *example*, where *(i)* $x \in \mathcal{X}$ is drawn with respect to the distribution D and *(ii)* $y \in \{0,1\}$ is drawn with respect to the Bernoulli distribution over $\{0,1\}$ that assigns the probability $c(x)$ to 1 and the probability $1 - c(x)$ to 0. In an example (x, y), x is usually called the *instance* and y is called the *label*. Note that repeated calls to $\mathrm{EX}(c, D)$ are always treated as independent samplings. If $c : \mathcal{X} \to \{0,1\}$ is a concept, then the label y of the example (x, y) returned by $\mathrm{EX}(c, D)$ is uniquely defined by the instance x.

The following is a standard definition in many mathematics textbooks such as the book by Rudin [30].

Definition 2.1. *Let $\mathcal{X}$ be an input space and D a distribution. The* support of distribution *D, denoted by $\mathrm{supp}(D)$, is the smallest closed set $X \subseteq \mathcal{X}$ with $Pr_{x \sim D}[x \in \mathcal{X} - X] = 0$.*

Therefore, all the instances of examples returned from an oracle are in $\mathrm{supp}(D)$ and no instance can be sampled from $\mathcal{X} - \mathrm{supp}(D)$.

Every multi-set $\mathcal{S}$ of elements in $\mathcal{X} \times \{0,1\}$ is called a *sample* over $\mathcal{X}$. Note that it is important to consider samples as multi-sets, since, as is typical in statistical machine learning, the multiplicity with which examples occur in a sample will contain important information.

The problem that we are dealing with in this thesis is a classification task, which is an important machine learning problem. In a classification task, upon seeing a sample of examples (known as the training set) from an oracle $\mathrm{EX}(c, D)$, a procedure predicts the label of potentially unseen instances (known as the test set) drawn from D. This procedure is usually called the *learning algorithm*[1] and c is called the *target concept*. If $c : \mathcal{X} \to \{0, 1\}$, the classification task is called *binary classification*.

For the rest of this document, we do not explicitly mention the input space $\mathcal{X}$ that concept classes, distributions and samples are defined over, when $\mathcal{X}$ is clear from the context.

2.1 PAC-learning framework

We focus on a specific framework for classification, called PAC-learning, which is due to Valiant [37]. The formal definition of this framework is as follows.

Definition 2.2. ***(Valiant [37])*** *A concept class $\mathcal{C}$ is* probably approximately correctly learnable *(*PAC*-learnable), if there exists a learning algorithm $\mathcal{L}$ and a $m : (0, \frac{1}{2}) \times (0, \frac{1}{2}) \to \mathbb{N}$ such that: for any target concept $c^* \in \mathcal{C}$, for all $\epsilon, \delta \in (0, \frac{1}{2})$ and for any distribution $D \in \mathcal{D}_{\mathcal{X}}$, if $\mathcal{L}$ is given access to $\mathrm{EX}(c^*, D)$ and inputs ϵ and δ, then with probability at least $1 - \delta$, after seeing a sample $\mathcal{S}$ of $m(\epsilon, \delta)$ examples, where $m(\epsilon, \delta)$ is polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$, $\mathcal{L}$ outputs a concept $c \in \mathcal{C}$ satisfying $Pr_{x \sim D}[c(x) \neq c^*(x)] \leq \epsilon$.*

A concept class is PAC-learnable if a polynomial number of examples is, with high probability, sufficient to find a concept in the class that disagrees with the target concept only in a low-probability subset of the input space, no matter what the underlying distribution is on the input space. The error of such a concept, ϵ, is caused by the fact that the learner sees only a finite number of examples. The learner is not required to succeed all the time, it is allowed to fail with probability of at most δ. This will happen when the examples are not representative of the underlying distribution[2]. By changing the size of the sample in the PAC-learning framework, the error (ϵ) and the probability of failure of the learner (δ) can be set to be arbitrarily small.

In the literature, for some concept classes polynomial-time learning algorithms $\mathcal{L}$ are proposed (for example [2, 37]). A concept class $\mathcal{C}$ is *efficiently* PAC-learnable if it is PAC-learnable with a learning algorithm $\mathcal{L}$ that runs in time polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$.[3]

The complexity of learning in the PAC-learning framework depends on the underlying concept class. The following two definitions introduce the measure of complexity of a concept class called the VC-dimension.

Definition 2.3. ***(Kearns and Vazirani [25])*** *Let $\mathcal{X}$ be an input space. Let $X = \{x_1, \ldots, x_m\} \subseteq \mathcal{X}$. For any concept class $\mathcal{C}$, let*

$$\Pi_{\mathcal{C}}(X) = \{(c(x_1), \ldots, c(x_m)) : c \in \mathcal{C}\}.$$

If $\Pi_{\mathcal{C}}(X) = \{0, 1\}^m$ then we say X is shattered *by $\mathcal{C}$.*

"Therefore, X is shattered by $\mathcal{C}$ if $\mathcal{C}$ realizes all possible dichotomies of X" [25].

Definition 2.4. ***(Vapnik and Chervonenkis [40])*** *Let $\mathcal{X}$ be an input space and $\mathcal{C}$ a concept class over $\mathcal{X}$. The* Vapnik-Chervonenkis dimension (VC-dimension*) of $\mathcal{C}$ is defined as follows. If arbitrarily large finite sets $X \subseteq \mathcal{X}$ can be shattered by $\mathcal{C}$, then the* VC *dimension of $\mathcal{C}$ is ∞. Otherwise the* VC*-dimension is the cardinality d of the largest set $X \subseteq \mathcal{X}$ that is shattered by $\mathcal{C}$.*

[1] In this document we use the terms "learning algorithm" and "learner" interchangeably.

[2] In the literature, sometimes the probability of *success* of the algorithm is considered instead of the probability of *failure*. This probability, which is called *confidence*, is at least $1 - \delta$.

[3] Often the concept class $\mathcal{C}$ is parameterized by a parameter n, that is $\mathcal{C} = \cup_{n \geq 1} \mathcal{C}_n$ and all concepts in $\mathcal{C}_n$ share a subdomain $\mathcal{X}_n$ and $\mathcal{X} = \cup_{n \geq 1} \mathcal{X}_n$. In such cases a polynomial dependence on n is also allowed [25]. In this thesis, this additional parameter is not relevant and hence omitted.

Blumer *et al.* [7] showed that the sufficient and necessary condition of learning in the PAC framework is determined by the VC-dimension of the concept class.

Theorem 2.5. ***(Blumer, Ehrenfeucht, Haussler and Warmuth [7])*** *A concept class $\mathcal{C}$ of* VC-*dimension d is* PAC-*learnable iff $d < \infty$.*

The proof of Theorem 2.5 is not presented in this thesis and the reader is referred to the article by Blumer *et al.* [7] for more details. However, a short description of the idea of the proof can be found in Section 5.1.

2.2 Noise in PAC-learning

In many real world applications examples are not usually noise-free. Informally, noise can be considered as any type of distortion in the examples returned to the learner by the oracle. In this section, we define a specific kind of noise, called label noise, in which only the label of the examples can be flipped.

Definition 2.6. *A* label noise model *over $\mathcal{X}$ is a mapping*

$$\Phi : 2^{\mathcal{X}} \times \mathcal{D}_{\mathcal{X}} \times \mathcal{X} \to 2^{[0,1]}. \tag{2.1}$$

A label noise model Φ over $\mathcal{X}$ is called deterministic *if $|\Phi(c, D, x)| = 1$ for all $c \in 2^{\mathcal{X}}$, $D \in \mathcal{D}_{\mathcal{X}}$, and $x \in \mathcal{X}$. Otherwise, Φ is called non-deterministic.*

Deterministic label noise models immediately induce probabilistic concepts in the following way.

Definition 2.7. *Let Φ be a deterministic label noise model over $\mathcal{X}$. For $D \in \mathcal{D}_{\mathcal{X}}$ and $c \in 2^{\mathcal{X}}$, the function $\Phi_{c,D} : \mathcal{X} \to [0, 1]$, determined by $\{\Phi_{c,D}(x)\} = \Phi(c, D, x)$ for all $x \in \mathcal{X}$, is called the* Φ-noisy concept *with respect to D and c. Then* $\mathrm{EX}_{\Phi}(c, D) = \mathrm{EX}(\Phi_{c,D}, D)$.

For convenience, we do not explicitly specify the input space $\mathcal{X}$ that the noise model is defined over when $\mathcal{X}$ is clear from the context. Also in this thesis, we only focus on label noise and not any other type of noise. The reader is referred to Sections 7.2.2 and 7.3 for more details on other types of noise.

A noisy concept $\Phi_{c,D}$ resulting from a deterministic label noise model Φ by fixing a distribution D and a concept c can be thought of as a probabilistic concept that results from c by applying a certain label noise process described by Φ. In particular, sampling according to c and D (through $\mathrm{EX}(c, D)$), followed by applying the label noise model, is defined as sampling according to the corresponding noisy concept $\Phi_{c,D}$ and D (through $\mathrm{EX}(\Phi_{c,D}, D)$).

For instance, if $\Phi(c, D, x) = 0.4$ then $\mathrm{EX}_{\Phi}(c, D)$ has a probability of 40% to label x with 1 if x is sampled. If the non-noisy concept c satisfies $c(x) = 1$, this corresponds to a probability of 60% of mislabeling the particular instance x. This probability of mislabeling of instance x is formally defined as follows.

Definition 2.8. *Let $\mathcal{X}$ be an input space, $\mathcal{C}$ a concept class, $D \in \mathcal{D}_{\mathcal{X}}$ a distribution and Φ a deterministic label noise model. For any $x \in \mathcal{X}$ and $c \in \mathcal{C}$, the* noise rate *$nr_{c,D}(x)$ is defined as follows.*

$$\mathrm{nr}_{c,D}(x) = |c(x) - \Phi_{c,D}(x)| \tag{2.2}$$

So we can think of the oracle $\mathrm{EX}_{\Phi}(c, D)$ as executing the following sequence of instructions:

1. Simulate $\mathrm{EX}(c, D)$ and let $(x, c(x))$ denote the resulting example.
2. With probability $\mathrm{nr}_{c,D}(x)$, let $y = 1 - c(x)$. Otherwise, let $y = c(x)$.
3. Return (x, y).

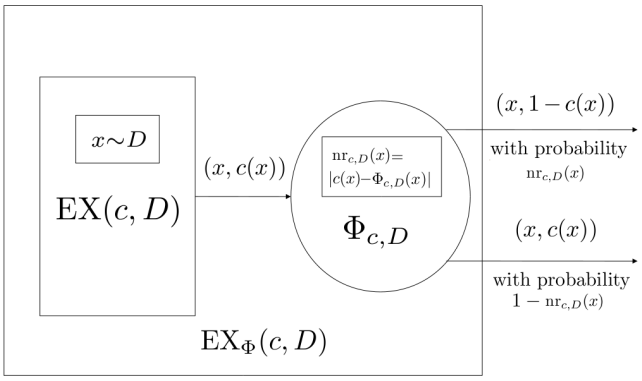

Figure 2.1: Deterministic label noise oracle

For illustration, see Figure 2.1.

If Φ is a non-deterministic label noise model then there exist D, c, and x such that $\Phi(c, D, x)$ contains more than one value in $[0, 1]$. This models a situation in which the noise model applied to the examples returned by $\mathrm{EX}(c, D)$ has a non-deterministically chosen noise rate. In such a case, we can think of the oracle $\mathrm{EX}_\Phi(c, D)$ as executing the following sequence of instructions:

1. Simulate $\mathrm{EX}(c, D)$ and let $(x, c(x))$ denote the resulting labeled example.
2. Non-deterministically, pick any value $p' \in \Phi(c, D, x)$.
3. With probability $p = |c(x) - p'|$, let $y = 1 - c(x)$. Otherwise, let $y = c(x)$.
4. Return (x, y).

For illustration, see Figure 2.2.

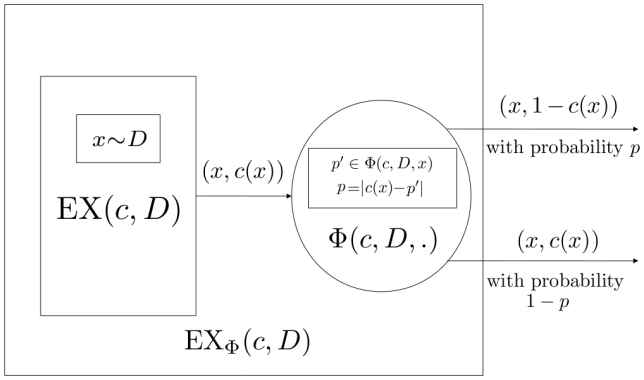

Figure 2.2: Non-deterministic label noise oracle

We use the definition of label noise models in order to model noise flipping the observed labels in examples presented to a learner that tries to identify an unknown target concept c^* in a given concept class $\mathcal{C}$ from a sample drawn $i.i.d.$ with respect to some unknown but fixed probability distribution D. Our corresponding definition of learnability is a variant of Valiant's original model of PAC learning [37], $cf.$ Definition 2.2.

Definition 2.9. *Let $\underline{\epsilon}, \underline{\delta} \in [0, \frac{1}{2})$. Let $\mathcal{C}$ be a concept class and $\mathcal{D} \subseteq \mathcal{D}_{\mathcal{X}}$ a class of distributions. Let Φ be a label noise model. $\mathcal{C}$ is* $\mathrm{PAC}_{\underline{\epsilon},\underline{\delta}}$-learnable *with respect to Φ and $\mathcal{D}$ if there exists a learning algorithm $\mathcal{L}$[4] and a function $m : (0, \frac{1}{2}) \times (0, \frac{1}{2}) \to \mathbb{N}$ such that for all $\epsilon \in (\underline{\epsilon}, \frac{1}{2})$, for all $\delta \in (\underline{\delta}, \frac{1}{2})$, for all $D \in \mathcal{D}$, and for all target concepts $c^* \in \mathcal{C}$ the following property is fulfilled:*

> *$\mathcal{L}$, given ϵ and δ, requests a sample $\mathcal{S}$ of $m(\epsilon, \delta)$ many independent draws from $\mathrm{EX}_{\Phi}(c^*, D)$, where $m(\epsilon, \delta)$ is polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$. Then $\mathcal{L}$, with probability of at least $1 - \delta$, returns a concept $c \in \mathcal{C}$ such that $Pr_{x \sim D}[c(x) \neq c^*(x)] \leq \epsilon$.*

The following definition determines under what condition two noisy oracles can potentially display the same stochastic behavior.

Definition 2.10. *Let Φ and Φ' be label noise models, c, c' two concepts, $D \in \mathcal{D}_{\mathcal{X}}$ a distribution and $x \in \mathcal{X}$. We say that* $\mathrm{EX}_{\Phi}(c, D)$ and $\mathrm{EX}_{\Phi'}(c', D)$ have potentially equivalent behavior on x *if $\Phi(c, D, x) \cap \Phi'(c', D, x) \neq \emptyset$.* $\mathrm{EX}_{\Phi}(c, D)$ *and* $\mathrm{EX}_{\Phi'}(c', D)$ *have potentially equivalent behavior on D if they have potentially equivalent behavior on all $x \in \mathrm{supp}(D)$.*

The following two lemmas state the learnability result for the case that two noisy oracles can potentially produce the same set of examples. These lemmas will be used frequently in Chapters 3 and 5.

Lemma 2.11. *Let $\underline{\epsilon}, \underline{\delta} \in [0, \frac{1}{2})$. Let $\mathcal{C}$ be a concept class. Let $\mathcal{D} \subseteq \mathcal{D}_{\mathcal{X}}$ be a class of distributions. Let Φ be a deterministic label noise model and Φ' be a non-deterministic label noise model such that for any $c \in \mathcal{C}$ and any $D \in \mathcal{D}$,* $\mathrm{EX}_{\Phi}(c, D)$ *and* $\mathrm{EX}_{\Phi'}(c, D)$ *have potentially equivalent behavior on D. Let $\mathcal{L}$ be a learning algorithm that* $\mathrm{PAC}_{\underline{\epsilon},\underline{\delta}}$*-learns $\mathcal{C}$ with respect to Φ' and $\mathcal{D}$. Then $\mathcal{L}$* $\mathrm{PAC}_{\underline{\epsilon},\underline{\delta}}$*-learns $\mathcal{C}$ with respect to Φ and $\mathcal{D}$.*

Proof. Let $D \in \mathcal{D}$ be a distribution. Let $x \in \mathcal{X}$ be any instance drawn with respect to D and $c^* \in \mathcal{C}$ the target concept. $\mathrm{EX}_{\Phi'}(c^*, D)$ can always choose $p' = \Phi(c^*, D, x)$ in its non-deterministic step when x is drawn with respect to D.

Therefore, $\mathrm{EX}_{\Phi'}(c^*, D)$ can label any sample $\mathcal{S}$ the same way as $\mathrm{EX}_{\Phi}(c^*, D)$. Since $\mathcal{L}$ can $\mathrm{PAC}_{\underline{\epsilon},\underline{\delta}}$-learn $\mathcal{C}$ with any sample $\mathcal{S}$ drawn from $\mathrm{EX}_{\Phi'}(c^*, D)$ (given a sufficient sample size), then $\mathcal{L}$ can also $\mathrm{PAC}_{\underline{\epsilon},\underline{\delta}}$-learn $\mathcal{C}$ with the specific kind of sample just described. Therefore, $\mathcal{L}$ can also $\mathrm{PAC}_{\underline{\epsilon},\underline{\delta}}$-learn $\mathcal{C}$ with respect to Φ and $\mathcal{D}$. □

Lemma 2.12. *Let Φ be a label noise model. Let $\mathcal{C}$ be a concept class, $c, c' \in \mathcal{C}$ with $c \neq c'$ and $\mathcal{D} \subseteq \mathcal{D}_{\mathcal{X}}$. Let $\underline{\epsilon} < \frac{1}{2} Pr_{x \sim D}[c(x) \neq c'(x)]$. If there is some $D \in \mathcal{D}$ such that* $\mathrm{EX}_{\Phi}(c, D)$ *and* $\mathrm{EX}_{\Phi}(c', D)$ *have potentially equivalent behavior on D, then $\mathcal{C}$ is not* $\mathrm{PAC}_{\underline{\epsilon},\underline{\delta}}$*-learnable with respect to Φ and $\mathcal{D}$ for any $\underline{\delta} \in (0, \frac{1}{2})$.*

Proof. Let $Pr_{x \sim D}[c(x) \neq c'(x)] = \epsilon$. Since $\mathrm{EX}_{\Phi}(c, D)$ and $\mathrm{EX}_{\Phi}(c', D)$ have potentially equivalent behavior, these two oracles can label any set of instances drawn with respect to D in the same way. Therefore, no learning algorithm $\mathcal{L}$ can distinguish between the examples drawn from $\mathrm{EX}_{\Phi}(c, D)$ and $\mathrm{EX}_{\Phi}(c', D)$. Since we only consider deterministic learning algorithms in this thesis, in the best case, $\mathcal{L}$ will have an error of $\frac{\epsilon}{2}$. This will happen, if $\mathcal{L}$ returns a concept $c'' \in \mathcal{C}$ (if such a concept exists) such that $Pr_{x \sim D}[c(x) \neq c''(x)] = Pr_{x \sim D}[c'(x) \neq c''(x)] = \frac{\epsilon}{2}$ because otherwise the error of $\mathcal{L}$ is at least greater than $\frac{\epsilon}{2}$ for one of the cases that c or c' is the target concept. Therefore, no learning algorithm $\mathcal{L}$ can $\mathrm{PAC}_{\underline{\epsilon},\underline{\delta}}$-learn $\mathcal{C}$ with respect to Φ and $\mathcal{D}$ for any $\underline{\epsilon} < \frac{\epsilon}{2}$. □

2.3 Classic examples of label noise models

In this section we introduce some of the classic label noise models in the literature, reformulated in our generic label noise framework introduced in Definition 2.6. We consider the model of random

[4] In this thesis we just consider deterministic learning algorithms, not randomized ones.

classification noise, as defined by Angluin and Laird [2], its generalization called constant-partition classification noise, introduced by Decatur [14] and studied by Ralaivola, Denis, and Magnan [29], and the model of malicious misclassification noise, which is due to Sloan [32].

We start with the most benign type of label noise introduced by Angluin and Laird.

Definition 2.13. ***(Angluin and Laird [2])*** *Let $\eta \in [0, 1)$. The* η-random classification noise model *is a label noise model $\Phi_{\mathrm{rcn}(\eta)}$, defined by*

$$\Phi_{\mathrm{rcn}(\eta)}(c, D, x) = \begin{cases} \{1 - \eta\}, & \text{if } c(x) = 1, \\ \{\eta\}, & \text{if } c(x) = 0, \end{cases}$$

where $c \in 2^{\mathcal{X}}$, $D \in \mathcal{D}_{\mathcal{X}}$, and $x \in \mathcal{X}$.

The random classification noise model has also been called the *classification noise model* [2] and the *random misclassification noise model* [32] in the literature.

For every example (x, y) that a learning algorithm $\mathcal{L}$ draws from the oracle $\mathrm{EX}_{\Phi_{\mathrm{rcn}(\eta)}}(c, D)$, $y = c(x)$ will hold with probability $1 - \eta$. With probability η, the label will be flipped, *i.e.*, $y = 1 - c(x)$.

In particular, the model of random classification noise does not depend on the distribution D. Moreover, the value of a noisy concept in an instance $x \in \mathcal{X}$ does not depend on values of the underlying target concept c other than the value $c(x)$ itself. Furthermore, the model of random classification noise is a deterministic label noise model.

The first learnability result with respect to random classification noise is due to Angluin and Laird.

Theorem 2.14. ***(Angluin and Laird [2])*** *Let $\mathcal{C}$ be a finite concept class. Let $\eta \in [0, \frac{1}{2})$. Then $\mathcal{C}$ is $\mathrm{PAC}_{0,0}$-learnable with respect to η-random classification noise and $\mathcal{D}_{\mathcal{X}}$.*

The proof of Theorem 2.14 can be found in Section 5.1. This proof gives us insights for many of the proofs and arguments in the rest of this thesis.

Angluin and Laird [2] have a more general learning algorithm than the one introduced in the proof of Theorem 2.14. Their algorithm can $\mathrm{PAC}_{0,0}$-learn any finite concept class under random classification noise knowing only an upper bound η_b on η such that $\eta \leq \eta_b < \frac{1}{2}$ instead of knowing η itself. They also have an algorithm for estimating η_b from the sample. They show that estimating this upper bound increases the required sample size only slightly. The reader is referred to the article by Angluin and Laird [2] for more details on the algorithm and its correctness.

Laird [27] proved a stronger version of Theorem 2.14, for concept classes of finite VC-dimension instead of finite concept classes.[5]

Theorem 2.15. ***(Laird [27])*** *Let $\mathcal{C}$ be a concept class of* VC*-dimension $d < \infty$. Let $\eta \in [0, \frac{1}{2})$. Then $\mathcal{C}$ is $\mathrm{PAC}_{0,0}$-learnable with respect to η-random classification noise and $\mathcal{D}_{\mathcal{X}}$.*

The proof of Theorem 2.15 can be found in Section 5.1.

Next, we consider a generalization of the random classification noise model, introduced by Decatur.

Definition 2.16. ***(Decatur [14])*** *Let $k \in \mathbb{N}$ and $\eta = (\eta_1, \ldots, \eta_k) \in [0, 1)^k$. Let $\pi = (\pi_1, \ldots, \pi_k) \subseteq (\mathcal{X} \times \{0, 1\})^k$ be a k-tuple of pairwise disjoint sets such that $\pi_1 \cup \cdots \cup \pi_k = \mathcal{X} \times \{0, 1\}$. The* (η, π)-constant-partition classification noise (CPCN) model *is a label noise model $\Phi_{\mathrm{cpcn}(\eta,\pi)}$, defined by*

$$\Phi_{\mathrm{cpcn}(\eta,\pi)}(c, D, x) = \begin{cases} \{1 - \eta_i\}, & \text{if } c(x) = 1, \\ \{\eta_i\}, & \text{if } c(x) = 0, \end{cases}$$

where $c \in 2^{\mathcal{X}}$, $D \in \mathcal{D}_{\mathcal{X}}$, $x \in \mathcal{X}$, and $i \in \{1, \ldots, k\}$ is such that $(x, c(x)) \in \pi_i$.

[5] Finite concept classes also have finite VC-dimension but concept classes of finite VC-dimension may have an infinite number of concepts.

For every example (x, y) that a learning algorithm $\mathcal{L}$ draws from the oracle $\mathrm{EX}_{\Phi_{\mathrm{cpcn}(\eta,\pi)}}(c, D)$, the probability with which the label y disagrees with $c(x)$ is determined by which partition π_i of $\mathcal{X} \times \{0, 1\}$ the example $(x, c(x))$ belongs to.

Other than that, the CPCN model behaves like the random classification noise model, $i.e.$, it is deterministic, independent of the distribution D, and produces noisy concepts whose value for an instance x does not depend on values of the underlying target concept c other than the value $c(x)$ itself.

Later, Ralaivola, Denis and Magnan proved that CPCN is equivalent to random classification noise as far as learnability is concerned [29].

Theorem 2.17. ***(Ralaivola, Denis, Magnan [29])*** *Let $\mathcal{C}$ be a concept class. Let $k \in \mathbb{N}$, $\eta = (\eta_1, \ldots, \eta_k) \in [0, \frac{1}{2})^k$ and $\eta' \in [0, \frac{1}{2})$. Let $\pi = (\pi_1, \ldots, \pi_k) \subseteq (\mathcal{X} \times \{0, 1\})^k$ be pairwise disjoint sets such that $\pi_1 \cup \cdots \cup \pi_k = \mathcal{X} \times \{0, 1\}$. Then the following statements are equivalent.*

1. *$\mathcal{C}$ is* PAC*-learnable.*
2. *$\mathcal{C}$ is $\mathrm{PAC}_{0,0}$-learnable with respect to $\Phi_{\mathrm{rcn}(\eta')}$ and $\mathcal{D}_{\mathcal{X}}$.*
3. *$\mathcal{C}$ is $\mathrm{PAC}_{0,0}$-learnable with respect to $\Phi_{\mathrm{cpcn}(\eta,\pi)}$ and $\mathcal{D}_{\mathcal{X}}$.*
4. *$\mathcal{C}$ has finite* VC*-dimension.*

The reader is referred to the article by Ralaivola *et al.* [29] for the proof.

The last model of label noise discussed in this section is a weaker version of random classification noise that is due to Sloan.

Definition 2.18. ***(Sloan [32])*** *Let $\eta \in [0, 1)$. The* η-malicious classification noise *model is a label noise model $\Phi_{\mathrm{mcn}(\eta)}$, defined by*

$$\Phi_{\mathrm{mcn}(\eta)}(c, D, x) = \begin{cases} [1 - \eta, 1], & \text{if } c(x) = 1, \\ [0, \eta], & \text{if } c(x) = 0, \end{cases}$$

where $c \in 2^{\mathcal{X}}$, $D \in \mathcal{D}_{\mathcal{X}}$, and $x \in \mathcal{X}$.

Malicious classification noise has also been called *malicious misclassification noise [32]* in the literature.

For every example (x, y) that a learning algorithm $\mathcal{L}$ draws from the oracle $\mathrm{EX}_{\Phi_{\mathrm{mcn}(\eta)}}(c, D)$, $y = c(x)$ will hold with probability at least $1 - \eta$. Beyond this probability of at least $1 - \eta$, we cannot make any assumption at all about the label y.

In particular, as is the case for random classification noise, the label noise model in malicious classification noise does not depend on the distribution D. Also, the value of a noisy concept for an instance $x \in \mathcal{X}$ does not depend on values of the underlying target concept c other than the value $c(x)$ itself. In contrast to random classification noise, the model of malicious classification noise is a non-deterministic label noise model.

Sloan proved the first learnability result with respect to malicious classification noise.

Theorem 2.19. ***(Sloan [31, 32])*** *Let $\mathcal{C}$ be a finite concept class. Let $\eta \in [0, \frac{1}{2})$. Then $\mathcal{C}$ is $\mathrm{PAC}_{0,0}$-learnable with respect to $\Phi_{\mathrm{mcn}(\eta)}$ and $\mathcal{D}_{\mathcal{X}}$.*

Sloan proved this theorem with an approach similar to that used by Angluin and Laird [2] for the random classification noise model. He argued that the η-malicious noise oracle will be less harmful than an η-random classification process, because it flips the label of at most as many examples as are flipped by the random classification noise. The reader is referred to the article by Sloan [31] for more details.

Similar to the case of random classification noise, the learnability results for finite concept classes can be generalized to the classes of finite VC-dimension.

Theorem 2.20. *Let $\mathcal{C}$ be a concept class of* VC*-dimension $d < \infty$. Let $\eta \in [0, \frac{1}{2})$. Then $\mathcal{C}$ is* $\mathrm{PAC}_{0,0}$*-learnable with respect to $\Phi_{\mathrm{mcn}(\eta)}$ and $\mathcal{D}_{\mathcal{X}}$.*

It should be mentioned that this theorem has not been proved in the literature before. But it is not hard to show that the idea of the proof of Theorem 2.19 together with the generalization technique in converting the proof of Theorem 2.14 to the proof of Theorem 2.15 can be combined here as well to prove Theorem 2.20. The proof is omitted due to this similarity.

Some other classic noise models, such as noise models generated by the malicious error oracle [23, 38] and the random attribute error oracle [32] cannot be modeled with Definition 2.6, since they are not purely label noise models. In these models, the instance $x \in \mathcal{X}$ can also be distorted. Such models are beyond the scope of this thesis. Therefore, for the remainder of this thesis we simply use the terms "noise" and "label noise" interchangebly unless explicitly stated otherwise.

Chapter 3

Locally Variable Noise

One of our main criticisms of the classic noise models introduced in Section 2.3 is that most of them assume that the noise rate is independent of the instances, distribution and the target concept. However, in many real world applications the noise rate depends on at least one of these parameters [8]. For example, instances may be less likely to be mislabeled if all instances in their local neighbourhood have the same label, as opposed to the case when the target concept labels half of the points in the neighbourhood with label 0 and the other half with label 1 (see Figure 3.1).

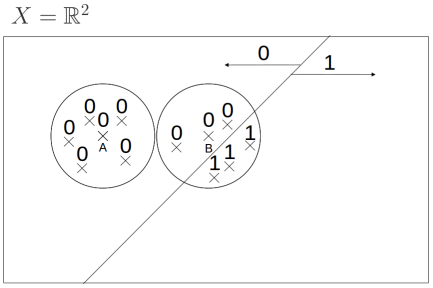

Figure 3.1: Distance Ball Noise Model

In this chapter, we introduce four new label noise models in which the noise rate for any instance depends on the instance itself, on the underlying target concept and (in all but one model) also on the underlying distribution. These noise models are a step from conventional noise models towards noise models that appear in real world applications. Also, these noise models are local in the sense that they yield noisy concepts that assign a label to a point $x \in \mathcal{X}$ depending on the labels of other points in the proximity of x. The measure of proximity can be either a distance metric or the probability mass. The next two sections introduce these noise models formally based on these two measures of proximity.

3.1 Distance ball noise models

For our first noise model, we use the underlying distance metric, dist, to define the measure of proximity.

Definition 3.1. *For any* radius $\rho \geq 0$ *and any instance* $x \in X$*, the* ρ-distance ball $\mathrm{DB}_\rho(x)$ *around* x *is defined by*

$$\mathrm{DB}_\rho(x) = \{x' \in \mathcal{X} \mid \mathrm{dist}(x, x') \leq \rho\},$$

where dist *is an arbitrary metric on* $\mathcal{X}$ *that has been fixed a priori.*

Then, we introduce our first label noise model, called the ρ-distance malicious classification noise model[1]. This model goes back to a suggestion made by Shai Ben-David (personal correspondence with Sandra Zilles).

Definition 3.2. *Let $\rho \geq 0$. The* ρ-distance malicious classification noise model $\Phi_{\mathrm{dmball}(\rho)}$ *is a label noise model defined as*

$$\Phi_{\mathrm{dmball}(\rho)}(c, D, x) = \begin{cases} \{c(x)\}, & \text{if } c(x) = c(x') \text{ for all } x' \in \mathrm{DB}_\rho(x), \\ [0,1], & \text{otherwise}, \end{cases}$$

where $c \in 2^{\mathcal{X}}$, $D \in \mathcal{D}_{\mathcal{X}}$, and $x \in \mathcal{X}$.

For every example (x, y) that a learning algorithm $\mathcal{L}$ draws from the oracle $\mathrm{EX}_{\Phi_{\mathrm{dmball}(\rho)}}(c, D)$, $y = c(x)$ will be guaranteed if all points in the ρ-distance ball around x have the same label under c (see Point A in Figure 3.1). If both positively and negatively labeled points lie in the ρ-distance ball around x, we cannot make any assumption at all about the label y (see Point B in Figure 3.1).

The label noise model here does not depend on the distribution D. However, unlike for the classic models considered in Section 2.3, the value of a noisy concept for an instance $x \in \mathcal{X}$ depends on some of the values of the underlying target concept c other than the value $c(x)$ itself.

Like malicious classification noise, the model of distance malicious classification noise is in general a non-deterministic label noise model.

Distance malicious classification noise is a strong adversarial noise model. Even simple concept classes may not be PAC-learnable with respect to such a noise model. One such simple concept class is the class of one-dimensional threshold functions defined as follows.

Definition 3.3. *Let $\mathcal{X} \subseteq \mathbb{R}$ and $\mathcal{C}^1_{\mathrm{thr}} = \{c_\theta \mid \theta \in \mathbb{R}\}$, where*

$$c_\theta(x) = \begin{cases} 1, & \text{if } x \geq \theta, \\ 0, & \text{if } x < \theta. \end{cases}$$

$\mathcal{C}^1_{\mathrm{thr}}$ *is called the class of* one-dimensional threshold functions.

We show that the class of one-dimensional threshold functions is not $\mathrm{PAC}_{\underline{\epsilon},\underline{\delta}}$-learnable with respect to distance malicious classification noise and $\mathcal{D}_{\mathcal{X}}$ for any arbitrary $\underline{\epsilon}, \underline{\delta} \in [0, \frac{1}{2})$.

Proposition 3.4. *Let $\rho > 0$ and $\underline{\epsilon}, \underline{\delta} \in [0, \frac{1}{2})$. Let $\mathcal{X} = \mathbb{R}$. $\mathcal{C}^1_{\mathrm{thr}}$ is not $\mathrm{PAC}_{\underline{\epsilon},\underline{\delta}}$-learnable with respect to $\Phi_{\mathrm{dmball}(\rho)}$ and $\mathcal{D}_{\mathcal{X}}$.*

Proof. Let $x_1, x_2 \in \mathcal{X}$ such that $\mathrm{dist}(x_1, x_2) \leq \frac{\rho}{2}$ and $x_1 < x_2$. The distance between x_1 and x_2 guarantees that the ρ-distance ball around either of x_1 or x_2 contains the other point. Let D be such that $Pr_{x\sim D}[x = x_1] = \underline{\epsilon} + \frac{1-\underline{\epsilon}}{2}$, $Pr_{x\sim D}[x = x_2] = \frac{1-\underline{\epsilon}}{2}$ and $Pr_{x\sim D}[x \notin \{x_1, x_2\}] = 0$. Therefore, $\mathrm{supp}(D) = \{x_1, x_2\}$.

Let $\mathcal{C} = \{c, c'\}$ where $c(x) = c_{x_1-\rho}(x)$ and $c'(x) = c_{x_2}(x)$ for all $x \in \mathcal{X}$ (as illustrated in Figure 3.2). We show that $\mathrm{EX}_{\Phi_{\mathrm{dmball}(\rho)}}(c, D)$ and $\mathrm{EX}_{\Phi_{\mathrm{dmball}(\rho)}}(c', D)$ have potentially equivalent behavior on D. Therefore, we can use Lemma 2.12 to show that $\mathcal{C}^1_{\mathrm{thr}}$ is not $\mathrm{PAC}_{\underline{\epsilon},\underline{\delta}}$-learnable with respect to $\Phi_{\mathrm{dmball}(\rho)}$ and $\mathcal{D}_{\mathcal{X}}$.

For any $x \in \mathrm{supp}(D)$, $\Phi_{\mathrm{dmball}(\rho)}(c, D, x) = \{1\}$ because the ρ-distance ball around x only contains points with label 1. Also, for any $x \in \mathrm{supp}(D)$, $\Phi_{\mathrm{dmball}(\rho)}(c', D, x) = [0, 1]$ because the ρ-distance ball around x contains points with both labels 0 and 1. Thus, $\Phi_{\mathrm{dmball}(\rho)}(c, D, x) \cap \Phi_{\mathrm{dmball}(\rho)}(c', D, x) \neq \emptyset$ for all $x \in \mathrm{supp}(D)$. Therefore, $\mathrm{EX}_{\mathrm{dmball}(\rho)}(c, D)$ and $\mathrm{EX}_{\mathrm{dmball}(\rho)}(c', D)$ have potentially equivalent behavior on D based on Definition 2.10.

Since $Pr_{x\sim D}[c(x) \neq c'(x)] > \underline{\epsilon}$, therefore, $\mathcal{C}^1_{\mathrm{thr}}$ is not $\mathrm{PAC}_{\underline{\epsilon},\underline{\delta}}$-learnable with respect to $\Phi_{\mathrm{dmball}(\rho)}$ and $\mathcal{D}_{\mathcal{X}}$. □

[1] We simply write distance malicious classification noise instead of ρ-distance malicious classification noise when ρ is clear from the context.

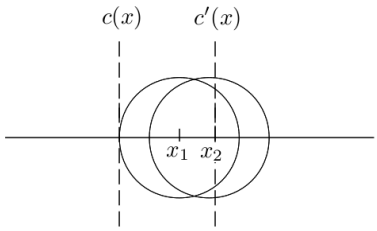

Figure 3.2: $\mathcal{C}^1_{\text{thr}}$ is not PAC-learnable with respect to $\Phi_{\text{dmball}(\rho)}$ and $\mathcal{D}_{\mathcal{X}}$.

It seems like choosing an arbitrary label from the distance ball around a point as the label of that point gives too much freedom to the oracle. We mitigate this a bit in our second noise model, called the ρ-distance random classification noise model[2]. This model is a deterministic version of ρ-distance malicious classification noise and was also inspired by a suggestion by Shai Ben-David.

Definition 3.5. *Let $\rho \geq 0$. The* ρ-distance random classification noise model $\Phi_{\text{drball}(\rho)}$ *is a label noise model defined as*

$$\Phi_{\text{drball}(\rho)}(c, D, x) = \{Pr_{x' \sim D}[c(x') = 1 \mid x' \in \text{DB}_\rho(x)]\},$$

where $c \in 2^{\mathcal{X}}$, $D \in \mathcal{D}_{\mathcal{X}}$, and $x \in \text{supp}(D)$. $\Phi_{\text{drball}(\rho)}(c, D, x) = \{0\}$ for $x \notin \text{supp}(D)$.

For every example (x, y) that a learning algorithm $\mathcal{L}$ draws from the oracle $\text{EX}_{\Phi_{\text{drball}(\rho)}}(c, D)$, $y = c(x)$ will be guaranteed if all points in the ρ-distance ball around x have the same label under c (see Point A in Figure 3.1). If both positively and negatively classified points lie in the ρ-distance ball around x, then the label y will be drawn from $\{0, 1\}$ according to the distribution of labels within the ρ-distance ball around x (see Point B in Figure 3.1).

This label noise model depends on the distribution D. Moreover, the value of a noisy concept for an instance $x \in \mathcal{X}$ depends on some of the values of the underlying target concept c other than the value $c(x)$ itself.

One simple observation about the relation of these two distance noise models is as follows. Any concept class that is PAC-learnable with respect to distance malicious classification noise and a class of distributions is also PAC-learnable with respect to distance random classification noise and the same class of distributions.

Proposition 3.6. *Let $\rho \geq 0$ and $\underline{\epsilon}, \underline{\delta} \in [0, \frac{1}{2})$. Let $\mathcal{D} \subseteq \mathcal{D}_{\mathcal{X}}$ be a class of distributions. Any concept class $\mathcal{C}$ that is* $\text{PAC}_{\underline{\epsilon},\underline{\delta}}$*-learnable with respect to $\Phi_{\text{dmball}(\rho)}$ and $\mathcal{D}$ is also* $\text{PAC}_{\underline{\epsilon},\underline{\delta}}$*-learnable with respect to $\Phi_{\text{drball}(\rho)}$ and $\mathcal{D}$.*

Proof. Let $\mathcal{L}$ be a learning algorithm that $\text{PAC}_{\underline{\epsilon},\underline{\delta}}$-learns $\mathcal{C}$ with respect to $\Phi_{\text{dmball}(\rho)}$ and $\mathcal{D}$. Let $D \in \mathcal{D}$ be a distribution and $c^* \in \mathcal{C}$ the target concept.

In the rest of this proof, we show that the two oracles $\text{EX}_{\Phi_{\text{dmball}(\rho)}}(c^*, D)$ and $\text{EX}_{\Phi_{\text{drball}(\rho)}}(c^*, D)$ have potentially equivalent behavior on D. Since $\Phi_{\text{dmball}(\rho)}$ is a non-deterministic label noise model and $\Phi_{\text{drball}(\rho)}$ is a deterministic one, we can apply Lemma 2.11 to show that $\mathcal{L}$ $\text{PAC}_{\underline{\epsilon},\underline{\delta}}$-learns $\mathcal{C}$ with respect to $\Phi_{\text{drball}(\rho)}$ and $\mathcal{D}$.

Let us analyze a possible behavior of $\text{EX}_{\Phi_{\text{dmball}(\rho)}}$ according to the following cases.

1. $c^*(x) = c^*(x')$ for all $x' \in \text{DB}_\rho(x)$. In this case $\Phi_{\text{dmball}(\rho)}(c^*, D, x) = \Phi_{\text{drball}(\rho)}(c^*, D, x) = \{c^*(x)\}$ based on Definitions 3.2 and 3.5 respectively. Therefore, $\text{EX}_{\Phi_{\text{drball}(\rho)}}(c^*, D)$ and $\text{EX}_{\Phi_{\text{dmball}(\rho)}}(c^*, D)$ have potentially equivalent behavior on all such x.

[2]We simply write distance random classification noise model instead of ρ-distance random classification noise model when ρ is clear from the context.

2. Otherwise, $\Phi_{\mathrm{dmball}(\rho)}(c^*, D, x) = [0, 1]$, in particular $\Phi_{\mathrm{drball}(\rho)}(c^*, D, x)$ is contained in $\Phi_{\mathrm{dmball}(\rho)}(c^*, D, x)$. Therefore, $\mathrm{EX}_{\Phi_{\mathrm{dmball}(\rho)}}(c^*, D)$ and $\mathrm{EX}_{\Phi_{\mathrm{drball}(\rho)}}(c^*, D)$ also have potentially equivalent behavior on all x that are not satisfying the condition of case 1.

And, hence, $\mathrm{EX}_{\Phi_{\mathrm{dmball}(\rho)}}(c^*, D)$ and $\mathrm{EX}_{\Phi_{\mathrm{drball}(\rho)}}(c^*, D)$ have potentially equivalent behavior on D.

Note that since the number of examples drawn by the learning algorithm does not play any role in our argument, the polynomial sample bounds needed in the definition of PAC-learnability are preserved. □

Unlike the case of distance malicious classification noise, the class of one-dimensional threshold functions is $\mathrm{PAC}_{0,0}$-learnable with respect to distance random classification noise and $\mathcal{D}_{\mathcal{X}}$. To prove this we need some preliminaries. We first show that the noisy concept resulting from applying distance random classification noise to any concept in the class of one-dimensional threshold functions is a non-decreasing function.

Lemma 3.7. *Let $\rho \geq 0$. Let $\mathcal{X} \subseteq \mathbb{R}$. Let $c^* \in \mathcal{C}^1_{\mathrm{thr}}$ be the target concept and $D \in \mathcal{D}_{\mathcal{X}}$ a distribution. $\Phi_{\mathrm{drball}(\rho)_{c^*,D}}(x)$ is a non-decreasing function with respect to $x \in \mathcal{X}$.*

Proof. We show that $\Phi_{\mathrm{drball}(\rho)_{c^*,D}}(x)$ never decreases when x is increasing. Let $\theta^* \in \mathbb{R}$ such that $x \geq \theta^*$ implies $c^*(x) = 1$ and $x < \theta^*$ implies $c^*(x) = 0$.

Let $x_1, x_2 \in \mathcal{X}$ such that $x_1 \leq x_2$. Based on the positions of x_1 and x_2 in $\mathbb{R}$ the following three cases can happen:

1. $x_1 < \theta^* - \rho$: based on Definition 3.5, $\Phi_{\mathrm{drball}(\rho)_{c^*,D}}(x_1) = c^*(x_1) = 0$. Since Φ is always in $[0, 1]$, $\Phi_{\mathrm{drball}(\rho)_{c^*,D}}(x_1) \leq \Phi_{\mathrm{drball}(\rho)_{c^*,D}}(x_2)$.

2. $x_2 > \theta^* + \rho$: based on Definition 3.5, $\Phi_{\mathrm{drball}(\rho)_{c^*,D}}(x_2) = c^*(x_2) = 1$. Since Φ is always in $[0, 1]$, $\Phi_{\mathrm{drball}(\rho)_{c^*,D}}(x_1) \leq \Phi_{\mathrm{drball}(\rho)_{c^*,D}}(x_2)$.

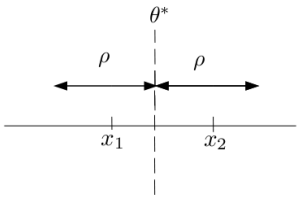

Figure 3.3: When $x_1 \geq \theta^* - \rho$ and $x_2 \leq \theta^* + \rho$

3. $x_1 \geq \theta^* - \rho$ and $x_2 \leq \theta^* + \rho$: in this case (as illustrated in Figure 3.3),

$$Pr_{x\sim D}[\theta^* \leq x \leq x_1 + \rho] \leq Pr_{x\sim D}[\theta^* \leq x \leq x_2 + \rho], \tag{3.1}$$

because $Pr_{x\sim D}[\theta^* \leq x \leq x_2 + \rho] = Pr_{x\sim D}[\theta^* \leq x \leq x_1 + \rho] + Pr_{x\sim D}[x_1 + \rho \leq x \leq x_2 + \rho]$. Similarly,

$$Pr_{x\sim D}[x_1 - \rho \leq x < \theta^*] \geq Pr_{x\sim D}[x_2 - \rho \leq x < \theta^*]. \tag{3.2}$$

Therefore, using Equations 3.1 and 3.2 respectively,

$$\begin{aligned}
\Phi_{\mathrm{drball}(\rho)_{c^*,D}}(x_1) &= \frac{Pr_{x\sim D}[\theta^* \leq x \leq x_1 + \rho]}{Pr_{x\sim D}[x_1 - \rho \leq x < \theta^*] + Pr_{x\sim D}[\theta^* \leq x \leq x_1 + \rho]} \\
&\leq \frac{Pr_{x\sim D}[\theta^* \leq x \leq x_2 + \rho]}{Pr_{x\sim D}[x_1 - \rho \leq x < \theta^*] + Pr_{x\sim D}[\theta^* \leq x \leq x_2 + \rho]} \\
&\leq \frac{Pr_{x\sim D}[\theta^* \leq x \leq x_2 + \rho]}{Pr_{x\sim D}[x_2 - \rho \leq x < \theta^*] + Pr_{x\sim D}[\theta^* \leq x \leq x_2 + \rho]} \\
&= \Phi_{\mathrm{drball}(\rho)_{c^*,D}}(x_2)
\end{aligned}$$

□

Next, we present a theorem by Kearns and Schapire [24] and then we describe the proof sketch. The idea that they used in their proof can be applied in our proof of $\mathrm{PAC}_{0,0}$-learnability of a class of one-dimensional thresholds on the line with respect to distance random classification noise and $\mathcal{D}_{\mathcal{X}}$.

As we mentioned in Chapter 2, any set of functions $f : \mathcal{X} \to [0,1]$ can be considered as a concept class. Therefore, an oracle can be defined when any of the functions in the set is selected as the target concept in the same way as the oracle was defined when a probabilistic concept is chosen as the target concept from a concept class. The following theorem shows that for a specific type of such a set of functions, an algorithm exists that, with high probability, returns good estimates of the value of the target probabilistic concept for a given fraction of probability mass on $\mathcal{X}$. Note that the learning framework is different here than the learning framework of PAC-learning with respect to noise (Definition 2.9) because the learner returns a function of the form $f : \mathcal{X} \to [0,1]$ instead of a function of the form $c : \mathcal{X} \to \{0,1\}$.

Theorem 3.8. ***(Kearns and Schapire [24])*** *Let $\epsilon, \delta, \gamma \in (0, \frac{1}{2})$. Let $\mathcal{X} = \mathbb{R}$ and F the set of all non-decreasing functions $f : \mathcal{X} \to [0,1]$. There exists an algorithm L that, for any target function $f^* \in F$ and $D \in \mathcal{D}_{\mathcal{X}}$, using*

$$m = \lceil \frac{4}{\epsilon\gamma} \rceil \lceil max(\frac{64 \ln(\frac{2^{21}}{(\epsilon\gamma)^2\delta})}{\epsilon\gamma}, \frac{2\ln(\frac{4\lceil \frac{4}{\epsilon\gamma} \rceil}{\delta})}{\gamma^2}) \rceil$$

examples drawn with respect to $\mathrm{EX}(f^*, D)$, will return a function $\hat{f} : \mathcal{X} \to [0,1]$ with the property that

$$Pr_{x\sim D}[|\hat{f}(x) - f^*(x)| > \gamma] \leq \epsilon$$

with probability of at least $1 - \delta$.

The proof sketch below, except for slight changes in wording and notation, is the same as in the article by Kearns and Schapire [24].

Sketch of the Proof. (Kearns and Schapire [24]) Let

$$s_1 = \lceil \frac{4}{\epsilon\gamma} \rceil,$$

and

$$s_2 = \lceil \max(\frac{64 \ln(\frac{2^{21}}{(\epsilon\gamma)^2\delta})}{\epsilon\gamma}, \frac{2\ln(\frac{4\lceil \frac{4}{\epsilon\gamma} \rceil}{\delta})}{\gamma^2}) \rceil.$$

L draws a sample of $m = s_1 s_2$ examples (x_i, y_i) where $x_i \in \mathcal{X}$ and $y_i \in \{0,1\}$ for $1 \leq i \leq m$. Let the examples be sorted by their instances such that $x_1 \leq \ldots \leq x_m$. First, let us assume $x_1 < \ldots < x_m$. Later we show how to remove this assumption.

$\mathcal{X}$ can be partitioned into s_1 disjoint intervals, each interval I_j, $1 \leq j \leq s_1$, containing exactly s_2 instances of $\mathcal{S}$ *i.e.*, $I_1 = (-\infty, x_{s_2}]$, $I_j = (x_{(j-1)s_2}, x_{js_2}]$ for $2 \leq j \leq s_1 - 1$ and $I_{s_1} = (x_{(t-1)s_2}, +\infty)$. Let $\phi_j = \frac{1}{s_2} \sum_{i=(j-1)s_2+1}^{js_2} y_i$. L will return a step function $\hat{f}$ such that for any $x \in I_j$, $f(x) = \phi_j$.

The high level idea of the proof can be described as follows. We show that for $1 \leq j \leq s_1$, $Pr_{x\sim D}[x \in I_j]$ is at most $\frac{\epsilon\gamma}{2}$ with high probability. Next we show that if f^* increases by at most $\frac{\gamma}{2}$ on any interval I_j , then $\hat{f}$ is close to f^* on all points $x \in I_j$. On the other hand, since f^* is non-decreasing and can only take values between 0 and 1, f^* can increase by more than $\frac{\gamma}{2}$ in at most $\frac{2}{\gamma}$ of the intervals. Since an interval has weight of at most $\frac{\epsilon\gamma}{2}$ under the distribution, the sum of weights of all such intervals under the distribution is at most $\frac{2}{\gamma} \times \frac{\epsilon\gamma}{2} = \epsilon$ with probability of at least $1 - \frac{\delta}{2}$.

To remove the assumption that no instance is sampled more than once, we replace the input space $\mathcal{X}$ and distribution D with a new input space $\mathcal{X}'$ and distribution D' under which an instance is very unlikely to be sampled more than once. Let $\mathcal{X}' = \mathcal{X} \times T$ and $D' = D \times U$ where U is the uniform distribution over the set $T = \{0, \ldots, 2^k - 1\}$ for $k = \lceil 2\log(m) + \log(\frac{1}{\delta})\rceil$. Then $\mathcal{X}'$ is linearly ordered under the lexicographic ordering *i.e.*, $(x_1, t_1) \leq (x_2, t_2)$ iff $x_1 < x_2$ or $x_1 = x_2$ and $t_1 < t_2$. Therefore, the chance that any instance is being labeled twice in a sample of size m is $\binom{m}{2}2^{-k} \leq m^2 2^{-k-1} \leq \frac{\delta}{2}$.

Therefore, L with probability of at least $(1 - \frac{\delta}{2})(1 - \frac{\delta}{2}) > 1 - \delta$ will return a function $\hat{f}$ such that $Pr_{x\sim D}[|f^*(x) - \hat{f}(x)| > \gamma] \leq \epsilon$. □

Using the above theorem, we can easily obtain the following corollary.

Corollary 3.9. *Let $\rho \geq 0$ and $\epsilon, \delta, \gamma \in (0, \frac{1}{2})$. Let $\mathcal{X} = \mathbb{R}$. There exists an algorithm L, that for any target concept $c^* \in \mathcal{C}^1_{\text{thr}}$ and $D \in \mathcal{D}_{\mathcal{X}}$, using*

$$m = \lceil \frac{4}{\epsilon\gamma} \rceil \lceil max(\frac{64\ln(\frac{2^{21}}{(\epsilon\gamma)^2\delta})}{\epsilon\gamma}, \frac{2\ln(\frac{4\lceil\frac{4}{\epsilon\gamma}\rceil}{\delta})}{\gamma^2}) \rceil$$

examples drawn with respect to $\mathrm{EX}(c^*, D)$, will return a function $\hat{f} : \mathcal{X} \to [0, 1]$ with the property that

$$Pr_{x\sim D}[|\hat{f}(x) - \Phi_{\mathrm{drball}(\rho)_{c^*,D}}(x)| > \gamma] \leq \epsilon$$

with probability of at least $1 - \delta$.

Proof. In Lemma 3.7, we proved that the noisy (probabilistic) concepts resulting from applying ρ-distance random classification noise model to concepts in $\mathcal{C}^1_{\text{thr}}$ from a set of non-decreasing functions. Therefore, we can directly apply Theorem 3.8. □

Now, we have all the prerequisites to show that the class of one-dimensional threshold functions is $\mathrm{PAC}_{0,0}$-learnable with respect to distance random classification noise and $\mathcal{D}_{\mathcal{X}}$.

Proposition 3.10. *Let $\rho \geq 0$. Let $\mathcal{X} = \mathbb{R}$. $\mathcal{C}^1_{\text{thr}}$ is $\mathrm{PAC}_{0,0}$-learnable with respect to $\Phi_{\mathrm{drball}(\rho)}$ and $\mathcal{D}_{\mathcal{X}}$.*

Proof. Let $\epsilon, \delta \in (0, \frac{1}{2})$ and $D \in \mathcal{D}_{\mathcal{X}}$ a distribution. Let

$$s_1 = \lceil \frac{64}{\epsilon^2\delta} \rceil,$$

and

$$s_2 = \lceil \max\,(\frac{2^{10}\ln(\frac{2^{31}}{\epsilon^4\delta^3})}{\epsilon^2\delta}, \frac{2^5\ln(\frac{2^4 s_1}{\delta})}{\epsilon^4}) \rceil.$$

The learning algorithm $\mathcal{L}$ starts by drawing a sample $\mathcal{S}$ of $m = s_1 s_2$ examples (x_i, y_i) where $x_i \in \mathcal{X}$ and $y_i \in \{0, 1\}$ for $1 \leq i \leq m$. Let the examples be sorted by their instances such that $x_1 \leq \ldots \leq x_m$. First let us assume $x_1 < \ldots < x_m$. Later we show how to remove this assumption. $\mathcal{X}$ can be partitioned into s_1 disjoint intervals, I_j, $1 \leq j \leq s_1$, as in the proof sketch of Theorem 3.8. Also, using the same argument as in the proof sketch of Theorem 3.8, $Pr_{x\sim D}[x \in I_j]$ is at most $\frac{\epsilon^2\delta}{32}$ with probability of at least $1 - \frac{\delta}{8}$ for all $1 \leq j \leq s_1$.

After drawing the sample two cases can happen based on whether the sample has a clear boundary between points with different labels or not. We will define $\mathcal{L}$ by a case distinction and for each case we prove the learner will fulfill the learnability conditions of Definition 2.9.

1. There exists at least one set of three examples $\{(x_1, y), (x_2, 1 - y), (x_3, y)\} \subseteq \mathcal{S}$ such that $x_1 < x_2 < x_3$ for some $y \in \{0, 1\}$ (see Figure 3.4).
   Let $\theta^* \in \mathbb{R}$ such that c_{θ^*} is the target concept and c the concept returned by $\mathcal{L}$. In this case,

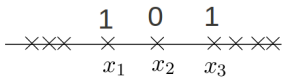

Figure 3.4: Case 1 in the proof of Proposition 3.10 for $y = 1$

$\text{dist}(x_2, \theta^*) \leq \rho$ because at least the label of one of the instances x_1, x_2 or x_3 is flipped by $\Phi_{\text{drball}(\rho)}$. Then based on the number of intervals that intersect with $\text{DB}_\rho(x_2)$, two subcases can happen.

1(a). $\text{DB}_\rho(x_2)$ has intersection with at most $\lfloor \frac{32}{\epsilon\delta} \rfloor$ intervals $i.e.$, there exists no $n_1, n_2 \in \mathbb{N}, n_1 \geq \lfloor \frac{32}{\epsilon\delta} \rfloor$ and $n_2 \leq s_1 - n_1$ such that for any $0 \leq i \leq n_1 - 1$

$$\text{DB}_\rho(x_2)_{|\mathcal{S}} \cap I_{n_2+i} \neq \emptyset.$$

Let $c \in \mathcal{C}^1_{\text{thr}}$, returned by $\mathcal{L}$, be such that $c(x) = 1$ iff $x \geq x_2$. Note that because $\text{dist}(x_2, \theta^*) \leq \rho$, $Pr_{x\sim D}[c(x) \neq c_{\theta^*}(x)] \leq Pr_{x\sim D}[x \in \text{DB}_\rho(x_2)]$. We show that c has error of at most ϵ with probability of at least $1 - \frac{\delta}{8}$. This happens because $\text{DB}_\rho(x_2)$ intersects with at most $\lfloor \frac{32}{\epsilon\delta} \rfloor$ intervals and, with probability of at least $1 - \frac{\delta}{8}$, $Pr_{x\sim D}[x \in I_j] \leq \frac{\epsilon^2\delta}{32}$ for all $1 \leq j \leq s_1$. Therefore, $Pr_{x\sim D}[c(x) \neq c_{\theta^*}(x)] \leq Pr_{x\sim D}[x \in \text{DB}_\rho(x_2)] \leq \lfloor \frac{32}{\epsilon\delta} \rfloor \times \frac{\epsilon^2\delta}{32} \leq \frac{32}{\epsilon\delta} \times \frac{\epsilon^2\delta}{32} = \epsilon$ with probability of $1 - \frac{\delta}{8}$.

1(b). $\text{DB}_\rho(x_2)$ has intersection with more than $\lfloor \frac{32}{\epsilon\delta} \rfloor$ intervals $i.e.$, there exists some $n_1, n_2 \in \mathbb{N}, n_1 \geq \lfloor \frac{32}{\epsilon\delta} \rfloor$ and $n_2 \leq s_1 - n_1$ such that for any $0 \leq i \leq n_1 - 1$

$$\text{DB}_\rho(x_2) \cap I_{n_2+i} \neq \emptyset.$$

We can use algorithm L in Corollary 3.9 to find a function $\hat{f} : \mathcal{X} \to [0, 1]$ such that with probability of at least[3] $1 - \frac{\delta}{8}$, for x in at least $\lfloor (1 - \frac{\delta}{4})s_1 \rfloor$ of the intervals we have $|\hat{f}(x) - \Phi_{\text{drball}(\rho)_{c_{\theta^*},D}}(x)| \leq \frac{\epsilon^2}{4} < \frac{\epsilon}{4}$.

Let $\text{DB}_\rho(x_2)_{|\mathcal{S}}$ be the set of instances in $\mathcal{S}$ that are in $\text{DB}_\rho(x_2)$. Let $\mathcal{L}$ return a (not necessarily unique) concept $c \in \mathcal{C}^1_{\text{thr}}$ that labels the leftmost $\lfloor (1-\hat{f}(x_2))|\text{DB}_\rho(x_2)_{|\mathcal{S}}| \rfloor$ points in $\text{DB}_\rho(x_2)_{|\mathcal{S}}$ with 0 and labels the rightmost $\lceil \hat{f}(x_2)|\text{DB}_\rho(x_2)_{|\mathcal{S}}| \rceil$ points in $\text{DB}_\rho(x_2)_{|\mathcal{S}}$ with 1 (see Figure 3.5). We now show that c has error of at most $\frac{\epsilon}{2}$ with probability of at least $1 - \frac{3\delta}{8}$.

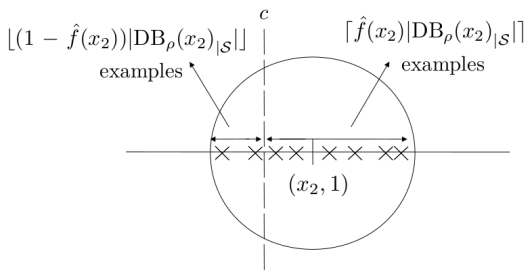

Figure 3.5: Case 1(b). in the proof of Proposition 3.10

[3]This probability is not $1 - \frac{\delta}{4}$ because we have already assumed that no instance is sampled more than once.

Since more than $\lfloor \frac{32}{\epsilon\delta} \rfloor$ intervals intersect with $\mathrm{DB}_\rho(x_2)$, we can find a lower bound on the cardinality of $\mathrm{DB}_\rho(x_2)_{|\mathcal{S}}$.

$$\begin{aligned}
|\mathrm{DB}_\rho(x_2)_{|\mathcal{S}}| &\geq (\lfloor \frac{32}{\epsilon\delta} \rfloor + 1 - 2) \times s_2 + 2 \\
&\geq (\frac{32}{\epsilon\delta} - 2) \times \lceil \max(\frac{2^{10}\ln(\frac{2^{31}}{\epsilon^4\delta^3})}{\epsilon^2\delta}, \frac{2^5 \ln(\frac{2^4 s_1}{\delta})}{\epsilon^4}) \rceil + 2 \\
&\geq (\frac{32}{\epsilon\delta} - 2) \times \max(\frac{2^{10}\ln(\frac{2^{31}}{\epsilon^4\delta^3})}{\epsilon^2\delta}, \frac{2^5 \ln(\frac{2^4 s_1}{\delta})}{\epsilon^4}) \\
&> (\frac{32}{\epsilon\delta} - 2) \times \frac{32}{\epsilon^2}\ln(\frac{2^{10}}{\epsilon^2\delta}) \\
&= \frac{2^{10}}{\epsilon^3\delta}\ln(\frac{2^{10}}{\epsilon^2\delta}) - \frac{64}{\epsilon^2}\ln(\frac{2^{10}}{\epsilon^2\delta}) \\
&> \frac{2^5}{\epsilon^2}\ln(\frac{4}{\delta})
\end{aligned}$$

Using Lemma A.1 in the Appendix, $\frac{2^5}{\epsilon^2}\ln(\frac{4}{\delta})$ examples are sufficient so that with probability of at least $1 - \frac{\delta}{4}$, $Pr_{x\sim D}[x \in \mathrm{DB}_\rho(x_2)$ and $c(x) = 0]$ deviates no more than $\frac{\epsilon}{4}$ from $1 - \hat{f}(x_2)$. Since $\hat{f}$ has error of at most $\frac{\epsilon}{4}$ with probability of at least $1 - \frac{\delta}{8}$, therefore, $Pr_{x\sim D}[x \in \mathrm{DB}_\rho(x_2)$ and $c(x) = 0]$ deviates at most by $\frac{\epsilon}{4} + \frac{\epsilon}{4} = \frac{\epsilon}{2}$ from $1 - \hat{f}(x_2)$ with probability of at least $(1 - \frac{\delta}{8})(1 - \frac{\delta}{4}) > 1 - \frac{3\delta}{8}$. Therefore, with probability of at least $1 - \frac{3\delta}{8}$, $Pr_{x\sim D}[c(x) \neq c_{\theta^*}(x)] \leq \frac{\epsilon}{2}\Pr_{x\sim D}[x \in \mathrm{DB}_\rho(x_2)] \leq \frac{\epsilon}{2}$.

2. There exists no set of three examples $\{(x_1, y), (x_2, 1-y), (x_3, y)\} \subseteq \mathcal{S}$ such that $x_1 < x_2 < x_3$ for any $y \in \{0, 1\}$.

Figure 3.6: x_- and x_+ in the proof of Proposition 3.10

Let $\theta^* \in \mathbb{R}$ such that c_{θ^*} is the target concept. Also let us define x_- and x_+ as follows (see Figure 3.6).

$$x_- = \begin{cases} \max\{x \mid (x, 0) \in \mathcal{S}\}, & \text{if } \exists (x, 0) \in \mathcal{S}, \\ x_+ - 1, & \text{if } \nexists (x, 0) \in \mathcal{S}, \end{cases}$$

and

$$x_+ = \begin{cases} \min\{x \mid (x, 1) \in \mathcal{S}\}, & \text{if } \exists (x, 1) \in \mathcal{S}, \\ x_- + 1, & \text{if } \nexists (x, 1) \in \mathcal{S}. \end{cases}$$

Let $\mathcal{L}$ return a (not necessarily unique) concept $c \in \mathcal{C}^1_{\mathrm{thr}}$ such that $c(x) = 1$ iff $x \geq x_+$.

We know that $\theta^* \in [x_- - \rho, x_+ + \rho)$ because otherwise neither x_- nor x_+ could be sampled with labels 0 and 1 respectively.

We can use the same case distinction that we used for the number of intervals that intersect $\mathrm{DB}_\rho(x_2)$ for the number of intervals that intersect $\theta^* \in [x_- - \rho, x_+ + \rho)$. The proof is similar in each of the subcases to the corresponding subcases in Case 1. Therefore, c has an error of at most $\epsilon = \max(\epsilon, \frac{3\epsilon}{8})$ with probability of at least $\min(1 - \frac{\delta}{8}, 1 - \frac{3\delta}{8}) = 1 - \frac{3\delta}{8}$.

To remove the assumption that no instance is sampled more than once, we can use the same technique as in the sketch of the proof of Theorem 3.8. We can replace the input space $\mathcal{X}$ and distribution D with a new input space $\mathcal{X}'$ and distribution D' under which an instance is very unlikely to be sampled more than once. Let $\mathcal{X}' = \mathcal{X} \times T$ and $D' = D \times U$ where U is the uniform distribution over the set $T = \{0, \ldots, 2^k - 1\}$ for $k = \lceil 2\log(m) + \log(\frac{1}{\delta})\rceil$. Therefore, the chance that any instance is being labeled twice in a sample of size m is $\binom{m}{2}2^{-k} \leq m^2 2^{-k-1} \leq \frac{\delta}{8}$.

Therefore, $\mathcal{L}$ returns a concept that has error of at most ϵ with probability of at least $(1 - \frac{3\delta}{8})(1 - \frac{\delta}{8}) > 1 - \frac{\delta}{2} > 1 - \delta$. □

3.2 Weight ball noise models

In the next two noise models, we use the mass under the underlying distribution as the measure of proximity.

Definition 3.11. *For any $D \in \mathcal{D}_{\mathcal{X}}$, $\omega \in [0, 1]$, and $x \in \mathcal{X}$, the* ω-weight ball $\mathrm{WB}_\omega(x)$ *around x is defined as*

$$\mathrm{WB}_\omega(x) = \mathrm{DB}_\rho(x)\,, \text{ where } \rho = \sup\{\rho' \mid \Pr_{x' \sim D}[x' \in DB_{\rho'}(x)] \leq \omega\}\,.$$

Our first weight ball label noise model is called ω-weight malicious classification noise[4].

Definition 3.12. *Let $\omega \in [0, 1]$. The* ω-weight malicious classification noise model $\Phi_{\mathrm{wmball}(\omega)}$ *is a label noise model, defined by*

$$\Phi_{\mathrm{wmball}(\omega)}(c, D, x) = \begin{cases} \{c(x)\}\,, & \text{if } c(x) = c(x') \text{ for all } x' \in \mathrm{WB}_\omega(x)\,, \\ [0, 1]\,, & \text{otherwise}\,, \end{cases}$$

where $D \in \mathcal{D}_{\mathcal{X}}$, $c \in 2^{\mathcal{X}}$, and $x \in \mathcal{X}$.

For every example (x, y) that a learning algorithm $\mathcal{L}$ draws from the oracle $\mathrm{EX}_{\Phi_{\mathrm{wmball}(\omega)}}(c, D)$, $y = c(x)$ will be guaranteed if all points in the ω-weight ball around x have the same label under c. If both positively and negatively labeled points lie in the ω-weight ball around x, we cannot make any assumption at all about the label y.

The label noise model here depends on the distribution D. Moreover, the value of a noisy concept for an instance $x \in \mathcal{X}$ depends on some of the values of the underlying target concept c other than the value $c(x)$ itself.

Like the malicious classification noise and the distance malicious classification noise models, the model of weight malicious classification noise is a non-deterministic label noise model.

The adversarial power in weight malicious classification noise is not as strong as the adversarial power in distance malicious classification noise. To illustrate this, we show that the class of one-dimensional threshold-functions, which is not $\mathrm{PAC}_{\underline{\epsilon},0}$-learnable with respect to ρ-distance malicious classification noise and $\mathcal{D}_{\mathcal{X}}$ for any $\underline{\epsilon} \in [0, \frac{1}{2})$, is $\mathrm{PAC}_{\omega,0}$-learnable with respect to ω-weight malicious classification noise and $\mathcal{D}_{\mathcal{X}}$.

Proposition 3.13. *Let $\omega \in [0, \frac{1}{2})$. Let $\mathcal{X} = \mathbb{R}$. $\mathcal{C}^1_{\mathrm{thr}}$ is $\mathrm{PAC}_{\omega,0}$-learnable with respect to $\Phi_{\mathrm{wmball}(\omega)}$ and $\mathcal{D}_{\mathcal{X}}$.*

Proof. Let $\epsilon \in (\omega, \frac{1}{2})$, $\delta \in (0, \frac{1}{2})$ and $\omega' = \epsilon - \omega$. Let the learning algorithm $\mathcal{L}$ be as follows. $\mathcal{L}$ upon seeing a sample $\mathcal{S}$ returns any concept in the concept class that has the smallest number of wrong label predictions for instances in $\mathcal{S}$.[5] We will show that using a sample $\mathcal{S}$ of

$$m \geq \frac{2}{\epsilon - \omega}(\ln(\frac{1}{\delta}) + \ln(\frac{2}{\epsilon - \omega})) = \frac{2}{\omega'}(\ln(\frac{1}{\delta}) + \ln(\frac{2}{\omega'}))$$

[4] We simply write weight malicious classification noise instead of ω-weight malicious classification noise when ω is clear from the context.

[5] This is the minimum pn-disagreement strategy (Definition 5.9) that will be introduced in Chapter 5.

examples, the concept returned by $\mathcal{L}$ will have an error of at most ϵ with probability of at least $1-\delta$.

Let $M = \{x \mid (x,y) \in \mathcal{S} \text{ for some } y \in \{0,1\}\} \cup \{-\infty, +\infty\}$ be the union of the set of distinct instances in the sample with $\{-\infty, +\infty\}$. Then $|M| \leq m+2$. Let x_1 and x_2 be any two consecutive points in M $i.e.$, there is no $x_3 \in M$ such that $x_1 < x_3 < x_2$. Using Lemma A.5 from the Appendix, m examples guarantee that

$$\text{with probability of at least } 1-\delta,\ Pr_{x\sim D}[x_1 < x < x_2] \leq \omega'. \tag{3.3}$$

Also let

$$\mathcal{X}_- = \{x \mid \exists x' \geq x : (x',0) \in \mathcal{S} \text{ and } \forall x'' \leq x : (x'',1) \notin \mathcal{S}\},$$
$$\mathcal{X}_+ = \{x \mid \exists x' \leq x : (x',1) \in \mathcal{S} \text{ and } \forall x'' \geq x : (x'',0) \notin \mathcal{S}\},$$
$$\mathcal{X}_{-+} = \mathcal{X} - \mathcal{X}_- - \mathcal{X}_+.$$

Note that $\mathcal{L}$ can always return a concept c_θ where $\theta \in \mathcal{X}_{-+}$. θ does not need to be in $\mathcal{X}_-$ because for such θ there exists $e > 0$ such that $c_{\theta+e}$ where $\theta + e \in \mathcal{X}_{-+}$ has at most the same number of wrong label predictions on the instances of $\mathcal{S}$ than c_θ. Also θ does not need to be in $\mathcal{X}_+$ because for such θ there exists $e > 0$ such that $c_{\theta-e}$ where $\theta - e \in \mathcal{X}_{-+}$ has at most the same number of wrong label predictions on the instances of $\mathcal{S}$ than c_θ.

After drawing the sample, two cases can happen based on whether the sample has a clear boundary between points with different labels or not $i.e.$, whether $M \cap \mathcal{X}_{-+} = \emptyset$ or $M \cap \mathcal{X}_{-+} \neq \emptyset$. We consider these two cases separately below.

1. $M \cap \mathcal{X}_{-+} = \emptyset$. Also let us assume $M \cap \mathcal{X}_- \neq \emptyset$ and $M \cap \mathcal{X}_+ \neq \emptyset$. The same argument as here can be used when either $M \cap \mathcal{X}_+ = \emptyset$ or $M \cap \mathcal{X}_- = \emptyset$, following Cases 1(a) and 1(b), respectively.

   Let c_θ be the concept returned by $\mathcal{L}$ and c_{θ^*} the target concept. Based on the position of θ and θ^* the following two subcases (1(a) and 1(b)) can be considered.

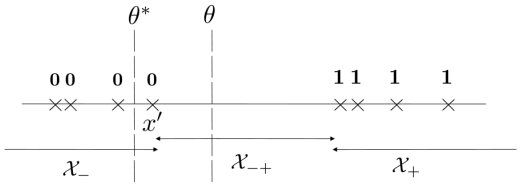

Figure 3.7: Case 1(a) part (i) in the proof of Proposition 3.13

1(a). $\theta^* < \theta$. Let x' be such that $(x',1) \in \mathcal{S}$ and, for all $(x,1) \in \mathcal{S}$, $\text{dist}(x',\theta) \leq \text{dist}(x,\theta)$. Now, based on the position of θ^* and x' the following two cases can happen.

(i). If $\theta^* \leq x'$ (see Figure 3.7) then $Pr_{x\sim D}[c_\theta(x) \neq c_{\theta^*}(x)] \leq Pr_{x\sim D}[\theta \leq x < x'] = Pr_{x\sim D}[\theta < x < x'] \leq \omega' < \epsilon$ with probability of at least $1-\delta$, based on the property in Equation 3.3.

(ii). Otherwise (see Figure 3.8) $Pr_{x\sim D}[c_\theta(x) \neq c_{\theta^*}(x)] = Pr_{x\sim D}[\theta \leq x < x'] + Pr_{x\sim D}[x' \leq x < \theta^*] = Pr_{x\sim D}[\theta < x < x'] + Pr_{x\sim D}[x' \leq x < \theta^*]$ because $\theta^* > x'$. Furthermore, $Pr_{x\sim D}[\theta < x < x'] \leq \omega'$ with probability of at least $1-\delta$ based on the property in Equation 3.3. Also $Pr_{x\sim D}[x' \leq x < \theta^*] \leq Pr_{x\sim D}[x' \leq x \leq \theta^*] \leq \omega$ because otherwise x' would have been labeled 0 in $\mathcal{S}$. Therefore, $Pr_{x\sim D}[c_\theta(x) \neq c_{\theta^*}(x)] \leq \omega' + \omega = \epsilon$ with probability of at least $1-\delta$.

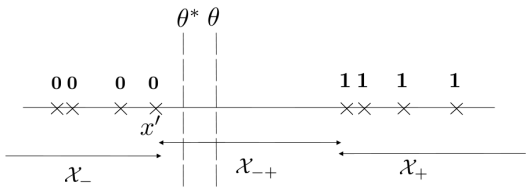

Figure 3.8: Case 1(a) part (ii) in the proof of Proposition 3.13

1(b). $\theta^* \geq \theta$. Let x' such that $(x', 0) \in \mathcal{S}$ and, for all $(x, 0) \in \mathcal{S}$, $\text{dist}(x', \theta) \leq \text{dist}(x, \theta)$. Now, based on the position of θ^* and x' the following two cases can happen.

(i). If $\theta^* \leq x'$ then $Pr_{x\sim D}[c_\theta(x) \neq c_{\theta^*}(x)] = Pr_{x\sim D}[\theta^* \leq x \leq x'] + Pr_{x\sim D}[x' < x < \theta]$ because $\theta^* > x'$. Furthermore, $Pr_{x\sim D}[\theta^* \leq x \leq x'] \leq \omega$ because otherwise x' would have been labeled 1 in $\mathcal{S}$. Also $Pr_{x\sim D}[x' < x < \theta] \leq \omega'$ with probability of at least $1 - \delta$ based on the property in Equation 3.3. Therefore, $Pr_{x\sim D}[c_\theta(x) \neq c_{\theta^*}(x)] \leq \omega + \omega' = \epsilon$ with probability of at least $1 - \delta$.

(ii). Otherwise, based on the property in Equation 3.3, $Pr_{x\sim D}[c_\theta(x) \neq c_{\theta^*}(x)] = Pr_{x\sim D}[\theta^* \leq x < \theta] < Pr_{x\sim D}[x' < x < \theta] \leq \omega' < \epsilon$ with probability of at least $1 - \delta$.

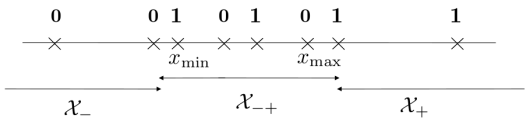

Figure 3.9: Case 2 in the proof of Proposition 3.13

2. $M \cap \mathcal{X}_{-+} \neq \emptyset$. Let $x_{\text{max}} = \max\{x \mid x \in \mathcal{X}_{-+} \wedge (x, 0) \in \mathcal{S}\}$ and $x_{\text{min}} = \min\{x \mid x \in \mathcal{X}_{-+} \wedge (x, 1) \in \mathcal{S}\}$ (see Figure 3.9). Note that both x_{min} and x_{max} exists (if only one of x_{min} and x_{max} exist, it would have been contained in either $\mathcal{X}_-$ or $\mathcal{X}_+$ and thus not it $\mathcal{X}_{-+}$.) Let c_θ be the concept returned by $\mathcal{L}$ and c_{θ^*} the target concept. If θ^* is greater than or equal to θ then $Pr_{x\sim D}[c_\theta(x) \neq c_{\theta^*}(x)] \leq \omega < \epsilon$ because otherwise x_{max} would have been labeled 1. Similarly, if $\theta^* < \theta$ then $Pr_{x\sim D}[c_\theta(x) \neq c_{\theta^*}(x)] \leq \omega < \epsilon$ because otherwise x_{min} would have been labeled 0. Therefore, c_θ will have an error of at most ϵ with probability 1 in this case.

□

However, the class of one-dimensional threshold functions is still not $\text{PAC}_{\underline{\epsilon},0}$-learnable with respect to ω-weight malicious classification noise and $\mathcal{D}_\mathcal{X}$ when $\underline{\epsilon} \in [0, \omega)$.

Proposition 3.14. *Let $\omega \in (0, \frac{1}{2})$, $\underline{\epsilon} \in [0, \omega)$ and $\underline{\delta} \in [0, \frac{1}{2})$. Let $\mathcal{X} = \mathbb{R}$. $\mathcal{C}^1_{\text{thr}}$ is not $\text{PAC}_{\underline{\epsilon},\underline{\delta}}$-learnable with respect to $\Phi_{\text{wmball}(\omega)}$ and $\mathcal{D}_\mathcal{X}$.*

Proof. Let $\Omega \in (\underline{\epsilon}, \omega)$. Let $x_1, x_2, x_3 \in \mathcal{X}$ such that

$$x_1 < x_2 < x_3,$$

and

$$\text{dist}(x_1, x_2) < \text{dist}(x_2, x_3).$$

Let D be such that

$$Pr_{x\sim D}[x = x_1] = \Omega, Pr_{x\sim D}[x = x_2] = \omega - \Omega, Pr_{x\sim D}[x = x_3] = 1 - \omega,$$

and

$$Pr_{x\sim D}[x \notin \{x_1, x_2, x_3\}] = 0.$$

Therefore, $\text{supp}(D) = \{x_1, x_2, x_3\}$.

Let $\mathcal{C} = \{c, c'\}$ with $c(x) = c_{x_1-\text{dist}(x_1,x_2)}(x)$ and $c'(x) = c_{x_2}(x)$ for all $x \in \mathcal{X}$. We show that $\text{EX}_{\Phi_{\text{wmball}(\omega)}}(c, D)$ and $\text{EX}_{\Phi_{\text{wmball}(\omega)}}(c', D)$ have potentially equivalent behavior on D. Therefore, we can use Lemma 2.12 to show that $\mathcal{C}^1_{\text{thr}}$ is not $\text{PAC}_{\underline{\epsilon},\underline{\delta}}$-learnable with respect to $\Phi_{\text{wmball}(\omega)}$ and $\mathcal{D}_{\mathcal{X}}$.

For any $x \in \text{supp}(D)$, $\Phi_{\text{wmball}(\omega)}(c, D, x) = \{1\}$ because the ω-weight ball around x only contains points with label 1. For x_3, $\Phi_{\text{wmball}(\omega)}(c', D, x_3) = \{1\}$. Also for any $x \in \{x_1, x_2\}$, $\Phi_{\text{wmball}(\omega)}(c', D, x) = [0, 1]$ because the ω-weight ball around x contains points with both labels 0 and 1. Thus, $\Phi_{\text{wmball}(\omega)}(c, D, x) \cap \Phi_{\text{wmball}(\omega)}(c', D, x) \neq \emptyset$ for all $x \in \text{supp}(D)$. Therefore, based on Definition 2.10, $\text{EX}_{\Phi_{\text{wmball}(\omega)}}(c, D)$ and $\text{EX}_{\Phi_{\text{wmball}(\omega)}}(c', D)$ have potentially equivalent behavior on D.

Since $Pr_{x\sim D}[c(x) \neq c'(x)] = \Omega > \underline{\epsilon}$, therefore, $\mathcal{C}^1_{\text{thr}}$ is not $\text{PAC}_{\underline{\epsilon},\underline{\delta}}$-learnable with respect to $\Phi_{\text{wmball}(\omega)}$ and $\mathcal{D}_{\mathcal{X}}$. □

Although not as strong as distance malicious classification noise, still the adversarial characteristics of distance weight classification noise is strong enough to make simple concept classes not PAC-learnable with respect to such a noise model. One such concept class is the class of two-dimensional axis-parallel halfspaces defined as follows.

Definition 3.15. *Let $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{C}^2_{\text{thr}} = \{c_\theta \mid \theta \in \mathbb{R}\}$, where*

$$c_\theta(x, x') = \begin{cases} 1\,, & \text{if } x \geq \theta\,, \\ 0\,, & \text{if } x < \theta\,. \end{cases}$$

$\mathcal{C}^2_{\text{thr}}$ is called the class of two-dimensional axis-parallel halfspaces.

We show that the class of two-dimensional axis-parallel halfspaces is not $\text{PAC}_{\underline{\epsilon},\underline{\delta}}$-learnable with respect to weight malicious classification noise and $\mathcal{D}_{\mathcal{X}}$ for any $\underline{\epsilon}, \underline{\delta} \in (0, \frac{1}{2})$.

Proposition 3.16. *Let $\omega \in (0, \frac{1}{2})$ and $\underline{\epsilon}, \underline{\delta} \in [0, \frac{1}{2})$. Let $\mathcal{X} = \mathbb{R}^2$. $\mathcal{C}^2_{\text{thr}}$ is not $\text{PAC}_{\underline{\epsilon},\underline{\delta}}$-learnable with respect to $\Phi_{\text{wmball}(\omega)}$ and $\mathcal{D}_{\mathcal{X}}$.*

Proof. Let $m = \lceil \frac{1}{\omega} \rceil$. Let $\{x_1, \ldots, x_{2m}\} \in \mathcal{X}$ such that for $1 \leq i \leq m$, $x_i = (0, i)$ and for $1 + m \leq i \leq 2m$, $x_i = (\frac{1}{2}, i - m)$ (see Figure 3.10). Let D be a probability distribution such that for $1 \leq i \leq m$, $Pr_{x\sim D}[x = x_i] = \frac{1+\omega}{2m}$ and for $1 + m \leq i \leq 2m$, $Pr_{x\sim D}[x = x_i] = \frac{1-\omega}{2m}$. Therefore, $\text{supp}(D) = \{x_1, \ldots, x_{2m}\}$.

Let $\mathcal{C} = \{c, c'\}$ where $c(x) = c_{-1}(x)$ and $c'(x) = c_1(x)$ for all $x \in \mathcal{X}$. We show that $\text{EX}_{\Phi_{\text{wmball}(\omega)}}(c, D)$ and $\text{EX}_{\Phi_{\text{wmball}(\omega)}}(c', D)$ have potentially equivalent behavior on D. Therefore, we can use Lemma 2.12 to show that $\mathcal{C}^2_{\text{thr}}$ is not $\text{PAC}_{\underline{\epsilon},\underline{\delta}}$-learnable with respect to $\Phi_{\text{wmball}(\omega)}$ and $\mathcal{D}_{\mathcal{X}}$.

For any $x \in \text{supp}(D)$, $\Phi_{\text{wmball}(\omega)}(c, D, x) = \{1\}$ because the ω-weight ball around x only contains points with label 1. Moreover, for any $x \in \text{supp}(D)$, $\Phi_{\text{wmball}(\omega)}(c', D, x) = [0, 1]$ because the ω-weight ball around x contains points with both labels 0 and 1.[6] Thus, $\Phi_{\text{wmball}(\omega)}(c, D, x) \cap \Phi_{\text{wmball}(\omega)}(c', D, x) \neq \emptyset$ for all $x \in \text{supp}(D)$. Using Definition 2.10, $\text{EX}_{\Phi_{\text{wmball}(\omega)}}(c, D)$ and $\text{EX}_{\Phi_{\text{wmball}(\omega)}}(c', D)$ have potentially equivalent behavior on D.

Since $Pr_{x\sim D}[c(x) \neq c'(x)] = 0.5 > \underline{\epsilon}$ as $\underline{\epsilon}[0, \frac{1}{2})$, therefore, $\mathcal{C}^2_{\text{thr}}$ is not $\text{PAC}_{\underline{\epsilon},\underline{\delta}}$-learnable with respect to $\Phi_{\text{wmball}(\omega)}$ and $\mathcal{D}_{\mathcal{X}}$. □

[6] Because for any point x_i, $1 \leq i \leq m$, the ball around x_i only contains x_i and x_{i+m} and for any point x_i, $1 + m \leq i \leq 2m$, the ball around x_i only contains x_i and x_{i-m}.

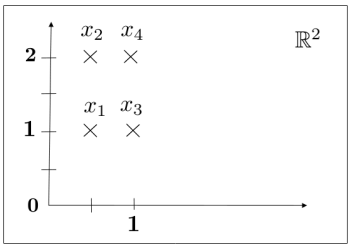

Figure 3.10: Geometry of the input space when $\omega = \frac{7}{12}$ and $m = 2$.

Like distance malicious classification noise, it seems like choosing an arbitrary label from the weight ball around a point as the label of that point gives too much freedom to the oracle. We mitigate this a bit in our next noise model, called the ω-weight random classification noise[7]. This model is a deterministic version of ω-weight malicious classification noise.

Definition 3.17. *Let $\omega \in [0, 1]$. The* ω-weight random classification noise model $\Phi_{\mathrm{wrball}(\omega)}$ *is a label noise model defined as*

$$\Phi_{\mathrm{wrball}(\omega)}(c, D, x) = \{\Pr_{x' \sim D}[c(x') = 1 \mid y \in \mathrm{WB}_\omega(x)]\},$$

where $D \in \mathcal{D}_\mathcal{X}$, $c \in 2^\mathcal{X}$, and $x \in \mathrm{supp}(D)$. $\Phi_{\mathrm{wrball}(\omega)}(c, D, x) = \{0\}$ for $x \notin \mathrm{supp}(D)$.

For every example (x, y) that a learning algorithm $\mathcal{L}$ draws from the oracle $\mathrm{EX}_{\Phi_{\mathrm{wrball}(\rho)}}(c, D)$, $y = c(x)$ will be guaranteed if all points in the ω-weight ball around x have the same label under c. If both positively and negatively labeled points lie in the ω-weight ball around x, then the label y will be drawn from $\{0, 1\}$ according to the distribution of labels within the ω-weight ball around x.

This label noise model depends on the distribution D. Moreover, the value of the noise function in an instance $x \in \mathcal{X}$ depends on some of the values of the underlying target concept c other than the value $c(x)$ itself.

Similar to distance ball models, we can show that any concept class that is PAC-learnable with respect to Φ_{wmball} and a class of distributions is also PAC-learnable with respect to Φ_{wrball} and the same class of distributions.

Proposition 3.18. *Let $\omega \in [0, 1]$ and $\underline{\epsilon}, \underline{\delta} \in [0, \frac{1}{2})$. Let $\mathcal{D} \subseteq \mathcal{D}_\mathcal{X}$ be a class of distributions. Any concept class $\mathcal{C}$ that is $\mathrm{PAC}_{\underline{\epsilon},\underline{\delta}}$-learnable with respect to $\Phi_{\mathrm{wmball}(\omega)}$ and $\mathcal{D}$ is also $\mathrm{PAC}_{\underline{\epsilon},\underline{\delta}}$-learnable with respect to $\Phi_{\mathrm{wrball}(\omega)}$ and $\mathcal{D}$.*

The proof is exactly the same as the proof of Proposition 3.6 replacing $\Phi_{\mathrm{dmball}(\rho)}$ and $\Phi_{\mathrm{drball}(\rho)}$ with $\Phi_{\mathrm{wmball}(\omega)}$ and $\Phi_{\mathrm{wrball}(\omega)}$ respectively.

one-dimensional threshold functions		
	ρ-distance	ω-weight
malicious classification noise model	No	Yes ($\underline{\epsilon} \in [\omega, \frac{1}{2}), \underline{\delta} \in [0, \frac{1}{2})$) No ($\underline{\epsilon} \in [0, \omega), \underline{\delta} \in [0, \frac{1}{2})$)
random classification noise model	Yes ($\underline{\epsilon}, \underline{\delta} \in [0, \frac{1}{2})$)	Yes ($\underline{\epsilon} \in [\omega, \frac{1}{2}), \underline{\delta} \in [0, \frac{1}{2})$) Open ($\underline{\epsilon} \in [0, \omega), \underline{\delta} \in [0, \frac{1}{2})$)

Table 3.1: Summary of learnability results with our noise models on $\mathcal{C}^1_{\mathrm{thr}}$.

[7]We simply write weight random classification noise instead of ω-weight random classification noise when ω is clear from the context.

Table 3.1 summarizes the learnability results of our noise models for the class of one-dimensional threshold functions. The question of whether one-dimensional threshold functions are $\mathrm{PAC}_{\underline{\epsilon},\underline{\delta}}$-learnable with respect to ω-weight random classification noise remains open for $\underline{\epsilon} \in [0, \omega)$ and $\underline{\delta} \in [0, \frac{1}{2})$.

Note that noise models that depend both on the target concept and on the instances make learning very difficult. We do not have any strong general learnability results for our new models. Still we hope that they can serve as first steps in the direction of modeling noise in a more realistic way.

Chapter 4

Learning under Label Noise in Finite Input Spaces

In this chapter, we assume that $\mathcal{X}$ is finite. Therefore, the VC-dimension of any concept class and the support of any distribution over $\mathcal{X}$ is finite. Having a finite input space makes the learning easier, as a learner can easily estimate the weight of the important parts of the distribution by sampling. Also throughout this chapter we assume the noise model is deterministic.

First, in the following corollary, we show that the learnability result with respect to the CPCN model (Theorem 2.17) implies learnability with respect to any arbitrary noise model when the input space is finite and the noise rate for any point is strictly less than a half. A condition on the noise model is that its value on instance x does not depend on the distribution and not on any other values of the target concept c^* than the value $c^*(x)$ itself.

Corollary 4.1. *Let $\mathcal{X}$ be a finite input space, $\mathcal{C}$ a concept class and $\mathcal{D}_\mathcal{X}$ the class of all distributions. Let Φ be a deterministic noise model such that*

1. $\mathrm{nr}_{c,D}(x) < \frac{1}{2}$ *for all $x \in \mathcal{X}$, $c \in \mathcal{C}$ and $D \in \mathcal{D}_\mathcal{X}$, and*
2. $\mathrm{nr}_{c,D}(x) = \mathrm{nr}_{c',D'}(x)$ *for all $D, D' \in \mathcal{D}_\mathcal{X}$ and all $c, c' \in \mathcal{C}$ with $c(x) = c'(x)$.*

Then $\mathcal{C}$ is $\mathrm{PAC}_{0,0}$-learnable with respect to Φ and $\mathcal{D}_\mathcal{X}$.

Proof. Let $c^* \in \mathcal{C}$ be the target concept and $D \in \mathcal{D}_\mathcal{X}$ a distribution. Let $k = |\{\mathrm{nr}_{c,D'}(x) \mid x \in \mathcal{X}\}|$ for some $c \in \mathcal{C}$ and $D' \in \mathcal{D}_\mathcal{X}$. Note that k is finite because $\mathcal{X}$ is finite. Let $\eta = (\eta_1, \ldots, \eta_k)$ be such that

1. $\eta_i \in \{\mathrm{nr}_{c^*,D}(x) \mid x \in \mathcal{X}\}$ for all $1 \leq i \leq k$.
2. $\eta_i \neq \eta_j$ for all $1 \leq i, j \leq k$ and $i \neq j$.

Let $\pi = (\pi_1, \ldots, \pi_k)$ such that $\pi_i = \{x \mid x \in \mathcal{X} \text{ and } \mathrm{nr}_{c^*,D}(x) = \eta_i\} \times \{0, 1\}$. Let Φ' be a CPCN model with parameters η and π. This is well-defined because of condition 2 of the corollary. $\mathrm{EX}_\Phi(c^*, D)$ and $\mathrm{EX}_{\Phi'}(c^*, D)$ have potentially equivalent behavior on D because $\Phi(c^*, D, x) = \Phi'(c^*, D, x)$ due to the way we construct Φ'.

Since $\pi_1 \cup \ldots \cup \pi_k = \mathcal{X} \times \{0, 1\}$ and $\eta \in [0, \frac{1}{2})^k$ (due to the assumption of the corollary), Theorem 2.17 can be used to show that $\mathcal{C}$ is $\mathrm{PAC}_{0,0}$-learnable with respect to Φ' and $\mathcal{D}_\mathcal{X}$. Using Lemma 2.11, $\mathcal{C}$ is also $\mathrm{PAC}_{0,0}$-learnable with respect to Φ and $\mathcal{D}_\mathcal{X}$ because $\mathrm{EX}_\Phi(c^*, D)$ and $\mathrm{EX}_{\Phi'}(c^*, D)$ have potentially equivalent behavior on D, and Φ and Φ' are both deterministic label noise models. □

Next we show that, when the input space is finite, we can generalize the result of Corollary 4.1 to the case that the noise rate for any point just needs to be different than a half instead of strictly less than a half. However, as in Corollary 4.1, the noise rate still depends only on the point itself.

Like Corollary 4.1 this problem can also be cast into a problem in the CPCN framework. However, we cannot use the learnability results with respect to CPCN anymore, because the learnability result with respect to CPCN only holds when the noise rate is strictly less than a half. First we need to state and prove the following lemma which is a generalization of Lemma A.5 in the Appendix.

Lemma 4.2. ***(Generalized Coupon Collector Problem)*** *Let $n \in \mathbb{N}$. Let $A_1, \ldots, A_k$ be events with probability greater than or equal to p. Then in a sequence of*

$$m = \frac{n}{p} \ln(\frac{k}{1-(1-\delta)^{\frac{1}{n}}}) \tag{4.1}$$

independent trials, the probability that every event occurs at least n times is at least $1-\delta$.

Proof. Using Lemma A.5 in the Appendix, replacing δ by $1-(1-\delta)^{\frac{1}{n}}$, a sequence of

$$m' = \frac{1}{p} \ln(\frac{k}{1-(1-\delta)^{\frac{1}{n}}})$$

independent trials is sufficient such that, with probability of at least $(1-\delta)^{\frac{1}{n}}$, every event $A_1, .., A_k$ occurs at least once.

Let the following denote a sequence of $n \times m'$ trials, partitioned into n pairwise disjoint sets of size m'.

$$\underbrace{t_{11}, \ldots, t_{1m'}}_{1} \underbrace{t_{21}, \ldots, t_{2m'}}_{2} \ldots \underbrace{t_{n1}, \ldots, t_{nm'}}_{n}$$

In each of these sets, with probability of at least $(1-\delta)^{\frac{1}{n}}$ each event $A_1, \ldots, A_k$ occurs at least once. The probability that in all of the sets each event $A_1, \ldots, A_k$ occurs at least once is at least $((1-\delta)^{\frac{1}{n}})^n = 1-\delta$ due to the independence of the trials. Therefore, with probability of at least $1-\delta$ each of the events $A_1, \ldots, A_k$ occurs at least n times in a sample of size $m = n \times m'$, as denoted in Equation 4.1. □

The statement of the generalization of Corollary 4.1 is as follows.

Proposition 4.3. *Let $\mathcal{X}$ be a finite input space. Let $\mathcal{C}$ be a concept class and Φ a deterministic noise model such that*

1. $\mathrm{nr}_{c,D}(x) \neq \frac{1}{2}$ *for all $x \in \mathcal{X}$, $c \in \mathcal{C}$ and $D \in \mathcal{D}_{\mathcal{X}}$, and*
2. $\mathrm{nr}_{c,D}(x) = \mathrm{nr}_{c',D'}(x)$ *for all $D, D' \in \mathcal{D}_{\mathcal{X}}$ and all $c, c' \in \mathcal{C}$ with $c(x) = c'(x)$.*

Then C is $\mathrm{PAC}_{0,0}$-learnable with respect to Φ and $\mathcal{D}_{\mathcal{X}}$.

Proof. Let $\mathcal{X} = \{x_1, \ldots, x_n\}$. Without loss of generality, let for all $i \in \{1, \ldots, n-1\}$, $Pr_{x\sim D}[x = x_i] \geq Pr_{x\sim D}[x = x_{i+1}]$. Then, for any $\epsilon > 0$, there exists a value $k_\epsilon \in \{1, \ldots, n\}$ such that $Pr_{x\sim D}[x = x_i] \geq \frac{\epsilon}{n}$ iff $i \leq k_\epsilon$. Thus

$$\sum_{i=k_\epsilon+1}^{n} Pr_{x\sim D}[x = x_i] < (n-k_\epsilon)\frac{\epsilon}{n} \leq (n-1)\frac{\epsilon}{n} < \epsilon$$

Let

$$\eta = \min_{1\leq i\leq n} |\frac{1}{2} - \mathrm{nr}_{c^*,D}(x_i)|$$

for arbitrary $c^* \in \mathcal{C}$ and $D \in \mathcal{D}_{\mathcal{X}}$ and $\mathcal{L}$ a learning algorithm with the following strategy. $\mathcal{L}$ draws a sample $\mathcal{S}$ of size m (as in Equation 4.7) examples from $\mathrm{EX}_\Phi(c^*, D)$. $\mathcal{L}$ returns a concept c that labels any point $x \in X$ as follows.

1. If x is sampled at least m_1 times in $\mathcal{S}$, where m_1 is as in Equation 4.2 and $\mathrm{nr}_{c^*,D}(x) < \frac{1}{2}$, c will return the label that is observed more often for x in the sample (if one label is observed more often than the other).

2. If x is sampled at least m_1 times in $\mathcal{S}$, where m_1 is as in Equation 4.2 and $\mathrm{nr}_{c^*,D}(x) > \frac{1}{2}$, c will return the label that is observed less often for x in the sample (if one label is observed more often than the other).

3. Otherwise, $c(x)$ will be defined such that $c \in \mathcal{C}$ if possible.[1]

In the rest of the proof, we exploit the finiteness of $\mathcal{X}$ to show that in order for c to have an error of at most ϵ with probability of at least $1-\delta$, it is sufficient for c to label points with higher mass under D ($\{x_1, \ldots, x_{k_\epsilon}\}$) correctly. Moreover, in order to predict a label of any point correctly with high probability, that point should be sampled often enough so that the learner can use the noise rate of that point to predict its label. Since $\mathcal{X}$ is finite, a large enough sample size will guarantee that with high probability any point that the learner is interested in will be sampled a sufficient number of times. The rest of this proof deals with computing a large enough sample size to achieve this goal.

When examples are drawn from $\mathrm{EX}_\Phi(c^*, D)$, the average of the labels observed for any instance x in $\mathcal{S}$ is a random variable with expectation $\Phi_{c^*,D}(x)$. Let for any instance x in $\mathcal{S}$, $\hat{\Phi}_{c^*,D}(x)$ denote an estimate of the expectation of this random variable computed from the sample. $\hat{\Phi}_{c^*,D}(x)$ can be easily calculated by summing up the labels that are observed for instance x, and dividing the sum by the total number of times that x is sampled.

If with high probability, for any $x \in \{x_1, \ldots, x_{k_\epsilon}\}$, $\hat{\Phi}_{c^*,D}(x)$ deviates less than η from its expectation, $\Phi_{c^*,D}(x)$, then x is seen with one label more often than with the other one in the sample. Knowing whether $\mathrm{nr}_{c^*,D}(x)$ is less or greater than a half, $\mathcal{L}$ can easily decide which label (the more observed one or the less observed one) is the correct label of x. Therefore, with high probability, $\mathcal{L}$ returns the correct label for the points $x_1, \ldots, x_{k_\epsilon}$. As we described, predicting the $\hat{\Phi}_{c^*,D}$ values with any amount of deviation strictly less than η is sufficient for $\mathcal{L}$ to return a concept that with probability of at least $1-\delta$ has error of at most ϵ. For simplicity, we choose the deviation of $\frac{\eta}{2}$.

We can use the Hoeffding inequality (Lemma A.1 in the Appendix) to bound the probability of deviation of $\hat{\Phi}_{c^*,D}$ from $\Phi_{c^*,D}$ by at most $\frac{\eta}{2}$. Based on this inequality, if any point $x \in \{x_1, \ldots, x_{k_\epsilon}\}$ is sampled at least

$$m_1 \geq \frac{2}{\eta^2 \ln(1-(1-\frac{\delta}{2})^{\frac{1}{n}})} \tag{4.2}$$

times, the estimation of $\hat{\Phi}_{c^*,D}(x)$ is guaranteed to deviate at most $\frac{\eta}{2}$ from $\Phi_{c^*,D}(x)$ with probability of at least $1-\frac{\delta}{2}$, as we will see next. Using the Hoeffding inequality, for any point $x \in \{x_1, \ldots, x_{k_\epsilon}\}$

$$Pr[|\hat{\Phi}_{c^*,D}(x) - \Phi_{c^*,D}(x)| \geq \frac{\eta}{2}] \leq e^{-2(\frac{\eta}{2})^2 m_1}$$

where the probability is taken over a sample of of size m, $i.i.d.$ from D. Inserting m_1 from Equation 4.2

$$Pr[|\hat{\Phi}_{c^*,D}(x) - \Phi_{c^*,D}(x)| \geq \frac{\eta}{2}] \leq 1-(1-\frac{\delta}{2})^{\frac{1}{n}}$$

Let $\delta' = 1-(1-\frac{\delta}{2})^{\frac{1}{n}}$. Therefore, for any point $x \in \{x_1, \ldots, x_{k_\epsilon}\}$

$$Pr[|\hat{\Phi}_{c^*,D}(x) - \Phi_{c^*,D}(x)| \leq \frac{\eta}{2}] \geq 1-\delta' \tag{4.3}$$

If we want Equation 4.3 to be satisfied for all the points $x \in \{x_1, \ldots, x_{k_\epsilon}\}$ at the same time

$$Pr[\forall x \in \{x_1, \ldots, x_{k_\epsilon}\} : |\hat{\Phi}_{c^*,D}(x) - \Phi_{c^*,D}(x)| \leq \frac{\eta}{2}] \geq (1-\delta')^{k_\epsilon} \geq (1-\delta')^n \tag{4.4}$$

[1] As we will see in the rest of this proof, with probability of at least $1-\delta$, this will be possible.

since $k_\epsilon \leq n$. Replacing δ' by $1-(1-\frac{\delta}{2})^{\frac{1}{n}}$ in Equation 4.4 and simplifying yields

$$Pr[\forall x \in \{x_1,\ldots,x_{k_\epsilon}\} : |\hat{\Phi}_{c^*,D}(x) - \Phi_{c^*,D}(x)| \leq \frac{\eta}{2}] \geq (1-\frac{\delta}{2}) \tag{4.5}$$

Now that we know how many times points with higher mass under D should be sampled during the sampling process, to get a good estimate of their $\Phi_{c^*,D}$ values, we need to compute a sample size such that any one of these points ($x \in \{x_1,\ldots,x_{k_\epsilon}\}$) occurs at least m_1 times in the sample with probability of at least $1-\frac{\delta}{2}$. Using Lemma 4.2 we need

$$m \geq \frac{m_1 n}{\epsilon} \ln(\frac{n}{1-(1-\frac{\delta}{2})^{\frac{1}{m_1}}}) \tag{4.6}$$

examples so that with probability at least $1-\frac{\delta}{2}$ every point $x \in \{x_1,\ldots,x_{k_\epsilon}\}$ is sampled at least m_1 times. This comes from the fact that in our problem $k \leq n$, $p \geq \frac{\epsilon}{n}$ and δ is replaced by $\frac{\delta}{2}$ in Lemma 4.2. Replacing m_1 in Equation 4.6 implies that

$$m \geq \frac{2n}{\epsilon\eta^2 \ln(1-(1-\frac{\delta}{2})^{\frac{1}{n}})} \ln(\frac{n}{1-(1-\frac{\delta}{2})^{\frac{\eta^2 \ln(1-(1-\frac{\delta}{2})^{\frac{1}{n}})}{2}}}) \tag{4.7}$$

examples are sufficient such that any point $x \in \{x_1,\ldots,x_{k_\epsilon}\}$ is at least sampled m_1 times in $\mathcal{S}$ with probability $1-\frac{\delta}{2}$.

The analysis of this proof consists of two phases. First we showed that with probability of at least $1-\frac{\delta}{2}$, drawing m_1 examples is sufficient to have an estimate of $\hat{\Phi}_{c^*,D}(x)$ that does not deviate more than $\frac{\eta}{2}$ from $\Phi_{c^*,D}(x)$ for any $x \in \{x_1,\ldots,x_{k_\epsilon}\}$. Then, we showed that drawing m examples is sufficient so that, with probability of at least $1-\frac{\delta}{2}$, any point $x \in \{x_1,\ldots,x_{k_\epsilon}\}$ is sampled m_1 times. Therefore, the probability of success in both phases of the algorithm is $(1-\frac{\delta}{2}) \times (1-\frac{\delta}{2}) \geq 1-\delta$.

Note that the error of c is at most ϵ with probability of at least $1-\delta$ because with probability of at least $1-\delta$, c predicts the labels of points $x \in \{x_1,..,x_{k_\epsilon}\}$ correctly and $\sum_{i=1}^{k_\epsilon} Pr_{x\sim D}[x=x_i] > 1-\epsilon$. □

In Proposition 4.3, the noise rate for any point in the input space only depends on the point itself. The following proposition deals with a situation where the noise rate for any point also depends on the underlying target concept. However, in this case, it is necessary to assume that for any point the noise rate is either less than or greater than a half for all concepts. Thus, it generalizes Proposition 4.3.

Proposition 4.4. *Let $\mathcal{X}$ be a finite input space, $\mathcal{C}$ a concept class and Φ a deterministic label noise model such that, for all $x \in \mathcal{X}$*

1. $\mathrm{nr}_{c^*,D}(x) \neq \frac{1}{2}$ *for all* $c^* \in \mathcal{C}$.
2. $[\mathrm{nr}_{c,D}(x) > \frac{1}{2} \Leftrightarrow \mathrm{nr}_{c',D}(x) > \frac{1}{2}]$ *for all* $c, c' \in \mathcal{C}$.

for all $D \in \mathcal{D}_\mathcal{X}$. Then $\mathcal{C}$ is $\mathrm{PAC}_{0,0}$-learnable with respect to Φ and $\mathcal{D}_\mathcal{X}$.

Sketch of the Proof. Let $c^* \in \mathcal{C}$ be the target concept and $D \in \mathcal{D}_\mathcal{X}$ a distribution. The idea of the proof is very similar to the one of Proposition 4.3. The learning algorithm estimates the values of $\Phi_{c^*,D}$ for points with higher mass under D and returns the concept whose noise rates are closest to $\mathrm{nr}_{c^*,D}$. The only difficulty may arise, because the learner does not know based on $\hat{\Phi}_{c^*,D}$ whether the noise rate is less than a half and the actual label is the more observed label or the noise rate is greater than a half and the actual label is the less often observed label. However, the second condition on Φ will prevent this situation from happening. It is not hard to show that the same number of examples as in the proof of Proposition 4.3 is sufficient here, too. □

Chapter 5

Minimum pn-disagreement Strategies

In this chapter we study a general learning algorithm with respect to noise (specifically label noise) called the minimum pn-disagreement strategy[1]. We consider the application of this strategy for learning in the presence of random classification noise. Also, we investigate the general characteristics of a label noise model that makes certain concept classes learnable with a minimum pn-disagreement strategy. In particular, we formulate a new sufficient condition for learnability with respect to label noise (Theorem 5.16).

First, we need to introduce some more notation. Also throughout this chapter we assume the label noise models are deterministic unless stated otherwise.

For a fixed target concept in a concept class and distribution, the pp-difference of a concept with respect to the target concept and the distribution is defined as follows.

Definition 5.1. *Let $\mathcal{C}$ be a concept class, D a distribution and $c^* \in \mathcal{C}$ the target concept. For any concept $c \in \mathcal{C}$, the $\Delta^{\mathrm{pp}}_{c^*,D}(c)$ between c and c^* is defined as*

$$\Delta^{\mathrm{pp}}_{c^*,D}(c) = E_{x\sim D}[|c(x) - c^*(x)|] \tag{5.1}$$

For the sake of convenience, we simply write $\Delta^{\mathrm{pp}}(c)$ instead of $\Delta^{\mathrm{pp}}_{c^*,D}(c)$ whenever c^* and D are clear from the context[2].

The pp-difference of a concept, c, has been considered as the error of c in the literature [2, 25] because for a large enough sample from a noise-free oracle, it is expected that c has error in predicting a fraction of $\Delta^{\mathrm{pp}}(c)$ of the instances. We can easily write Δ^{pp} in terms of probability as we did in the previous chapters.

Remark 5.2. *Let $\mathcal{C}$ be a concept class and D a distribution. Let $c^* \in \mathcal{C}$ be the target concept. Then*

$$\Delta^{\mathrm{pp}}(c) = Pr_{x\sim D}[c(x) \neq c^*(x)] \tag{5.2}$$

for all $c \in \mathcal{C}$.

For a fixed error parameter, a fixed distribution, and a fixed target concept, we can divide the concepts in a concept class into two disjoint groups.

Definition 5.3. ***(Angluin and Laird [2])*** *Let $\mathcal{C}$ be a concept class and D a distribution. Let $c^* \in \mathcal{C}$ be the target concept and $\epsilon \in (0, 1)$. A concept $c \in \mathcal{C}$ is if $\Delta^{\mathrm{pp}}(c) \leq \epsilon$. c is if $\Delta^{\mathrm{pp}}(c) > \epsilon$.*

[1] The term "strategy" in this chapter is referring to what we call a "learning algorithm" throughout this thesis.

[2] The designation pp is used here to show that the disagreement is measured between two "pure" concepts in the sense that both of these two concepts are not noisy and, therefore, are mappings from the input space to $\{0, 1\}$.

Since any target concept, c^*, has $\Delta^{\mathrm{pp}}(c^*)$ equal to zero, the target concept is always ϵ-good regardless of the value of ϵ. However, when we are dealing with a noisy oracle there may be no concept that has zero disagreement with the examples returned by a noisy oracle. But first we need to show how we can measure the disagreement between a concept and the target concept in the presence of noise.

Definition 5.4. *Let $\mathcal{C}$ be a concept class and D a distribution. Let $c^* \in \mathcal{C}$ be the target concept and Φ a deterministic noise model. For any concept $c \in C$, the $\Delta^{\mathrm{pn}}_{c^*,D,\Phi}(c)$ between c and c^* is defined as*

$$\Delta^{\mathrm{pn}}_{c^*,D,\Phi}(c) = E_{x\sim D}[|c(x) - \Phi_{c^*,D}(x)|] \tag{5.3}$$

For the sake of convenience, we write $\Delta^{\mathrm{pn}}(c)$ instead of $\Delta^{\mathrm{pn}}_{c^*,D,\Phi}(c)$ whenever c^*, D and Φ are clear from the context[3]. As in the case of pp-disagreement we can write pn-disagreement in terms of probability instead of expectation. But first we need the following definition.

Definition 5.5. *Let c be a concept and D a distribution. Let Φ be a deterministic noise model. We denote by $\hat{c}(x)$ the random variable defined as*

$$\hat{c}(x) = \begin{cases} 1 & \textit{with probability } \Phi_{c,D}(x) \\ 0 & \textit{otherwise} \end{cases} \tag{5.4}$$

Therefore, we can write the pn-disagreement of a concept c as the probability of disagreement between the labels $c(x)$ and a label drawn at random from $\hat{c}^*(x)$ where x is drawn at random with respect to the underlying distribution. The following proposition shows that this gives us the same outcome as Definition 5.4.

Proposition 5.6. *Let $\mathcal{C}$ be a concept class and D a distribution. Let $c^* \in \mathcal{C}$ be the target concept and Φ a noise model. Then*

$$\Delta^{\mathrm{pn}}_{c^*,D,\Phi}(c) = Pr_{x\sim D, y\sim\hat{c}^*(x)}[c(x) \neq y] \tag{5.5}$$

for all $c \in \mathcal{C}$.

Proof. We can break down the expectation based on whether $c(x) = 0$ or $c(x) = 1$ and then we can use Definition 5.5 to replace the $\Phi_{c^*,D}$ values.

$$\begin{aligned}
E_{x\sim D}[|c(x) - \Phi_{c^*,D}(x)|] &= \int_x |c(x) - \Phi_{c^*,D}(x)|dx \\
&= \int_{x|c(x)=0} \Phi_{c^*,D}(x)dx + \int_{x|c(x)=1} (1 - \Phi_{c^*,D}(x))dx \\
&= \int_{x|c(x)=0} Pr_{y\sim\hat{c}^*(x)}[y = 1]dx + \int_{x|c(x)=1} Pr_{y\sim\hat{c}^*(x)}[y = 0]dx \\
&= \int_{x|c(x)=0} Pr_{y\sim\hat{c}^*(x)}[c(x) \neq y]dx + \int_{x|c(x)=1} Pr_{y\sim\hat{c}^*(x)}[c(x) \neq y]dx \\
&= Pr_{x\sim D, y\sim\hat{c}^*(x)}[c(x) \neq y].
\end{aligned}$$

□

The following example shows how Δ^{pp} and Δ^{pn} can be computed.

Example 5.7. *Let $\mathcal{X} = \{x_1, x_2\}$, $\mathcal{C} = \{c_1, c_2, c_3\}$ the concept class described in Table 5.1 and $D \in \mathcal{D}_{\mathcal{X}}$ a distribution. Let $Pr_{x\sim D}[x = x_1] = 0.25$ and $Pr_{x\sim D}[x = x_2] = 0.75$. Let Φ be a deterministic label noise model with $\mathrm{nr}_{c^*,D}(x_1) = 0.75$ and $\mathrm{nr}_{c^*,D}(x_2) = 0.25$ for any $c^* \in \mathcal{C}$ (see Table 5.1).*

concept/label	x_1	x_2	concept/value	x_1	x_2
c_1	1	1	$\Phi_{c_1,D}$	0.25	0.75
c_2	0	1	$\Phi_{c_2,D}$	0.75	0.75
c_3	1	0	$\Phi_{c_3,D}$	0.25	0.25

Table 5.1: Concept class in Example 5.7

concept	Δ^{pp}	Δ^{pn}
c_1	$0.25\|1-0\|+0.75\|1-1\|=0.25$	$0.25\|0.25-0\|+0.75\|1-0.75\|=0.25$
$c^*=c_2$	$0.25\|0-0\|+0.75\|1-1\|=0$	$0.25\|0.75-0\|+0.75\|1-0.75\|=0.31$
c_3	$0.25\|1-0\|+0.75\|0-1\|=1$	$0.25\|0.25-0\|+0.75\|1-0.25\|=0.69$

Table 5.2: Δ^{pn} values in Example 5.7

We can use Definitions 5.1 and 5.4 to compute the Δ^{pp} and the Δ^{pn} values respectively for the case that c_2 is the target concept as in Table 5.2.

We can estimate the Δ^{pn} values with sampling. We call this estimation the pn-disagreement between a concept and the sample.

Definition 5.8. ***(Angluin and Laird [2])*** *Let c be a concept. Let $\mathcal{S}=\{(x_1,y_1),\ldots,(x_m,y_m)\}$ be a sample of size $m\in\mathbb{N}$. The $\mathcal{F}_{\mathrm{pn}}(c,\mathcal{S})$ between c and $\mathcal{S}$ is defined as follows*

$$\mathcal{F}_{\mathrm{pn}}(c,\mathcal{S})=\frac{1}{m}\sum_{i=1}^{m}|y_i-c(x_i)| \tag{5.6}$$

$\mathcal{F}_{\mathrm{pn}}(c,\mathcal{S})$ is an unbiased[4] estimator of Δ^{pn}. The minimum pn-disagreement strategy simply returns the concept that has the smallest pn-disagreement with the sample. We adapt the definition of this strategy from Angluin and Laird [2].

Definition 5.9. ***(Angluin and Laird [2])*** *Let $\mathcal{C}$ be a concept class and $\mathcal{S}$ a sample. A learning algorithm $\mathcal{L}$ is called a for $\mathcal{C}$ if upon seeing $\mathcal{S}$, $\mathcal{L}$ returns a (not necessarily unique) concept $\bar{c}\in\mathcal{C}$ such that*

$$\bar{c}\in \textit{arg min}_{c\in\mathcal{C}}\,\mathcal{F}_{\mathrm{pn}}(c,\mathcal{S}) \tag{5.7}$$

The minimum pn-disagreement strategy has also been called in the machine learning literature (see *e.g.,* [39]).

In the absence of noise, Δ^{pp} and Δ^{pn} values are equal for any pair of concept and target concept in the concept class. Thus, Δ^{pp} can be estimated using $\mathcal{F}_{\mathrm{pn}}$ when there is no noise. Since Δ^{pp} of the target concept is always zero in the noise-free case, any learning algorithm that upon sampling returns a concept that is consistent (has pn-disagreement of 0) with the sample is also a minimum pn-disagreement strategy.

5.1 Minimum pn-disagreement strategies for learning with random classification noise

In this section we consider the applications of minimum pn-disagreement strategies for learning concept classes with respect to the most benign label noise model, random classification noise.

[3]The designation pn is used here to show that the disagreement is measured between a "pure" concept and a "noisy" one. A concept is noisy in the sense that it is a mapping from the input space to $[0,1]$ resulting from applying a noise model when sampling from a "pure" concept.

[4]$\lim_{|\mathcal{S}|\to\infty}E[\mathcal{F}_{\mathrm{pn}}(c,\mathcal{S})]=\Delta^{\mathrm{pn}}(c)$

This will provide us some insight for the next section where we are seeking general features of a noise model in which learning with respect to noise is possible with the minimum pn-disagreement strategy.

In the random classification noise model, pn-disagreement values have a linear relationship with the corresponding pp-disagreement values. We state the following lemma and its proof from the article by Angluin and Laird [2].

Lemma 5.10. ***(Angluin and Laird [2])*** *Let $\eta \in [0, 1)$. Let $\mathcal{C}$ be a concept class and D a distribution. For any concept $c \in \mathcal{C}$ in the presence of η-random classification noise, $\Phi_{\mathrm{rcn}(\eta)}$,*

$$\Delta^{\mathrm{pn}}(c) = \eta + (1 - 2\eta)\Delta^{\mathrm{pp}}(c) \tag{5.8}$$

Proof. Let $c^* \in \mathcal{C}$ be the target concept. For any example (x, y) drawn from $\mathrm{EX}_{\mathrm{rcn}(\eta)}(c^*, D)$, the probability that the label y of the example disagrees with $c(x)$ is equal to the probability that (x, y) is drawn from the area in which $c(x) \neq c^*(x)$ and reported correctly by $\mathrm{EX}_{\mathrm{rcn}(\eta)}(c^*, D)$ (which happens with probability of $1 - \eta$) plus the probability that (x, y) is drawn from the area in which $c(x) = c^*(x)$ and reported incorrectly by $\mathrm{EX}_{\mathrm{rcn}(\eta)}(c^*, D)$ (which happens with probability of η). Therefore,

$$\Delta^{\mathrm{pn}}(c) = (1 - \eta)\Delta^{\mathrm{pp}}(c) + \eta(1 - \Delta^{\mathrm{pp}}(c)) = \eta + (1 - 2\eta)\Delta^{\mathrm{pp}}(c)$$

□

Next, we show that Theorem 2.14 can be proved using minimum pn-disagreement strategies. This proof is adapted from the article by Angluin and Laird [2]. As a reminder, the statement of the theorem is as follows.

> Let $\mathcal{C}$ be a finite concept class. Let $0 \leq \eta < \frac{1}{2}$. Then $\mathcal{C}$ is $\mathrm{PAC}_{0,0}$-learnable with respect to η-random classification noise and $\mathcal{D}_{\mathcal{X}}$.

Proof of Theorem 2.14. (Angluin and Laird [2]) Let $\epsilon, \delta \in (0, \frac{1}{2})$. Let $c^* \in \mathcal{C}$ be the target concept and $D \in \mathcal{D}_{\mathcal{X}}$ a distribution. We will show that if $\mathcal{L}$ draws a sample $\mathcal{S}$ of

$$m \geq \frac{2}{\epsilon^2(1 - 2\eta)^2} \ln(\frac{2n}{\delta})$$

examples from $\mathrm{EX}_{\mathrm{rcn}(\eta)}(c^*, D)$ and returns a (not necessarily unique) concept $\bar{c} \in \mathcal{C}$ such that

$$\bar{c} \in \arg\min_{c \in \mathcal{C}} \mathcal{F}_{\mathrm{pn}}(c, \mathcal{S})$$

then $\bar{c}$ has error of at most ϵ with probability at least $1 - \delta$.

Using Equation 5.8 in Lemma 5.10, for any concept $c \in \mathcal{C}$

$$\Delta^{\mathrm{pn}}(c) = \eta + (1 - 2\eta)\Delta^{\mathrm{pp}}(c)$$

Since $\eta < \frac{1}{2}$, for any $c \in \mathcal{C}$, $\Delta^{\mathrm{pn}}(c) \geq \eta$ and specifically for c^*, $\Delta^{\mathrm{pn}}(c^*) = \eta$ because $\Delta^{\mathrm{pp}}(c^*) = 0$. Also for any ϵ-bad concept $c \in \mathcal{C}$

$$\Delta^{\mathrm{pn}}(c) = \eta + (1 - 2\eta)\Delta^{\mathrm{pp}}(c) > \eta + (1 - 2\eta)\epsilon$$

As a result there is a separation of at least $(1 - 2\eta)\epsilon$ between the Δ^{pn} value of the target concept and the Δ^{pn} value of any ϵ-bad concept. For the rest of this proof, we assume any occurrence of c refers to an ϵ-bad concept.

We now show that m examples are sufficient so that, with probability of at least $1 - \delta$, no ϵ-bad concept minimizes the $\mathcal{F}_{\mathrm{pn}}$ value. In order for some ϵ-bad concept c to minimize $\mathcal{F}_{\mathrm{pn}}$ on a randomly drawn sample of size m, at least one of the following inequalities has to be fulfilled.

$$\mathcal{F}_{\mathrm{pn}}(c, \mathcal{S}) \leq \eta + \frac{(1 - 2\eta)\epsilon}{2} \tag{5.9}$$

$$\mathcal{F}_{\mathrm{pn}}(c^*, \mathcal{S}) \geq \eta + \frac{(1-2\eta)\epsilon}{2} \tag{5.10}$$

Otherwise an ϵ-good concept will minimize $\mathcal{F}_{\mathrm{pn}}$. We bound the probability that either Equation 5.9 or Equation 5.10 happens by at most δ. First let us start with Equation 5.9.

$$\begin{aligned} Pr_{x\sim D}[\mathcal{F}_{\mathrm{pn}}(c, \mathcal{S}) \leq & \eta + \frac{(1-2\eta)\epsilon}{2}] < \mathrm{LE}(\eta + (1-2\eta)\epsilon, m, \eta + \frac{(1-2\eta)\epsilon}{2}) \\ & = \mathrm{LE}(\eta + (1-2\eta)\epsilon, m, \eta + (1-2\eta)\epsilon - \frac{(1-2\eta)\epsilon}{2}) \end{aligned}$$

where LE is the function defined in Definition A.2 in the Appendix. Since $m \geq \frac{2}{\epsilon^2(1-2\eta)^2}\ln(\frac{2n}{\delta})$, by applying Lemma A.4 from the Appendix, with $s = \frac{\epsilon(1-2\eta)}{2}$ and replacing δ by $\frac{\delta}{2n}$,

$$\begin{aligned} Pr_{x\sim D}[\mathcal{F}_{\mathrm{pn}}(c, \mathcal{S}) \leq & \eta + \frac{(1-2\eta)\epsilon}{2}] \\ & < \mathrm{LE}(\eta + (1-2\eta)\epsilon, m, \eta + (1-2\eta)\epsilon - \frac{(1-2\eta)\epsilon}{2}) \\ & \leq \frac{\delta}{2n} \end{aligned}$$

Therefore, the probability that there exists an ϵ-bad concept that fulfills $\mathcal{F}_{\mathrm{pn}}(c, \mathcal{S}) \leq \eta + \frac{(1-2\eta)\epsilon}{2}$ is at most $(n-1)\frac{\delta}{2n} < \frac{\delta}{2}$ because the number of ϵ-bad concepts in $\mathcal{C}$ is at most $n-1$ (since c^*, which is ϵ-good, is assumed to be in $\mathcal{C}$).

Now, we consider Equation 5.10. By applying Lemma A.4 from the Appendix

$$Pr_{x\sim D}[\mathcal{F}_{\mathrm{pn}}(c^*, \mathcal{S}) \geq \eta + \frac{(1-2\eta)\epsilon}{2}] \leq \mathrm{GE}(\eta, m, \eta + \frac{(1-2\eta)\epsilon}{2}))$$

where GE is the function defined in Definition A.2 in the Appendix. Since $m \geq \frac{2}{\epsilon^2(1-2\eta)^2}\ln(\frac{2n}{\delta})$, by applying Lemma A.4 from the Appendix, with $s = \frac{\epsilon(1-2\eta)}{2}$ and replacing δ by $\frac{\delta}{2n}$,

$$\begin{aligned} \mathrm{GE}(\eta, m, \eta + \frac{(1-2\eta)\epsilon}{2})) & \leq \frac{\delta}{2n} \\ & \leq \frac{\delta}{2} \end{aligned}$$

Therefore, the probability that an ϵ-bad concept minimizes $\mathcal{F}_{\mathrm{pn}}$ is less than $\frac{\delta}{2} + \frac{\delta}{2} = \delta$. So with probability at least $1-\delta$, $\bar{c}$ has an error of at most ϵ. □

The same argument as in the proof of Theorem 2.14 cannot be applied when the concept class is not finite because then the sample size, m, becomes infinite. In the rest of this section, we show how Laird solved this problem. First, we show any subset of the input space can divide the concept class into finitely many equivalence classes.

Definition 5.11. ***(Laird [27])*** *Let $X \subseteq \mathcal{X}$. Any concept $c \in \mathcal{C}$ splits X into two disjoint sets $c \cap X$ and $X - c$. Two concepts c_1, c_2 are X- if they split X into the same sets. Any set of all pairwise X-equivalent concepts in $\mathcal{C}$ is called an X-.*

For the sake of convenience, we write equivalent and equivalence instead of X-equivalent and X-equivalence when X is clear from the context. The following lemma by Laird shows the relationship between the number of equivalence classes and the size of X.

Lemma 5.12. ***(Laird [27])*** *Let $X \subseteq \mathcal{X}$ and $\mathcal{C}$ a concept class of* VC*-dimension $d \in \mathbb{N}$. For any X with at least d elements, X divides $\mathcal{C}$ into at most $|X|^d + 1$ equivalence classes.*

For any pair of concept and target concept, the input space can be divided into two disjoint sets: the set of all the points that the concept labels the same as the target concept and the set of all the points that the concept labels differently from the target concept. The following theorem, by Blumer, Ehrenfeucht, Haussler and Warmuth [7], states that in a large enough sample, with high probability, at least one instance is drawn from the latter set for all ϵ-bad concepts.

Theorem 5.13. ***(Blumer, Ehrenfeucht, Haussler and Warmuth [7])*** *Let $\mathcal{C}$ be a concept class of* VC*-dimension of at most $d < \infty$, D a distribution and $c^* \in \mathcal{C}$ the target concept. Let $\epsilon, \delta \in (0, \frac{1}{2})$. Let $\mathcal{X}_c = \{x | c(x) \neq c^*(x)\}$ for $c \in \mathcal{C}$. Let $m_1 : [0,1] \times [0,1] \times \mathbb{N} \to \mathbb{N}$ be the function defined as follows.*

$$m_1(\epsilon, \delta, d) = \lceil max\ (\frac{4}{\epsilon} \log \frac{2}{\delta}, \frac{8d}{\epsilon} \log \frac{8d}{\epsilon}) \rceil \tag{5.11}$$

Then a set of $m_1(\epsilon, \delta, d)$ instances in $\mathcal{X}$, drawn $i.i.d.$ from D, will have a non-empty intersection with $\mathcal{X}_c$ for each ϵ-bad concept c, with probability at least $1 - \delta$.

Theorem 5.13 with Lemma 5.12 can be used to prove that any concept class of finite VC-dimension is PAC-learnable in the noise-free setting (see Theorem 2.5 for more details). The high level idea of the proof of Theorem 2.5 can be described as follows. $m_1(\epsilon, \delta, d)$ examples (m_1, for short) divide $\mathcal{C}$ into at most ${m_1}^d + 1$ equivalence classes (using Lemma 5.12). So although $\mathcal{C}$ may have an infinite number of concepts, there is a polynomial set of equivalence classes[5] of concepts given a finite sample. Then, using Theorem 5.13, for any ϵ-bad concept $c \in \mathcal{C}$, m_1 examples are sufficient so that with probability of at least $1 - \delta$, at least one example is sampled from the area in which c disagrees with the target concept. Therefore, with a sample $\mathcal{S}$ of size m_1, with probability of at least $1 - \delta$, all equivalence classes that contain an ϵ-bad concept will disagree with the target concept on at least one example from $\mathcal{S}$. Thus, all the equivalence classes that are consistent with the sample do not contain any ϵ-bad concepts. So it is sufficient for the learning algorithm to choose an arbitrary concept from any of those equivalence classes. Based on the argument above, any such concept is ϵ-good with probability of at least $1 - \delta$. The reader is referred to the article by Blumer *et al.* [7] for more details.

Although the proof is for a noise-free setting, the idea of the proof can be used in the noisy setting as well. In the presence of noise, there may be no concept that is consistent with the sample. But in the case of random classification noise, as we showed in the proof of Theorem 2.14, the target concept has the smallest Δ^{pn} value among all the concepts in the concept class, specifically, smaller than the Δ^{pn} value of any the ϵ-bad concepts.

We next show how Laird proved Theorem 2.15. In the proof of the theorem, he showed that a sufficiently large sample size will guarantee that with high probability no ϵ-bad concept will have a smaller pn-disagreement with the sample than the target concept. As a reminder, the statement of Theorem 2.15 is as follows.

> Let $\mathcal{C}$ be a concept class of VC-dimension $d < \infty$. Let $\eta \in [0, \frac{1}{2})$. Then $\mathcal{C}$ is $\mathrm{PAC}_{0,0}$-learnable with respect to η-random classification noise and $\mathcal{D}_{\mathcal{X}}$.

Proof of Theorem 2.15. (Laird [27]) Let $\epsilon, \delta \in (0, \frac{1}{2})$. Let $c^* \in \mathcal{C}$ be the target concept and $D \in \mathcal{D}_{\mathcal{X}}$ a distribution. The learner $\mathcal{L}$ first draws a sample $\mathcal{S}_1$ of size $m_1(\frac{\epsilon}{2}, \frac{\delta}{3}, d)$ (m_1, for short) from $\mathrm{EX}_{\mathrm{rcn}(\eta)}(c^*, D)$, where m_1 is the function in Equation 5.11. Using Lemma 5.12, $\mathcal{S}_1$ divides $\mathcal{C}$ into N equivalence classes where $N \leq m_1^d + 1$. Let $\mathcal{C}_N = \{c_1, \ldots, c_N\}$ be such that each concept $c_i \in \mathcal{C}$, for $1 \leq i \leq N$, belongs to a different equivalence class. These N concepts can be used as representatives for the N equivalence classes.

Theorem 5.13 will guarantee that, with probability of at least $1 - \frac{\delta}{3}$, there is at least one equivalence class that contains only $\frac{\epsilon}{2}$-good concepts because no $\frac{\epsilon}{2}$-bad concept can be in the same equivalence class as the target concept. Therefore, at least the equivalence class that contains the target

[5] This may sound incorrect at first glance, because X in Definition 5.11 and Lemma 5.12 is a subset of unlabeled instances of $\mathcal{X}$ and the sample $\mathcal{S}$ is a multi-set of (instance, label) pairs. Laird [27] shows that the same argument holds when the learner just considers the instance part of the sample $\mathcal{S}$ and ignores all the labels.

concept only contains $\frac{\epsilon}{2}$-good concepts. Therefore,

$$\text{with probability at least } 1-\frac{\delta}{3}\text{, there is at least one } \frac{\epsilon}{2}\text{-good concept in } \mathcal{C}_N. \tag{5.12}$$

In the rest of this proof we show that using a large enough sample, with high probability, the concept in $\mathcal{C}_N$ that minimizes $\mathcal{F}_{\text{pn}}$ is ϵ-good. This part of the proof is very similar to what we previously showed in the proof of Theorem 2.14. We, however, repeat in order to avoid references to that proof.

Let $m_2 : [0,1] \times [0,1] \times [0,1] \times \mathbb{N} \to \mathbb{N}$ be a function defined as follows.

$$m_2(\epsilon, \delta, \eta, N) = \lceil \frac{8}{\epsilon^2(1-2\eta)^2} \ln(\frac{3N}{\delta}) \rceil \tag{5.13}$$

Let $\mathcal{S}_2$ be a sample containing $\mathcal{S}_1$ and enough additional examples (if needed) so that it contains at least $m_2(\epsilon, \delta, \eta, N)$ examples (m_2, for short). Therefore, the total number of examples needed for $\mathcal{L}$ is $\max(m_1, m_2)$. Let $\mathcal{L}$ return a concept that has the minimum $\mathcal{F}_{\text{pn}}$ with $\mathcal{S}_2$.

Using Lemma 5.10, under random classification noise, for any ϵ-bad concept $c \in \mathcal{C}$, $\Delta^{pn}(c) = \eta + \Delta^{pp}(c)(1-2\eta) > \eta + \epsilon(1-2\eta)$. On the other hand, for any $\frac{\epsilon}{2}$-good concept $c' \in \mathcal{C}$, $\Delta^{pn}(c') = \eta + \Delta^{pp}(c')(1-2\eta) \leq \eta + \frac{\epsilon}{2}(1-2\eta)$. Therefore, there is a separation of at least $(\eta + \epsilon(1-2\eta)) - (\eta + \frac{\epsilon}{2}(1-2\eta)) = \frac{\epsilon}{2}(1-\eta)$ between the Δ^{pn} value of any $\frac{\epsilon}{2}$-good concept and the Δ^{pn} value of any ϵ-bad concept. For the rest of this proof, we assume any occurrence of c and c' refers to ϵ-bad and $\frac{\epsilon}{2}$-good concepts respectively.

In order for some ϵ-bad concept in $\mathcal{C}_N$ to minimize $\mathcal{F}_{\text{pn}}$, at least one of the following inequalities would have to be fulfilled. For at least one ϵ-bad concept in $\mathcal{C}_N$

$$\mathcal{F}_{\text{pn}}(c, \mathcal{S}) \leq \eta + \frac{3(1-2\eta)\epsilon}{4} \tag{5.14}$$

or for all $\frac{\epsilon}{2}$-good concepts in $\mathcal{C}_N$

$$\mathcal{F}_{\text{pn}}(c', \mathcal{S}) \geq \eta + \frac{3(1-2\eta)\epsilon}{4} \tag{5.15}$$

Otherwise an ϵ-good concept minimizes $\mathcal{F}_{\text{pn}}$. Therefore, $\mathcal{L}$ fails only if any of the following cases happens.

1. There is no $\frac{\epsilon}{2}$-good concept in $\mathcal{C}_N$. This will happen with probability of at most $\frac{\delta}{3}$ because of 5.12.

2. Considering Equation 5.14, for at least one ϵ-bad representative, $\mathcal{F}_{\text{pn}}(c, \mathcal{S}_2)$ is at most $\eta + \frac{3\epsilon}{4}(1-2\eta)$. Applying Lemma A.4 from the Appendix

$$\begin{aligned} Pr_{x\sim D}[\mathcal{F}_{\text{pn}}(c, \mathcal{S}_2) \leq \eta + \frac{3\epsilon(1-2\eta)}{4}] & \\ &< \text{LE}(\eta + (1-2\eta)\epsilon, \max(m_1, m_2), \eta + \frac{3(1-2\eta)\epsilon}{4}) \\ &\leq \frac{\delta}{3N} \end{aligned}$$

   Therefore, the probability that there exists an ϵ-bad concept in $\mathcal{C}_N$ that satisfies Equation 5.14 is at most $(N-1)\frac{\delta}{3N} < \frac{\delta}{3}$ because the number of ϵ-bad concepts is at most $N-1$. (There is at least one $\frac{\epsilon}{2}$-good concept in $\mathcal{C}_N$ with probability of at least $1-\frac{\delta}{3}$.)

3. Considering Equation 5.15, all of the $\frac{\epsilon}{2}$-good representatives have $\mathcal{F}_{\text{pn}}(c', \mathcal{S}_2)$ of at least $\eta +$

$\frac{3(1-2\eta)\epsilon}{4}$. Again by applying Lemma A.4 from the Appendix

$$\begin{aligned} Pr_{x\sim D}[\mathcal{F}_{\mathrm{pn}}(c', \mathcal{S}_2) \geq & \eta + \frac{3(1-2\eta)\epsilon}{4}] \\ & \leq \mathrm{GE}(\eta + \frac{(1-2\eta)\epsilon}{2}, \max(m_1, m_2), \eta + \frac{3(1-2\eta)\epsilon}{4}) \\ & \leq \frac{\delta}{3N} \end{aligned}$$

Therefore, the probability that all of the $\frac{\epsilon}{2}$-good concepts in $\mathcal{C}_N$ satisfy Equation 5.15 is at most $(\frac{\delta}{3N})^N \leq \frac{\delta}{3N} \leq \frac{\delta}{3}$ because the number of $\frac{\epsilon}{2}$-good concepts is at most N.

So the total probability of the failure of the algorithm is at most $3 \times \frac{\delta}{3} = \delta$. Therefore, with probability of at least $1-\delta$, the concept returned by $\mathcal{L}$ has error of at most ϵ. □

5.2 Sufficient conditions for learning with minimum pn-disagreement strategies

In this section, we investigate the properties of a label noise model that guarantee the minimum pn-disagreement strategy can be used to learn a concept class with respect to the noise model. pn-unambiguity is the first such property, defined as follows.

Definition 5.14. *Let $\mathcal{C}$ be a concept class, $\mathcal{D} \subseteq \mathcal{D}_{\mathcal{X}}$ a class of distributions and Φ a deterministic noise model. Φ is* with respect to $\mathcal{C}$ and $\mathcal{D}$ *if there exists a function $g : (0, \frac{1}{2}) \to (0, 1)$ such that for any target concept $c^* \in \mathcal{C}$, for any distribution $D \in \mathcal{D}$, for any $\epsilon \in (0, \frac{1}{2})$ and for any pair of concepts $c, c' \in \mathcal{C}$*

$$\Delta^{\mathrm{pp}}(c') - \Delta^{\mathrm{pp}}(c) > \epsilon \Rightarrow \Delta^{\mathrm{pn}}(c') - \Delta^{\mathrm{pn}}(c) \geq g(\epsilon). \tag{5.16}$$

Otherwise Φ is with respect to $\mathcal{C}$ and $\mathcal{D}$.

Proposition 5.15 shows that η-random classification noise is pn-unambiguous with respect to any concept class and $\mathcal{D}_{\mathcal{X}}$ when $\eta < \frac{1}{2}$.

Proposition 5.15. *Let $\mathcal{C}$ be a concept class. The η-random classification noise model, $\Phi_{\mathrm{rcn}(\eta)}$, is pn-unambiguous with respect to $\mathcal{C}$ and $\mathcal{D}_{\mathcal{X}}$ iff $\eta < \frac{1}{2}$.*

Proof. As we previously showed in Equation 5.8 of Lemma 5.10, for any concept $c \in \mathcal{C}$, under random classification noise

$$\Delta^{\mathrm{pn}}(c) = \eta + (1-2\eta)\Delta^{\mathrm{pp}}(c)$$

Let $\epsilon \in (0, \frac{1}{2})$ and $c, c' \in \mathcal{C}$ such that $\Delta^{\mathrm{pp}}(c') - \Delta^{\mathrm{pp}}(c) > \epsilon$. When $\eta < \frac{1}{2}$, $(1-2\eta) > 0$. Therefore,

$$\begin{aligned} \Delta^{\mathrm{pn}}(c') - \Delta^{\mathrm{pn}}(c) &= \eta + (1-2\eta)\Delta^{\mathrm{pp}}(c') - (\eta + (1-2\eta)\Delta^{\mathrm{pp}}(c)) \\ &= (1-2\eta)(\Delta^{\mathrm{pp}}(c') - \Delta^{\mathrm{pp}}(c)) > \epsilon(1-2\eta) > 0 \end{aligned}$$

The function g defined by $g(\epsilon) = \epsilon(1-2\eta)$ witnesses that Φ is pn-unambiguous. When $\eta \geq \frac{1}{2}$, $(1-2\eta) \leq 0$. Therefore,

$$\Delta^{\mathrm{pn}}(c') - \Delta^{\mathrm{pn}}(c) = (1-2\eta)(\Delta^{\mathrm{pp}}(c') - \Delta^{\mathrm{pp}}(c)) \leq 0$$

Thus no such function g as defined in Definition 5.14 exists and, therefore, Φ is pn-ambiguous. □

pn-unambiguity of any noise model with respect to a concept class of finite VC-dimension and a class of distributions is a sufficient condition for $\mathrm{PAC}_{0,0}$-learnability of that concept class with respect to the noise model and the class of distributions. This is stated in the following theorem, the main result of this section.

Theorem 5.16. *Let $\mathcal{C}$ be a concept class of* VC*-dimension $d < \infty$ and Φ a deterministic noise model. If Φ is* pn*-unambiguous with respect to $\mathcal{C}$ and $\mathcal{D}_{\mathcal{X}}$ then $\mathcal{C}$ is* $\mathrm{PAC}_{0,0}$*-learnable with respect to Φ and $\mathcal{D}_{\mathcal{X}}$.*

Proof. Let $\epsilon, \delta \in (0, \frac{1}{2})$. Let $c^* \in \mathcal{C}$ be the target concept , $D \in \mathcal{D}_{\mathcal{X}}$ a distribution and $\mathcal{L}$ a minimum pn-disagreement strategy. The total number of examples drawn by the learner, $\mathcal{L}$, is determined in exactly the same way as in the proof of Theorem 2.15. We briefly repeat the explanation for ease of reference.

As in the proof of Theorem 2.15 in Section 5.1, $\mathcal{L}$ first draws a sample $\mathcal{S}_1$ of size $m_1(\frac{\epsilon}{2}, \frac{\delta}{3}, d)$ (m_1, for short) to find a set of $N \leq {m_1}^d + 1$ representative concepts $\mathcal{C}_N = \{c_1, \ldots, c_N\}$ of N equivalence classes, among which at least one is $\frac{\epsilon}{2}$-good with probability of at least $1 - \frac{\delta}{3}$.

In the rest of this proof, we show that a large enough sample guarantees that, with high probability, the concept in $\mathcal{C}_N$ that minimizes $\mathcal{F}_{\mathrm{pn}}$ is $\frac{\epsilon}{2}$-good. The idea of this part is very similar to the corresponding part in the proof of Theorem 2.15.

Let $m_2 : [0,1] \times [0,1] \times \mathbb{N} \to \mathbb{N}$ be a function defined as follows.

$$m_2(\epsilon, \delta, N) = \lceil \frac{2}{g(\frac{\epsilon}{2})^2} \ln(\frac{3N}{\delta}) \rceil$$

Let $\mathcal{S}_2$ be a sample that contains $\mathcal{S}_1$ and enough additional examples (if needed) so that it contains at least $m_2(\epsilon, \delta, N)$ (m_2, for short) examples. Therefore, the total number of examples that $\mathcal{L}$ draws from the noisy oracle is $\max(m_1, m_2)$.

Since Φ is pn-unambiguous with respect to $\mathcal{C}$ and $\mathcal{D}_{\mathcal{X}}$, there exists a function g such that for any ϵ and for any pair of concepts $c, c' \in \mathcal{C}$ with $\Delta^{\mathrm{pp}}(c') - \Delta^{\mathrm{pp}}(c) > \frac{\epsilon}{2}$, $\Delta^{\mathrm{pn}}(c') - \Delta^{\mathrm{pn}}(c) \geq g(\frac{\epsilon}{2})$. Therefore, there is a separation of at least $g(\frac{\epsilon}{2})$ between the Δ^{pn} values of any pair of concepts that have at least a difference of $\frac{\epsilon}{2}$ between their Δ^{pp} values. For the rest of this proof, any occurrence of c' and c refers to ϵ-bad and $\frac{\epsilon}{2}$-good concepts respectively.

In order for some ϵ-bad concept in $\mathcal{C}_N$ to minimize $\mathcal{F}_{\mathrm{pn}}$ at least one of the following inequalities would have to be fulfilled. For all the $\frac{\epsilon}{2}$-good concepts in $\mathcal{C}_N$

$$\mathcal{F}_{\mathrm{pn}}(c, \mathcal{S}) \geq \Delta^{\mathrm{pn}}(c) + \frac{g(\frac{\epsilon}{2})}{2} \tag{5.17}$$

or for at least one ϵ-bad concept in $\mathcal{C}_N$

$$\mathcal{F}_{\mathrm{pn}}(c', \mathcal{S}) \leq \Delta^{\mathrm{pn}}(c) + \frac{g(\frac{\epsilon}{2})}{2} \tag{5.18}$$

because otherwise an ϵ-good concept minimizes $\mathcal{F}_{\mathrm{pn}}$. Therefore, $\mathcal{L}$ fails only if any of the following cases happens.

1. There is no $\frac{\epsilon}{2}$-good concept in $\mathcal{C}_N$. As in the proof of Theorem 2.15, this will happen with probability of at most $\frac{\delta}{3}$.

2. Considering Equation 5.17, all $\frac{\epsilon}{2}$-good concepts in $\mathcal{C}_N$ have $\mathcal{F}_{\mathrm{pn}}(c, \mathcal{S}_2)$ of at least $\Delta^{\mathrm{pn}}(c) + \frac{g(\epsilon)}{2}$. Applying Lemma A.4 from the Appendix

$$\begin{aligned} Pr_{x \sim D}[\mathcal{F}_{\mathrm{pn}}(c, \mathcal{S}_2) \geq & \Delta^{\mathrm{pn}}(c) + \frac{g(\frac{\epsilon}{2})}{2}] \\ & \leq \mathrm{GE}(\Delta^{\mathrm{pn}}(c), \max(m_1, m_2), \Delta^{\mathrm{pn}}(c) + \frac{g(\frac{\epsilon}{2})}{2})) \\ & \leq \frac{\delta}{3N} \end{aligned}$$

   Therefore, the probability that all of the $\frac{\epsilon}{2}$-good concepts satisfy Equation 5.17 is at most $(\frac{\delta}{3N})^N \leq \frac{\delta}{3N} \leq \frac{\delta}{3}$ because the number of $\frac{\epsilon}{2}$-good concepts in $\mathcal{C}_N$ is at most N.

3. Considering Equation 5.18, for at least one ϵ-bad concept in $\mathcal{C}_N$, $\mathcal{F}_{\mathrm{pn}}(c', \mathcal{S}_2)$ is at most $\Delta^{\mathrm{pn}}(c) + \frac{g(\frac{\epsilon}{2})}{2}$. Again by applying Lemma A.4 from the Appendix

$$\begin{aligned} Pr_{x\sim D}[\mathcal{F}_{\mathrm{pn}}(c', \mathcal{S}) &\leq \Delta^{\mathrm{pn}}(c) + \frac{g(\frac{\epsilon}{2})}{2}] \\ &< \mathrm{LE}(\Delta^{\mathrm{pn}}(c) + g(\frac{\epsilon}{2}), \max(m_1, m_2), \Delta^{\mathrm{pn}}(c) + \frac{g(\frac{\epsilon}{2})}{2}) \\ &\leq \frac{\delta}{3N} \end{aligned}$$

Therefore, the probability that there exists an ϵ-bad concept in $\mathcal{C}_n$ that satisfies Equation 5.18 is at most $(N-1)\frac{\delta}{3N} < \frac{\delta}{3}$ because the number of ϵ-bad concepts in $\mathcal{C}_N$ is at most $N-1$. (There is at least one $\frac{\epsilon}{2}$-good concept in $\mathcal{C}_N$ with probability of at least $1 - \frac{\delta}{3}$.)

So the total probability of failure of the algorithm is at most $3 \times \frac{\delta}{3} = \delta$. Therefore, with probability of at least $1 - \delta$ the concept returned by $\mathcal{L}$ has an error of at most ϵ. □

Corollary 5.17. *Let $\mathcal{C}$ be a* PAC*-learnable concept class and Φ a deterministic noise model. If Φ is* pn*-unambiguous with respect to $\mathcal{C}$ and $\mathcal{D}_{\mathcal{X}}$ then $\mathcal{C}$ is* $\mathrm{PAC}_{0,0}$*-learnable with respect to Φ and $\mathcal{D}_{\mathcal{X}}$.*

Now we introduce the second property of label noise models, called pn-monotonicity, defined as follows.

Definition 5.18. *Let $\mathcal{C}$ be a concept class, $\mathcal{D} \subseteq \mathcal{D}_{\mathcal{X}}$ a class of distributions and Φ a deterministic noise model. Φ is* with respect to $\mathcal{C}$ and $\mathcal{D}$ *if for any target concept $c^* \in \mathcal{C}$, for any distribution $D \in \mathcal{D}$ and for any pair of concepts $c, c' \in \mathcal{C}$*

$$\Delta^{\mathrm{pp}}(c') > \Delta^{\mathrm{pp}}(c) \Rightarrow \Delta^{\mathrm{pn}}(c') > \Delta^{\mathrm{pn}}(c) \tag{5.19}$$

Next, we show that pn-unambiguity of a noise model with respect to a concept class and a class of distributions implies the pn-monotonicity of such a noise model with respect to the same concept class and the same class of distributions.

Proposition 5.19. *Let $\mathcal{C}$ be a concept class, $\mathcal{D} \subseteq \mathcal{D}_{\mathcal{X}}$ a class of distributions and Φ a deterministic noise model. If Φ is* pn*-unambiguous with respect to $\mathcal{C}$ and $\mathcal{D}$ then Φ is* pn*-monotonic with respect to $\mathcal{C}$ and $\mathcal{D}$.*

Proof. Let $D \in \mathcal{D}$ be a distribution and $c^* \in \mathcal{C}$ the target concept. Let $c, c' \in \mathcal{C}$ be such that $\Delta^{\mathrm{pp}}(c') > \Delta^{\mathrm{pp}}(c)$. Let $\epsilon = \frac{\Delta^{\mathrm{pp}}(c') - \Delta^{\mathrm{pp}}(c)}{2}$. Since Φ is pn-unambiguous with respect to $\mathcal{C}$ and $\mathcal{D}$ there exists a function g such that $\Delta^{\mathrm{pn}}(c') - \Delta^{\mathrm{pn}}(c) \geq g(\epsilon) > 0$ and thus $\Delta^{\mathrm{pn}}(c') > \Delta^{\mathrm{pn}}(c)$. Therefore, Φ is pn-monotonic with respect to $\mathcal{C}$ and $\mathcal{D}$. □

Using Proposition 5.15 and then Proposition 5.19, we can easily show that the η-random classification noise model is pn-monotonic with respect to any concept class and $\mathcal{D}_{\mathcal{X}}$ when $\eta < \frac{1}{2}$.

Corollary 5.20. *Let $\mathcal{C}$ be a concept class. The η-random classification noise model, $\Phi_{rcn(\eta)}$, is* pn*-monotonic with respect to $\mathcal{C}$ and $\mathcal{D}_{\mathcal{X}}$ if $\eta < \frac{1}{2}$.*

We believe that the reverse direction of Proposition 5.19 is not true $i.e.$, the pn-monotonicity of a noise model with respect to a concept class and a class of distributions does not imply the pn-unambiguity of such a noise model with respect to the same concept class and the same class of distributions. This is stated in the following conjecture.

Conjecture 5.21. *There exists a concept class $\mathcal{C}$, a distribution $D \in \mathcal{D}_{\mathcal{X}}$ and a deterministic noise model Φ such that Φ is* pn*-monotonic with respect to $\mathcal{C}$ and $\{D\}$ but it is* pn*-ambiguous with respect to $\mathcal{C}$ and $\{D\}$.*

We also believe that not all concept classes of finite VC-dimension are $\mathrm{PAC}_{0,0}$-learnable with respect to a pn-monotonic noise model and $\mathcal{D}_{\mathcal{X}}$. This is stated in the following conjecture.

Conjecture 5.22. *There exists a concept class $\mathcal{C}$ of finite* VC*-dimension and a deterministic noise model Φ such that Φ is* pn*-monotonic with respect to $\mathcal{C}$ and $\mathcal{D}_{\mathcal{X}}$ but $\mathcal{C}$ is not $\mathrm{PAC}_{0,0}$-learnable with respect to Φ and $\mathcal{D}_{\mathcal{X}}$.*

However, any finite concept class is $\mathrm{PAC}_{0,0}$-learnable with respect to $\mathcal{D}_{\mathcal{X}}$ and a noise model that is pn-monotonic with respect to the concept class and $\mathcal{D}_{\mathcal{X}}$.

Proposition 5.23. *Let $\mathcal{C}$ be a finite concept class and Φ a deterministic noise model. If Φ is* pn*-monotonic with respect to $\mathcal{C}$ and $\mathcal{D}_{\mathcal{X}}$ then $\mathcal{C}$ is $\mathrm{PAC}_{0,0}$-learnable with respect to Φ and $\mathcal{D}_{\mathcal{X}}$.*

Sketch of the Proof. Let $k = \min_{c,c' \in \mathcal{C}} |\Delta^{\mathrm{pn}}(c) - \Delta^{\mathrm{pn}}(c')|$. We know k exists because the concept class is finite. Since Φ is pn-monotonic with respect to $\mathcal{C}$ and $\mathcal{D}_{\mathcal{X}}$, the target concept has the minimum Δ^{pn} value among all the concepts in $\mathcal{C}$. Therefore, there is a separation of at least k between the Δ^{pn} value of the target concept and that of any other concept in $\mathcal{C}$.

The rest of the proof is very similar to the proof of Theorem 5.16 because the minimum pn-disagreement learning algorithm can use this separation of Δ^{pn} values to find a sample size to guarantee that with high probability no ϵ-bad concept minimizes $\mathcal{F}_{\mathrm{pn}}$. □

Unfortunately, for all of our locally variable noise models, defined in Chapter 3, there exists a pair of concept class and distribution such that the concept class is not PAC-learnable with respect to any of the noise models and that distribution, even though the class is PAC-learnable in the noise-free setting.

Proposition 5.24. *Let $\underline{\epsilon}, \underline{\delta} \in [0, \frac{1}{2})$. For any of the noise models Φ introduced in Chapter 3, there exists a concept class $\mathcal{C}$ of finite* VC*-dimension and a distribution D such that $\mathcal{C}$ is not $\mathrm{PAC}_{0,0}$-learnable with respect to Φ and $\{D\}$.*

Proof. For $\Phi_{\mathrm{drball}(\rho)}$ and $\Phi_{\mathrm{dmball}(\rho)}$, let $\rho > 0$, $\mathcal{X} = \mathbb{R}$ and $x_1, x_2 \in \mathbb{R}$ with $\mathrm{dist}(x_1, x_2) < \rho$. Let D be a uniform distribution such that $Pr_{x \sim D}[x = x_1] = Pr_{x \sim D}[x = x_2] = 0.5$. Therefore, $\mathrm{supp}(D) = \{x_1, x_2\}$.

Let $\mathcal{C} = \{c, c'\}$ with $c = \{x_1\}$, $c = \{x_2\}$. In the rest of the proof, we show that $\mathrm{EX}_{\Phi_{\mathrm{drball}(\rho)}}(c, D)$ and $\mathrm{EX}_{\Phi_{\mathrm{drball}(\rho)}}(c', D)$ have potentially equivalent behavior on D. Therefore, Lemma 2.12 can be used to show that $\mathcal{C}$ is not $\mathrm{PAC}_{\underline{\epsilon},\underline{\delta}}$-learnable with respect to $\Phi_{\mathrm{drball}(\rho)}$ and $\mathcal{D}_{\mathcal{X}}$.

Based on the definition of ρ-distance random classification noise (Definition 3.5), for all $x \in \{x_1, x_2\}$, $\Phi_{\mathrm{drball}(\rho)}(c, D, x) = \Phi_{\mathrm{drball}(\rho)}(c', D, x) = \{0.5\}$. As a result, $\mathrm{EX}_{\mathrm{drball}(\rho)}(c, D)$ and $\mathrm{EX}_{\mathrm{drball}(\rho)}(c', D)$ have equivalent behavior on D based on Definition 2.10.

Since $Pr_{x \sim D}[c(x) \neq c'(x)] = 1 > \underline{\epsilon}$ for any $\underline{\epsilon} \in [0, \frac{1}{2})$, $\mathcal{C}$ is not $\mathrm{PAC}_{\underline{\epsilon},\underline{\delta}}$-learnable with respect to $\Phi_{\mathrm{drball}(\rho)}$ and $\mathcal{D}_{\mathcal{X}}$. Also, using the contrapositive of Proposition 3.6, $\mathcal{C}$ is not $\mathrm{PAC}_{\underline{\epsilon},\underline{\delta}}$-learnable with respect to $\Phi_{\mathrm{dmball}(\rho)}$ and $\{D\}$.

For $\Phi_{\mathrm{wrball}(\omega)}$ and $\Phi_{\mathrm{wmball}(\omega)}$, let $\omega \in (0, 1)$, $t = \lceil \frac{1}{\omega} \rceil$ and $\mathcal{X} = \{x_1, \ldots, x_{2t}\}$. Also let $\mathrm{dist}(x_i, x_{i+t}) < \mathrm{dist}(x_i, x)$ for $1 \leq i \leq t$ and $x \neq x_{i+t}$ and $x \neq x_i$. Let D be a uniform distribution such that $Pr_{x \sim D}[x = x_1] = \ldots = Pr_{x \sim D}[x = x_{2t}] = \frac{1}{2t}$. Therefore, $\mathrm{supp}(D) = \{x_1, \ldots, x_{2t}\}$.

Let $\mathcal{C} = \{c, c'\}$ such that $c(x_i) = 1$ for $1 \leq i \leq t$, $c(x_i) = 0$ for $t + 1 \leq i \leq 2t$ and $c'(x) = 1 - c(x)$ for all $x \in \mathcal{X}$. We show that $\mathrm{EX}_{\Phi_{\mathrm{wrball}(\omega)}}(c, D)$ and $\mathrm{EX}_{\Phi_{\mathrm{wrball}(\omega)}}(c', D)$ have potentially equivalent behavior on D. Therefore, we can use Lemma 2.12 to show that $\mathcal{C}$ is not $\mathrm{PAC}_{\underline{\epsilon},\underline{\delta}}$-learnable with respect to $\Phi_{\mathrm{wrball}(\omega)}$ and $\mathcal{D}_{\mathcal{X}}$.

Based on the definition of ω-weight random classification noise (Definition 3.17), for all $x \in \mathcal{X}$, $\Phi_{\mathrm{wrball}(\omega)}(c, D, x) = \Phi_{\mathrm{wrball}(\omega)}(c', D, x) = \{0.5\}$. As a result, based on Definition 2.10, $\mathrm{EX}_{\mathrm{wrball}(\omega)}(c, D)$ and $\mathrm{EX}_{\mathrm{wrball}(\omega)}(c', D)$ have potentially equivalent behavior on D.

Since $Pr_{x \sim D}[c(x) \neq c'(x)] = 1 > \underline{\epsilon}$ for any $\underline{\epsilon} \in [0, \frac{1}{2})$, $\mathcal{C}$ is not $\mathrm{PAC}_{\underline{\epsilon},\underline{\delta}}$-learnable with respect to $\Phi_{\mathrm{wrball}(\omega)}$ and $\mathcal{D}_{\mathcal{X}}$. Also, using the contrapositive of Proposition 3.18, $\mathcal{C}$ is not $\mathrm{PAC}_{\underline{\epsilon},\underline{\delta}}$-learnable with respect to $\Phi_{\mathrm{wmball}(\omega)}$ and $\{D\}$. □

Based on Theorem 5.16, any concept class of finite VC-dimension is $\mathrm{PAC}_{0,0}$-learnable with respect to $\mathcal{D}_{\mathcal{X}}$ and any noise model that is pn-unambiguous with respect to that concept class and $\mathcal{D}_{\mathcal{X}}$. Therefore, we can use Theorem 5.16 and Proposition 5.24 to conclude the following corollary.

Corollary 5.25. *For the deterministic noise models Φ introduced in Chapter 3, there exists a concept class $\mathcal{C}$ of finite* VC*-dimension and a distribution D such that Φ is* pn*-ambiguous with respect to $\mathcal{C}$ and $\{D\}$.*

Also, based on Proposition 5.23, any finite concept class is $\mathrm{PAC}_{0,0}$-learnable with respect to $\mathcal{D}_{\mathcal{X}}$ and any noise model that is pn-monotonic with respect to that concept class and $\mathcal{D}_{\mathcal{X}}$. Therefore, we can use Proposition 5.23 and Proposition 5.24 to conclude the following corollary because the counterexample in the proof of Proposition 5.24 is regarding a finite concept class.

Corollary 5.26. *For the deterministic noise models Φ introduced in Chapter 3, there exists a finite concept class $\mathcal{C}$ and a distribution D such that Φ is not* pn*-monotonic with respect to $\mathcal{C}$ and $\{D\}$.*

We can use Example 5.7 to show that there exists a concept class, a noise model and a distribution such that the noise model is neither pn-unambiguous nor pn-monotonic with respect to the concept class and the distribution. Therefore, the concept class is not $\mathrm{PAC}_{0,0}$-learnable with respect to the noise model and any class of distributions containing that specific distribution. However, the concept class is still $\mathrm{PAC}_{0,0}$-learnable with respect to that noise model and the class of all possible distributions. The example is repeated here for ease of reference.

Example 5.27. *Let $\mathcal{X} = \{x_1, x_2\}$, $\mathcal{C} = \{c_1, c_2, c_3\}$ the concept class described in Table 5.3 and $D \in \mathcal{D}_{\mathcal{X}}$ a distribution. Let $Pr_{x\sim D}[x = x_1] = 0.25$ and $Pr_{x\sim D}[x = x_2] = 0.75$. Let Φ be a deterministic label noise model with $\mathrm{nr}_{c^*,D}(x_1) = 0.75$ and $\mathrm{nr}_{c^*,D}(x_2) = 0.25$ for any $c^* \in \mathcal{C}$ (see Table 5.3).*

concept/label	x_1	x_2	concept/value	x_1	x_2
c_1	1	1	$\Phi_{c_1,D}$	0.25	0.75
c_2	0	1	$\Phi_{c_2,D}$	0.75	0.75
c_3	1	0	$\Phi_{c_3,D}$	0.25	0.25

Table 5.3: Concept class in Example 5.27

As we showed in Example 5.7, the Δ^{pp} and Δ^{pn} values for the case that c_2 is the target concept can be computed as in Table 5.4. Since $\Delta^{\mathrm{pn}}(c_2) > \Delta^{\mathrm{pn}}(c_1)$ but $\Delta^{\mathrm{pp}}(c_1) > \Delta^{\mathrm{pp}}(c_2)$, Φ is not pn*-monotonic with respect to $\mathcal{C}$ and $\{D\}$ (see Table 5.4). Using the contrapositive of Proposition 5.19, Φ is* pn*-ambiguous with respect to $\mathcal{C}$ and $\{D\}$. Since the minimizer of the Δ^{pn} is not the target concept, the minimum* pn*-disagreement strategy cannot be used to $\mathrm{PAC}_{0,0}$-learn $\mathcal{C}$ with respect to Φ and $\{D\}$. However, $\mathcal{C}$ is $\mathrm{PAC}_{0,0}$-learnable with respect to Φ and $\mathcal{D}_{\mathcal{X}}$ using Proposition 4.3.*

concept	Δ^{pp}	Δ^{pn}
c_1	0.25	0.25
$c^* = c_2$	0	0.31
c_3	1	0.69

Table 5.4: Δ^{pn} values in Example 5.27

Examples such as Example 5.27, that are $\mathrm{PAC}_{0,0}$-learnable but are not $\mathrm{PAC}_{0,0}$-learnable using the minimum pn-disagreement strategy, make us ask for a more general strategy. We introduce one such strategy in Chapter 6 and show that the new strategy can be used to $\mathrm{PAC}_{0,0}$-learn the concept class in Example 5.27 with respect to the noise model in the example and $\mathcal{D}_{\mathcal{X}}$.

Chapter 6

Minimum nn-disagreement Strategies

In this chapter, we study a new general learning algorithm with respect to label noise called the minimum nn-disagreement strategy. As in Chapter 5, we consider the application of this strategy with random classification noise. Furthermore, we investigate the general characteristics of a label noise model that makes the learnability of certain concept classes possible with the minimum nn-disagreement strategy.

As in Chapter 5, we assume the noise model is deterministic. We start by introducing more notation.

For any fixed concept and distribution, the nn-difference of a concept with respect to the target concept and the distribution is defined as follows.

Definition 6.1. ***(Kearns and Schapire [24])*** *Let $\mathcal{C}$ be a concept class and D a distribution. Let $c^* \in \mathcal{C}$ be the target concept and Φ a deterministic noise model. For any concept $c \in C$, the* nn-difference $\Delta^{\mathrm{nn}}_{c^*,D,\Phi}(c)$ *between c and c^* is defined as*

$$\Delta^{\mathrm{nn}}_{c^*,D,\Phi}(c) = E_{x\sim D}[|\Phi_{c,D}(x) - \Phi_{c^*,D}(x)|] \tag{6.1}$$

For the sake of convenience, we write $\Delta^{\mathrm{nn}}(c)$ instead of $\Delta^{\mathrm{nn}}_{c^*,D,\Phi}(c)$ whenever c^*, D and Φ are clear from the context[1].

Δ^{nn} has also been called *variational distance* in the literature [24]. Also it is not hard to verify that unlike the case of Δ^{pn}, for any target concept $c^* \in \mathcal{C}$, $\Delta^{\mathrm{nn}}(c^*)$ is always 0.

Unlike the case of Δ^{pp} and Δ^{pn}, Δ^{nn} cannot be written in terms of probability instead of expectation with the same pattern as Δ^{pp} and Δ^{pn}. This is shown in the following proposition.

Proposition 6.2. *There exists a concept class $\mathcal{C}$, a target concept $c^* \in \mathcal{C}$, a distribution D and a deterministic noise model Φ such that*

$$\Delta^{\mathrm{nn}}_{c^*,D,\Phi}(c) \neq Pr_{x\sim D,y\sim\hat{c},y'\sim\hat{c}^*}[y \neq y'] \tag{6.2}$$

for some $c \in \mathcal{C}$.

Proof. $Pr_{x\sim D,y\sim\hat{c},y'\sim\hat{c}^*}[y \neq y']$ is the probability that the $i.i.d.$ draws y and y', from the random variables $\hat{c}$ and $\hat{c}^*$ respectively, disagree with each other. Since each of these random variables can only take values 0 and 1, the disagreement happens when one variable takes the value 0 and the other one 1. For convenience and ease of readability, in the rest of the proof we write $Pr[]$ and $E[]$

[1] The designation nn is used here to show that the disagreement is measured between two "noisy" concepts in the sense that both of the concepts are mappings from the input space to $[0, 1]$.

instead of $Pr_{x\sim D,y\sim\hat{c},y'\sim\hat{c}^*}[]$ and $E_{x\sim D}[]$ respectively.

$$\begin{aligned}
&Pr[y \neq y'] \\
&\quad = Pr[y = 1 \text{ and } y' = 0] + Pr[y = 0 \text{ and } y' = 1] \\
&\quad = E[\Phi_{c,D}(x)(1 - \Phi_{c^*,D}(x))] + E[(1 - \Phi_{c,D}(x))\Phi_{c^*,D}(x)] \\
&\quad = E[\Phi_{c,D}(x) - \Phi_{c,D}(x)\Phi_{c^*,D}(x) + \Phi_{c^*,D}(x) - \Phi_{c^*,D}(x)\Phi_{c,D}(x)] \\
&\quad = E[\Phi_{c,D}(x) + \Phi_{c^*,D}(x) - 2\Phi_{c,D}(x)\Phi_{c^*,D}(x)]
\end{aligned}$$

We give an example in which $E[|\Phi_{c,D}(x) - \Phi_{c^*,D}(x)|]$ is not equal to $E[\Phi_{c,D}(x) + \Phi_{c^*,D}(x) - 2\Phi_{c,D}(x)\Phi_{c^*,D}(x)]$. Let $\mathcal{X} = \{x\}$, $\mathcal{C} = \{c\}$ with $c(x) = 1$ and the target concept $c^* = c$. The only distribution D over $\mathcal{X}$ assigns all the probability mass to point x. Let $\Phi_{c,D}(x) = 0.4$. As a result, $\Phi_{c^*,D}(x) = 0.4$. Therefore,

$$\begin{aligned}
Pr[y \neq y'] =& E[\Phi_{c,D}(x) + \Phi_{c^*,D}(x) - 2\Phi_{c,D}(x)\Phi_{c^*,D}(x)] \\
&= 0.4 + 0.4 - 2 \times 0.4 \times 0.4 = 0.48
\end{aligned}$$

but as we mentioned before, $\Delta^{\mathrm{nn}}(c^*) = 0$. □

We again use Example 5.7 to show how Δ^{nn} can be computed. We will reintroduce the example briefly.

Example 6.3. *Let $\mathcal{X} = \{x_1, x_2\}$, $\mathcal{C} = \{c_1, c_2, c_3\}$ the concept class described in Table 5.1 and $D \in \mathcal{D}_{\mathcal{X}}$ a distribution. Let $Pr_{x\sim D}[x = x_1] = 0.25$ and $Pr_{x\sim D}[x = x_2] = 0.75$. Let Φ be a deterministic label noise model with $\mathrm{nr}_{c^*,D}(x_1) = 0.75$ and $\mathrm{nr}_{c^*,D}(x_2) = 0.25$ for any $c^* \in \mathcal{C}$ (see Table 5.1).*

We can use Definitions 6.1 to compute the Δ^{nn} values for the case that c_2 is the target concept as in Table 6.1.

concept	Δ^{pp}	Δ^{nn}
c_1	p	$0.25\|0.25 - 0.75\| + 0.75\|0.75 - 0.75\| = 0.13$
$c^* = c_2$	0	$0.25\|0.75 - 0.75\| + 0.75\|0.75 - 0.75\| = 0$
c_3	1	$0.25\|0.25 - 0.75\| + 0.75\|0.25 - 0.75\| = 0.5$

Table 6.1: Δ^{nn} values in Example 6.3

Similar to the case of pn-disagreement, we can also define a measure of disagreement between a concept and a sample as follows.

Definition 6.4. *Let $\mathcal{C}$ be a concept class, D a distribution, and Φ a deterministic noise model. Let $\mathcal{S} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ be a sample of size m. For any concept $c \in \mathcal{C}$, $\mathcal{F}_{\mathrm{nn}}(c, \mathcal{S})$ is defined as follows.*

$$\mathcal{F}_{\mathrm{nn}}(c, \mathcal{S}) = \frac{1}{m}\sum_{i=1}^{m} \mid \Phi_{c,D}(x_i) - \frac{\#^+(x_i, \mathcal{S})}{\#(x_i, \mathcal{S})} \mid \tag{6.3}$$

where for all $x \in \mathcal{X}$, $\#^+(x, \mathcal{S}) = |\{j \in \{1, \ldots, m\} \mid x = x_j \wedge y_j = 1\}|$ and $\#(x, \mathcal{S}) = |\{j \in \{1, \ldots, m\} \mid x = x_j\}|$.

Unlike $\mathcal{F}_{\mathrm{pn}}$, the learner cannot compute $\mathcal{F}_{\mathrm{nn}}$ directly, because it does not know the distribution, and without that it cannot compute the $\Phi_{c,D}(x_i)$ values. However, the learner can estimate $\mathcal{F}_{\mathrm{nn}}$. But first we need the following definition.

Definition 6.5. *Let D be a distribution and $\mathcal{S} = \{x_1, \ldots, x_m\}$ a sample of m instances drawn $i.i.d.$ from D. We define the* estimated distribution of D, $\hat{D}$, *as follows.*

$$Pr_{x' \sim \hat{D}}[x' = x] = \#(x, \mathcal{S}) \cdot \frac{1}{m} \tag{6.4}$$

for all $x \in \mathcal{X}$ where $\#(x, \mathcal{S}) = |\{j \in \{1, \ldots, m\} \mid x = x_j\}|$.

Now, we can define the estimate $\hat{\mathcal{F}}_{\mathrm{nn}}$ of $\mathcal{F}_{\mathrm{nn}}$ as follows. We call this estimate the nn-disagreement between a concept and the sample.

Definition 6.6. *Let $\mathcal{C}$ be a concept class, D a distribution, and Φ a deterministic noise model. Let $\mathcal{S} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ be a sample of size m. Let $\hat{D}$ be the estimated distribution of D as in Definition 6.5 using the unlabeled instances of $\mathcal{S}$. For any concept $c \in \mathcal{C}$, the* nn-disagreement *$\hat{\mathcal{F}}_{\mathrm{nn}}(c, \mathcal{S})$ is defined as follows.*

$$\hat{\mathcal{F}}_{\mathrm{nn}}(c, \mathcal{S}) = \frac{1}{m} \sum_{i=1}^{m} \mid \Phi_{c,\hat{D}}(x_i) - \frac{\#^+(x_i, \mathcal{S})}{\#(x_i, \mathcal{S})} \mid \tag{6.5}$$

where for all $x \in \mathcal{X}$, $\#^+(x, \mathcal{S}) = |\{j \in \{1, \ldots, m\} \mid x = x_j \wedge y_j = 1\}|$ and $\#(x, \mathcal{S}) = |\{j \in \{1, \ldots, m\} \mid x = x_j\}|$.

Next, we define a smooth noise model as a noise model in which the estimate $\hat{\mathcal{F}}_{\mathrm{nn}}$ is close to $\mathcal{F}_{\mathrm{nn}}$ using only a polynomial number of examples.

Definition 6.7. *Let $\mathcal{C}$ be a concept class, $\mathcal{D} \in \mathcal{D}_{\mathcal{X}}$ a class of distributions, $\epsilon, \delta \in (0, \frac{1}{2})$ and Φ a deterministic noise model. Φ is* smooth with respect to $\mathcal{C}$ and $\mathcal{D}$ *iff there is a function $M : (0, \frac{1}{2}) \times (0, \frac{1}{2}) \to \mathbb{N}$ such that*

1. *$M(\epsilon, \delta)$ is polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$.*
2. *For all ϵ, δ, for all target concepts $c^* \in \mathcal{C}$ and for all D in $\mathcal{D}$ if $\mathcal{S}$ is a sample of at least $M(\epsilon, \delta)$ examples drawn from $\mathrm{EX}_{\Phi}(c^*, D)$ then, with probability of at least $1 - \delta$, for all $c \in \mathcal{C}$ we obtain $|\mathcal{F}_{\mathrm{nn}}(c, \mathcal{S}) - \hat{\mathcal{F}}_{\mathrm{nn}}(c, \mathcal{S})| < \epsilon$.*

The minimum nn-disagreement strategy simply returns a concept in the concept class that has the smallest estimated nn-disagreement with the sample.

Definition 6.8. *Let $\mathcal{C}$ be a concept class and $\mathcal{S}$ a sample. A learning algorithm $\mathcal{L}$ is called a* minimum nn-disagreement strategy *if upon seeing $\mathcal{S}$, $\mathcal{L}$ returns a (not necessarily unique) concept $\bar{c} \in \mathcal{C}$ such that:*

$$\bar{c} \in \textit{arg min}_{c \in \mathcal{C}} \, \hat{\mathcal{F}}_{\mathrm{nn}}(c, \mathcal{S}) \tag{6.6}$$

6.1 Minimum nn-disagreement strategies for learning with random classification noise

As in Section 5.1, we consider the application of minimum nn-disagreement strategies for learning with respect to random classification noise.

Similar to the case of pn-disagreement, the nn-disagreement of any concept also has a linear relationship with the pp-disagreement of the same concept in the presence of random classification noise.

Lemma 6.9. *Let $\eta \in [0, 1)$. Let $\mathcal{C}$ be a concept class and D a distribution. For any concept $c \in \mathcal{C}$ in the presence of η-random classification noise, $\Phi_{\mathrm{rcn}(\eta)}$,*

$$\Delta^{\mathrm{nn}}(c) = |1 - 2\eta| \Delta^{\mathrm{pp}}(c) \tag{6.7}$$

Proof. Let $c^* \in \mathcal{C}$ be the target concept. Based on Definition 2.13, $\Phi_{\mathrm{rcn}(\eta)_{c^*,D}}(x) = \Phi_{\mathrm{rcn}(\eta)_{c,D}}(x)$ iff $c^*(x) = c(x)$. Otherwise one of the values $\Phi_{\mathrm{rcn}(\eta)_{c^*,D}}(x)$ and $\Phi_{\mathrm{rcn}(\eta)_{c,D}}(x)$ is equal to $1-\eta$ and the other one is equal to η. Therefore,

$$\begin{aligned}\Delta^{\mathrm{nn}}(c) &= E_{x\sim D}[|\eta - (1-\eta)||c(x) - c^*(x)|] \\ &= |1-2\eta| E_{x\sim D}[|c(x) - c^*(x)|] \\ &= |1-2\eta| \Delta^{\mathrm{pp}}(c)\end{aligned}$$

□

Next we show that, in the presence of random classification noise, the concept that minimizes the pn-disagreement also minimizes the nn-disagreement.

Proposition 6.10. *Let $\eta \in [0, \frac{1}{2})$ and $\mathcal{C}$ a concept class. For any target concept $c^* \in C$, and any distribution $D \in \mathcal{D}_{\mathcal{X}}$ in the presence of random classification noise, $\Phi_{\mathrm{rcn}(\eta)}$, arg min$_{c\in\mathcal{C}}\Delta^{\mathrm{pn}}(c)$ and arg min$_{c\in\mathcal{C}}\Delta^{\mathrm{nn}}(c)$ exist and*

$$\textit{arg min}_{c\in\mathcal{C}}\Delta^{\mathrm{pn}}(c) = \textit{arg min}_{c\in\mathcal{C}}\Delta^{\mathrm{nn}}(c) \tag{6.8}$$

Proof. Note that both arg min$_{c\in\mathcal{C}}\Delta^{\mathrm{pn}}(c)$ and arg min$_{c\in\mathcal{C}}\Delta^{\mathrm{nn}}(c)$ exist because both Δ^{pn} and Δ^{nn} have a linear relationship with Δ^{pp} and the target concept always has the smallest Δ^{pp} in the concept class. Let $\bar{c} \in$ arg min$_{c\in\mathcal{C}}\Delta^{\mathrm{nn}}(c)$. Using Equation 6.7 in Lemma 6.9 and since $(1-2\eta) > 0$,

$$\Delta^{\mathrm{nn}}(c) = (1-2\eta)\Delta^{\mathrm{pp}}(c)$$

Therefore, $\bar{c} \in$ arg min$_{c\in\mathcal{C}}(1-2\eta)\Delta^{\mathrm{pp}}(c)$ and also $\bar{c} \in$ arg min$_{c\in\mathcal{C}}\eta + (1-2\eta)\Delta^{\mathrm{pp}}(c)$. Using Equation 5.8 in Lemma 5.10, $\bar{c} \in$ arg min$_{c\in\mathcal{C}}\Delta^{\mathrm{pn}}(c)$. We can use the similar technique to show that any minimizers of Δ^{pn} are also members of the set of all minimizers of Δ^{nn}. □

Therefore, the minimum pn-disagreement strategy and the minimum nn-disagreement strategy are equivalent for the random classification noise model.

6.2 Sufficient conditions for learning with minimum nn-disagreement strategy

In this section, we investigate the properties of a label noise model that guarantee that the minimum nn-disagreement strategy can be used to learn a concept class. nn-unambiguity is the first such property, defined as follows.

Definition 6.11. *Let $\mathcal{C}$ be a concept class, $\mathcal{D} \subseteq \mathcal{D}_{\mathcal{X}}$ a class of distributions and Φ a deterministic noise model. Φ is* nn-unambiguous with respect to $\mathcal{C}$ and $\mathcal{D}$ *if there exists a function $g : (0, \frac{1}{2}) \to (0, 1)$ such that for any target concept $c^* \in \mathcal{C}$, for any distribution $D \in \mathcal{D}$, for any $\epsilon \in (0, \frac{1}{2})$ and for any pair of concepts $c, c' \in \mathcal{C}$*

$$\Delta^{\mathrm{pp}}(c') - \Delta^{\mathrm{pp}}(c) > \epsilon \Rightarrow \Delta^{\mathrm{nn}}(c') - \Delta^{\mathrm{nn}}(c) \geq g(\epsilon). \tag{6.9}$$

Otherwise Φ is nn-ambiguous *with respect to $\mathcal{C}$ and $\mathcal{D}$.*

Next, we show that η-random classification noise is nn-unambiguous with respect to any concept class and $\mathcal{D}_{\mathcal{X}}$ when $\eta \neq \frac{1}{2}$. The proof of the following proposition is analogous to the proof of Proposition 5.15. However, the random classification noise model is nn-unambiguous not only when $\eta < \frac{1}{2}$ but also when $\eta > \frac{1}{2}$.

Proposition 6.12. *Let $\mathcal{C}$ be a concept class and $\eta \in (0, 1)$. The η-random classification noise model, $\Phi_{\mathrm{rcn}(\eta)}$, is nn-unambiguous with respect to $\mathcal{C}$ and $\mathcal{D}_{\mathcal{X}}$ iff $\eta \neq \frac{1}{2}$.*

Proof. As we previously showed in Lemma 6.9, in the presence of random classification noise

$$\Delta^{\mathrm{nn}}(c) = |1-2\eta|\Delta^{\mathrm{pp}}(c)$$

for all $c \in \mathcal{C}$. Let $c, c' \in \mathcal{C}$ such that $\Delta^{\mathrm{pp}}(c') > \Delta^{\mathrm{pp}}(c)$. Let $\epsilon = \frac{\Delta^{\mathrm{pp}}(c') - \Delta^{\mathrm{pp}}(c)}{2}$. When $\eta \neq \frac{1}{2}$, $|1-2\eta| > 0$. Therefore,

$$\begin{aligned}\Delta^{\mathrm{nn}}(c') - \Delta^{\mathrm{nn}}(c) &= |1-2\eta|\Delta^{\mathrm{pp}}(c') - |1-2\eta|\Delta^{\mathrm{pp}}(c) \\ &= |1-2\eta|(\Delta^{\mathrm{pp}}(c') - \Delta^{\mathrm{pp}}(c)) = 2\epsilon|1-2\eta| > 0\end{aligned}$$

The function g defined by $g(\epsilon) = \epsilon|1-2\eta|$ witnesses that Φ is nn-unambiguous with respect to $\mathcal{C}$ and $\mathcal{D}_{\mathcal{X}}$. When $\eta = \frac{1}{2}$, $|1-2\eta| = 0$. Therefore,

$$\Delta^{\mathrm{nn}}(c') - \Delta^{\mathrm{nn}}(c) = 0$$

Thus no such g as defined in Definition 6.11 exists and, therefore, Φ is nn-ambiguous. □

Smoothness and nn-unambiguity of a noise model with respect to a concept class and a class of distributions is a sufficient condition for $\mathrm{PAC}_{0,0}$-learnability of that concept class with respect to the noise model and the class of distributions. This is stated in the following Theorem.

Theorem 6.13. *Let $\mathcal{C}$ be a concept class of* VC*-dimension $d < \infty$ and Φ a deterministic noise model. If Φ is both* nn*-unambiguous and smooth with respect to $\mathcal{C}$ and $\mathcal{D}_{\mathcal{X}}$ then $\mathcal{C}$ is $\mathrm{PAC}_{0,0}$-learnable with respect to Φ and $\mathcal{D}_{\mathcal{X}}$.*

Proof. Let $D \in \mathcal{D}$, $\epsilon, \delta \in (0, \frac{1}{2})$ and $c^* \in \mathcal{C}$ the target concept. Let $\mathcal{L}$ be a minimum nn-disagreement strategy. We show that the concept returned by $\mathcal{L}$ has an error of at most ϵ with probability of at least $1-\delta$.

As in the proof of Theorem 2.15 in Section 5.1, $\mathcal{L}$ first draws a sample $\mathcal{S}_1$ of size $m_1(\frac{\epsilon}{2}, \frac{\delta}{4}, d)$ (m_1, for short) to find a set of $N \leq {m_1}^d + 1$ representative concepts $\mathcal{C}_N = \{c_1, \ldots, c_N\}$ of N equivalence classes, among which at least one is $\frac{\epsilon}{2}$-good with probability of at least $1 - \frac{\delta}{4}$.

Let $\mathcal{S}_2$ be a sample that contains $\mathcal{S}_1$ and enough additional examples (if needed) so that it contains at least $M(\frac{g(\frac{\epsilon}{2})}{4}, \frac{\delta}{4})$ (M, for short) examples. Since Φ is smooth with respect to $\mathcal{C}$ and $\mathcal{D}_{\mathcal{X}}$, $|\mathcal{S}_2|$ examples will guarantee that for all $c \in \mathcal{C}$ we can have an estimate $\hat{\mathcal{F}}_{\mathrm{nn}}(c, \mathcal{S}_2)$ for $\mathcal{F}_{\mathrm{nn}}(c, \mathcal{S}_2)$ such that $|\hat{\mathcal{F}}_{\mathrm{nn}}(c, \mathcal{S}_2) - \mathcal{F}_{\mathrm{nn}}(c, \mathcal{S}_2)| < \frac{g(\frac{\epsilon}{2})}{4}$, with probability of at least $1 - \frac{\delta}{4}$.

Let $\mathcal{S}_3$ be a sample that contains $\mathcal{S}_2$ and enough additional examples (if needed) so that it contains at least $m_2 \geq \frac{8}{g(\frac{\epsilon}{2})^2}\ln(\frac{4}{\delta})$ examples. Therefore, the total number of examples that $\mathcal{L}$ draws from the noisy oracle is $\max(m_1, m_2, M)$.

Since Φ is nn-unambiguous with respect to $\mathcal{C}$ and $\mathcal{D}_{\mathcal{X}}$, there exists a function g such that for any ϵ and for any pair of concepts $c, c' \in \mathcal{C}$ with $\Delta^{\mathrm{pp}}(c') - \Delta^{\mathrm{pp}}(c) > \frac{\epsilon}{2}$, $\Delta^{\mathrm{nn}}(c') - \Delta^{\mathrm{nn}}(c) \geq g(\frac{\epsilon}{2})$. Therefore, there is a separation of at least $g(\frac{\epsilon}{2})$ between the Δ^{nn} values of any pair of concepts that have at least a difference of $\frac{\epsilon}{2}$ between their Δ^{pp} values. For the rest of this proof, any occurrence of c' and c refers to ϵ-bad and $\frac{\epsilon}{2}$-good concepts respectively.

In order for some ϵ-bad concept in $\mathcal{C}_N$ to minimize $\mathcal{F}_{\mathrm{nn}}$ at least one of the following inequalities would have to be fulfilled. For all the $\frac{\epsilon}{2}$-good concepts in $\mathcal{C}_N$

$$\mathcal{F}_{\mathrm{nn}}(c, \mathcal{S}) \geq \Delta^{\mathrm{nn}}(c) + \frac{g(\frac{\epsilon}{2})}{4} \tag{6.10}$$

or for at least one ϵ-bad concept in $\mathcal{C}_N$

$$\mathcal{F}_{\mathrm{nn}}(c', \mathcal{S}) \leq \Delta^{\mathrm{nn}}(c) + \frac{3g(\frac{\epsilon}{2})}{4} \tag{6.11}$$

because otherwise an ϵ-good concept minimizes $\mathcal{F}_{\mathrm{nn}}$. Note that, unlike other proofs in Chapter 5, we need an additional gap of $2 \times \frac{g(\frac{\epsilon}{2})}{4} = \frac{g(\frac{\epsilon}{2})}{2}$ between the $\mathcal{F}_{\mathrm{nn}}$ values due to the error that may happen in estimating $\mathcal{F}_{\mathrm{nn}}$ values. As mentioned in Definition 6.8, this estimate, $\hat{\mathcal{F}}_{\mathrm{nn}}$, will be used by $\mathcal{L}$ instead of $\mathcal{F}_{\mathrm{nn}}$. Therefore, $\mathcal{L}$ fails only if any of the following cases happens.

1. There is no $\frac{\epsilon}{2}$-good concept in $\mathcal{C}_N$. As in the proof of Theorem 2.15, this will happen with probability of at most $\frac{\delta}{4}$.

2. The estimate $\hat{\mathcal{F}}_{\mathrm{nn}}$ of $\mathcal{F}_{\mathrm{nn}}$ has deviation of more than $\frac{g(\frac{\epsilon}{2})}{4}$ for at least one of the concepts. This will happen with probability of at most $\frac{\delta}{4}$ because of the sample size of at least $M(\frac{g(\frac{\epsilon}{2})}{4}, \frac{\delta}{4})$.

3. Considering Equation 6.10, all $\frac{\epsilon}{2}$-good concepts in $\mathcal{C}_N$ have $\mathcal{F}_{\mathrm{nn}}(c, \mathcal{S}_3)$ of at least $\Delta^{\mathrm{nn}}(c) + \frac{g(\frac{\epsilon}{2})}{4}$. Applying Lemma A.4 from the Appendix

$$\begin{aligned} Pr_{x\sim D}[\mathcal{F}_{\mathrm{nn}}(c, \mathcal{S}_3) \geq & \Delta^{\mathrm{nn}}(c) + \frac{g(\frac{\epsilon}{2})}{4}] \\ & \leq \mathrm{GE}(\Delta^{\mathrm{nn}}(c), |\mathcal{S}_3|, \Delta^{\mathrm{nn}}(c) + \frac{g(\frac{\epsilon}{2})}{4}) \\ & \leq \frac{\delta}{4N} \end{aligned}$$

   Therefore, the probability that all of the $\frac{\epsilon}{2}$-good concepts satisfy Equation 6.10 is at most $(\frac{\delta}{4N})^N \leq \frac{\delta}{4N} \leq \frac{\delta}{4}$ because the number of $\frac{\epsilon}{2}$-good concepts in $\mathcal{C}_N$ is at most N.

4. Considering Equation 6.11, for at least one ϵ-bad concept in $\mathcal{C}_N$, $\mathcal{F}_{\mathrm{nn}}(c', \mathcal{S}_3)$ is at most $\Delta^{\mathrm{nn}}(c) + \frac{3g(\frac{\epsilon}{2})}{4}$. Again by applying Lemma A.4 from the Appendix

$$\begin{aligned} Pr_{x\sim D}[\mathcal{F}_{\mathrm{nn}}(c', \mathcal{S}_3) \leq & \Delta^{\mathrm{nn}}(c) + \frac{3g(\frac{\epsilon}{2})}{4}] \\ & < \mathrm{LE}(\Delta^{\mathrm{nn}}(c) + g(\frac{\epsilon}{2}), |\mathcal{S}_3|, \Delta^{\mathrm{nn}}(c) + \frac{3g(\frac{\epsilon}{2})}{4}) \\ & \leq \frac{\delta}{4N} \end{aligned}$$

   Therefore, the probability that there exists an ϵ-bad concept in $\mathcal{C}_n$ that satisfies Equation 6.11 is at most $(N-1)\frac{\delta}{4N} < \frac{\delta}{4}$ because the number of ϵ-bad concepts in $\mathcal{C}_N$ is at most $N-1$. (There is at least one $\frac{\epsilon}{2}$-good concept in $\mathcal{C}_N$ with probability of at least $1 - \frac{\delta}{4}$.)

So the total probability of failure of the algorithm is at most $4 \times \frac{\delta}{4} = \delta$. Therefore, with probability of at least $1 - \delta$ the concept returned by $\mathcal{L}$ has an error of at most ϵ. □

Corollary 6.14. *Let $\mathcal{C}$ be a* PAC*-learnable concept class and Φ a deterministic noise model. If Φ is both* nn*-unambiguous and smooth with respect to $\mathcal{C}$ and $\mathcal{D}_{\mathcal{X}}$ then $\mathcal{C}$ is* $\mathrm{PAC}_{0,0}$*-learnable with respect to Φ and $\mathcal{D}_{\mathcal{X}}$.*

Next we introduce a second property of noise models called nn-monotonicity.

Definition 6.15. *Let $\mathcal{C}$ be a concept class, $\mathcal{D} \subseteq \mathcal{D}_{\mathcal{X}}$ a class of distributions and Φ a deterministic noise model. Φ is* nn-monotonic with respect to $\mathcal{C}$ and $\mathcal{D}$ *if for any target concept $c^* \in \mathcal{C}$, any distribution $D \in \mathcal{D}$ and any pair of concepts $c, c' \in \mathcal{C}$:*

$$\Delta^{\mathrm{pp}}(c') > \Delta^{\mathrm{pp}}(c) \Rightarrow \Delta^{\mathrm{nn}}(c') > \Delta^{\mathrm{nn}}(c) \tag{6.12}$$

With the same technique as in the proof of Proposition 5.19, we show that nn-unambiguity of a noise model with respect to a concept class and a class of distributions implies the nn-monotonicity of the noise model with respect to the same concept class and the same class of distributions.

Proposition 6.16. *Let $\mathcal{C}$ be a concept class, $\mathcal{D} \subseteq \mathcal{D}_{\mathcal{X}}$ a class of distributions and Φ a deterministic noise model. If Φ is* nn*-unambiguous with respect to $\mathcal{C}$ and $\mathcal{D}$ then Φ is* nn*-monotonic with respect to $\mathcal{C}$ and $\mathcal{D}$.*

Proof. Let $D \in \mathcal{D}$ be a distribution and $c^* \in \mathcal{C}$ the target concept. Let $c, c' \in \mathcal{C}$ such that $\Delta^{\mathrm{pp}}(c') > \Delta^{\mathrm{pp}}(c)$. Let $\epsilon = \frac{\Delta^{\mathrm{pp}}(c') - \Delta^{\mathrm{pp}}(c)}{2}$. Since Φ is nn-unambiguous there exists a function g such that $\Delta^{\mathrm{nn}}(c') - \Delta^{\mathrm{nn}}(c) \geq g(\epsilon)$, and since $g(\epsilon) > 0$, $\Delta^{\mathrm{nn}}(c') > \Delta^{\mathrm{nn}}(c)$. Therefore, Φ is nn-monotonic with respect to $\mathcal{C}$ and $\mathcal{D}$. □

Using Proposition 6.12 and then Proposition 6.16, we can show that η-random classification noise is nn-monotonic with respect to any concept class and $\mathcal{D}_{\mathcal{X}}$ when $\eta \neq \frac{1}{2}$.

Corollary 6.17. *Let $\mathcal{C}$ be a concept class and $\eta \in (0, 1)$. The η-random classification noise model, $\Phi_{\mathrm{rcn}(\eta)}$, is nn-monotonic with respect to $\mathcal{C}$ and $\mathcal{D}_{\mathcal{X}}$ if $\eta \neq \frac{1}{2}$.*

Similar to Chapter 5, we believe that the reverse direction of Proposition 6.16 is not true $i.e.$, the nn-monotonicity of a noise model with respect to a concept class and a class of distributions does not imply the nn-unambiguity of such a noise model with respect to the same concept class and the same class of distributions. This is stated in the following conjecture.

Conjecture 6.18. *There exists a concept class $\mathcal{C}$, a distribution $D \in \mathcal{D}_{\mathcal{X}}$ and a deterministic noise model Φ such that Φ is nn-monotonic with respect to $\mathcal{C}$ and $\{D\}$ but it is nn-ambiguous with respect to $\mathcal{C}$ and $\{D\}$.*

We also believe that not all concept classes of finite VC-dimension are $\mathrm{PAC}_{0,0}$-learnable with respect to an nn-monotonic noise model and $\mathcal{D}_{\mathcal{X}}$. This is stated in the following conjecture.

Conjecture 6.19. *There exists a concept class $\mathcal{C}$ of finite VC-dimension and a deterministic noise model Φ such that Φ is smooth and nn-monotonic with respect to $\mathcal{C}$ and $\mathcal{D}_{\mathcal{X}}$ but $\mathcal{C}$ is not $\mathrm{PAC}_{0,0}$-learnable with respect to Φ and $\mathcal{D}_{\mathcal{X}}$.*

However, any finite concept class is $\mathrm{PAC}_{0,0}$-learnable with respect to $\mathcal{D}_{\mathcal{X}}$ and a noise model that is both nn-monotonic and smooth with respect to the concept class and $\mathcal{D}_{\mathcal{X}}$.

Proposition 6.20. *Let $\mathcal{C}$ be a finite concept class and Φ a noise model. If Φ is both nn-monotonic and smooth with respect to $\mathcal{C}$ and $\mathcal{D}_{\mathcal{X}}$ then $\mathcal{C}$ is $\mathrm{PAC}_{0,0}$-learnable with respect to Φ and $\mathcal{D}_{\mathcal{X}}$.*

Sketch of the Proof. The sketch is exactly the same as the sketch of the proof of Proposition 5.23 replacing any occurrence of Δ^{pn} with Δ^{nn} and replacing the reference to Theorem 5.16 with a reference to Theorem 6.13. □

Note that we do not need the learning algorithm in the proof of Proposition 6.20 to be a minimum nn-disagreement strategy. However, the minimum nn-disagreement strategy can be used as the learning algorithm to prove the result. The same argument can be made for Proposition 5.23 and minimum pn-disagreement strategies.

We use Example 5.27 to show that there exists a concept class and a noise model and a distribution such that the noise model is nn-monotonic (nn-unambiguous) with respect to the concept class and $\mathcal{D}_{\mathcal{X}}$ but it is neither pn-monotonic nor pn-unambiguous with respect to the same concept class and any class of distributions containing that specific distribution. Also as we showed in Chapter 5, the concept class in the following example is not $\mathrm{PAC}_{0,0}$-learnable with any minimum pn-disagreement strategy. We repeat the example briefly for ease of reference.

Example 6.21. *Let $\mathcal{X} = \{x_1, x_2\}$, $\mathcal{C} = \{c_1, c_2, c_3\}$ the concept class described in Table 6.2 and $D \in \mathcal{D}_{\mathcal{X}}$ a distribution. Let $Pr_{x\sim D}[x = x_1] = p$ and $Pr_{x\sim D}[x = x_2] = 1 - p$ where $p \in [0, 1]$. Let Φ be a deterministic label noise model with $\mathrm{nr}_{c^*,D}(x_1) = 0.75$ and $\mathrm{nr}_{c^*,D}(x_2) = 0.25$ for any $c^* \in \mathcal{C}$ (see Table 6.2).*

Δ^{nn} values can be computed similar to Example 6.3 (See Table 6.3). Therefore, for any concept $c \in \mathcal{C}$, $\Delta^{\mathrm{nn}}(c) = 0.5\Delta^{\mathrm{pp}}(c)$ and Φ is nn-unambiguous (with $g(\epsilon) = \frac{\epsilon}{2}$), and also nn-monotonic (using Proposition 6.16) with respect to $\mathcal{C}$ and $\mathcal{D}_{\mathcal{X}}$[2].

[2]Because the parametric distribution D can represent all the distributions on $\mathcal{X}$.

concept/label	x_1	x_2	concept/value	x_1	x_2
c_1	1	1	$\Phi_{c_1,D}$	0.25	0.75
c_2	0	1	$\Phi_{c_2,D}$	0.75	0.75
c_3	1	0	$\Phi_{c_3,D}$	0.25	0.25

Table 6.2: Concept class in Example 6.21

concept	Δ^{pp}	Δ^{nn}
$c^* = c_1$	0	0
c_2	p	$0.5p$
c_3	$1-p$	$0.5(1-p)$
concept	Δ^{pp}	Δ^{nn}
c_1	p	$0.5p$
$c^* = c_2$	0	0
c_3	1	0.5
concept	Δ^{pp}	Δ^{nn}
c_1	$1-p$	$0.5(1-p)$
c_2	1	0.5
$c^* = c_3$	0	

Table 6.3: Δ^{nn} values in Example 6.21

But as we have already shown in Example 5.27, Φ is neither pn*-monotonic nor* pn*-unambiguous with respect to $\mathcal{C}$ and $\mathcal{D}_{\mathcal{X}}$.*

Finally, $\mathcal{C}$ is $\mathrm{PAC}_{0,0}$*-learnable with respect to Φ and $\mathcal{D}_{\mathcal{X}}$ using Theorem 6.13 because Φ is smooth with respect to $\mathcal{C}$ and $\mathcal{D}_{\mathcal{X}}$.*

Using Example 6.21, we can conclude the following corollary.

Corollary 6.22. *There exists a deterministic label noise model Φ and a concept class $\mathcal{C}$ such that $\mathcal{C}$ is* $\mathrm{PAC}_{0,0}$*-learnable by minimum* nn*-disagreement strategy with respect to Φ and $\mathcal{D}_{\mathcal{X}}$ but not by minimum* pn*-disagreement with respect to Φ and $\mathcal{D}_{\mathcal{X}}$.*

We can use Theorem 6.13 and Proposition 5.24 to conclude the following corollary.

Corollary 6.23. *For the deterministic noise models Φ introduced in Chapter 3, there exists a concept class $\mathcal{C}$ of finite* VC*-dimension and a distribution D such that Φ is* nn*-ambiguous with respect to $\mathcal{C}$ and $\mathcal{D}$.*

Also, using Corollary 6.23 and the contrapositive of Proposition 6.16, we can conclude the following corollary.

Corollary 6.24. *For the deterministic noise models Φ introduced in Chapter 3, there exists a concept class $\mathcal{C}$ of finite* VC*-dimension and a distribution D such that Φ is not* nn*-monotonic with respect to $\mathcal{C}$ and $\mathcal{D}$.*

Chapter 7

Related Work

While previous models of learning (*e.g.,* [16]) require the learner to exactly determine the target concept but allow the learner to run in unbounded time, Valiant's PAC model of learning [37] requires the learner to return a concept which is a close approximation of the target concept in a time-efficient way [13]. However, one of the criticisms of the PAC model is the assumption that the examples that the learner receives from the oracle are noise-free [13]. To compensate, many noise models (*e.g.,* [2, 31, 37]) have been introduced for PAC-learning. In this thesis, we divide the noise models for PAC-learning into two groups. The first type of noise is when the instances of examples are not noisy but the labels can be flipped by the noisy oracle (*e.g.,* [2, 14, 32]). We call this type of noise label noise (see Definition 2.6). The second type of noise is when the instances can be distorted by the noisy oracle (*e.g.,* [17, 23, 37]). As we will see in Section 7.2.2, in this type of noise the labels may or may not be distorted by the oracle.

This chapter is organized as follows. First, the model of learning from statistical queries is introduced. Although not a noise model itself, the statistical query model can be used to prove various learnability results in the PAC-learning framework with noise. In the next section, different noise models in PAC-learning are discussed in more detail. Finally, the last section of the chapter is about noise models outside of the PAC-learning framework.

7.1 Statistical query model

Kearns [22] introduced a model of learning called learning from statistical queries. In this model the standard PAC-learning oracle, EX, is replaced by a weaker oracle called STAT. Rather than returning individual examples like EX, STAT provides accurate estimates of probabilities over the sample space generated by EX.

Let a query $\mathcal{Q}$ be a function $\mathcal{Q} : \mathcal{X} \times \{0,1\} \rightarrow \{0,1\}$ and $\tau \in (0,1]$ be a parameter called tolerance. For any concept c and distribution D, $\mathrm{STAT}_{\mathcal{Q},\tau}(c, D)$ returns an estimate of the probability that $\mathcal{Q}(x, c(x)) = 1$ when x is drawn according to D. This estimate deviates at most by τ from the actual value, $Pr_{x\sim D}[\mathcal{Q}(x, c(x)) = 1]$. Chernoff bounds (see Lemma A.1 in the Appendix) immediately imply that EX with high probability can simulate STAT by estimating the probability that $\mathcal{Q}(x, c(x)) = 1$ using $O(\frac{1}{\tau^2})$ examples. Therefore the statistical query model is a restriction of the PAC model.

A concept class $\mathcal{C}$ is efficiently learnable from statistical queries if there exists a learning algorithm $\mathcal{L}$ such that for any target concept $c^* \in \mathcal{C}$, for any distribution $D \in \mathcal{D}_{\mathcal{X}}$, and for any $\epsilon \in (0, \frac{1}{2})$, the following holds: if $\mathcal{L}$ is given input ϵ and access to $\mathrm{STAT}(c^*, D)$, then *(1)* for every query $(\mathcal{Q}, \tau)$ made by $\mathcal{L}$, $\mathcal{Q}$ can be evaluated in time polynomial in $\frac{1}{\tau}$ where $\frac{1}{\tau}$ is bounded by a polynomial in $\frac{1}{\epsilon}$, and *(2)* $\mathcal{L}$ will halt in time bounded by a polynomial[1] in $\frac{1}{\epsilon}$ and output a a concept $c \in \mathcal{C}$ that satisfies

[1] More accurately, all the polynomials in the definition of learnability from statistical queries should also be polynomial in the size of the representation of the target concept and also the complexity parameter of the concept class as introduced in

$Pr_{x\sim D}[c(x) \neq c^*(x)] \leq \epsilon$ [22].

Kearns [22] showed that any learning algorithm that is based on statistical queries can be automatically converted to a learning algorithm in the PAC framework. Therefore, any concept class that is efficiently learnable from statistical queries is also PAC-learnable.

The statistical query model of learning is important because virtually all known PAC-learning algorithms in the presence of noise can be either obtained from statistical query model algorithms or can be easily cast into a problem in the statistical query model. This is discussed in Section 7.2 with more details for specific noise models.

7.2 Noise in the PAC-learning framework

As we said earlier in this chapter, we distinguish between two different types of noise in PAC-learning in this thesis. The next two sections introduce some models from each of these types.

7.2.1 Label noise models in the PAC-learning framework

Angluin and Laird [2] introduced the first label noise model. In this noise model, called the η-random classification noise, the label of each example is subject to being flipped with some fixed but unknown probability $\eta < \frac{1}{2}$ (see Definition 2.13).

η is limited to be strictly less than a half in the random classification noise model. Clearly, when $\eta = \frac{1}{2}$ the oracle will not convey more information to the learner about the label of examples than an unbiased coin. When $\eta > \frac{1}{2}$, however, there is still information about the target concept, but it is equal to the information about the complement of the target concept with $\eta' = 1 - \eta < \frac{1}{2}$. If the learner knows a priori that $\eta > \frac{1}{2}$ it can flip the label of all examples and use the new examples to learn an approximation of the target concept. Also this situation can be recognized in concept classes that are not closed under complement [2].

Angluin and Laird [2] showed that any finite concept class is $\mathrm{PAC}_{0,0}$-learnable with respect to random classification noise and $\mathcal{D}_{\mathcal{X}}$. In their method first they assumed the learner is provided with an upper bound of strictly less than a half on η. They showed that having this upper bound, instead of the exact noise rate itself, is sufficient for their learning algorithm to be able to learn in the presence of random classification noise. Then, they showed that this upper bound can be estimated from the sample. Later, Laird in his PhD thesis [27] showed that any concept class of finite VC-dimension is $\mathrm{PAC}_{0,0}$-learnable with respect to random classification noise and $\mathcal{D}_{\mathcal{X}}$.

Kearns [22] showed that using a sample of $O(\frac{1}{\tau^2(1-2\eta)^2}\log(\frac{1}{\delta}))$ examples, $\mathrm{EX}_{\Phi_{\mathrm{rcn}(\eta)}}$ can simulate STAT with probability at least $1-\delta$ by estimating any query, $Pr_{x\sim D}[\mathcal{Q}(x, c(x)) = 1]$. Therefore, any learning algorithm that is based on statistical queries can be automatically converted to a learning algorithm in the presence of η-random classification noise for any $\eta < \frac{1}{2}$. Thus, any concept class that is efficiently learnable from statistical queries is also $\mathrm{PAC}_{0,0}$-learnable with respect to random classification noise and $\mathcal{D}_{\mathcal{X}}$.

The second label noise model was introduced by Sloan [31]. In this noise model, called η-malicious classification noise, with unknown probability $\eta < \frac{1}{2}$ on each example, the adversary decides whether to flip the label of the example or not before returning it to the learner. Otherwise, the correct example will be returned to the learner (see Definition 2.18).

Sloan [31] showed that the malicious classification noise model is weaker than the random classification noise model, in the sense that in malicious classification noise the label of at most a fraction η of the examples will be flipped as opposed to random classification noise where the label of exactly a fraction η of the examples will be flipped. Using this simple observation, Sloan proved that finite concept classes are $\mathrm{PAC}_{0,0}$-learnable with respect to malicious classification noise and $\mathcal{D}_{\mathcal{X}}$ [31].

In both of the above models, it is assumed that the probability with which the label of any example is flipped (this probability is defined as the noise rate in Definition 2.8) is constant and,

footnote 3 in Chapter 2.

therefore, independent of the example. Although having a constant noise rate is more realistic than having no noise at all, the noise rate is usually not constant in real world data [8]. It has been shown that learning when the noise rate varies among examples is a difficult problem in general [2, 14]. However, specific kinds of variable noise rate have been studied in the literature.

The very first variable label noise model was proposed by Decatur [14]. In this noise model, called the constant partition classification noise (CPCN), $\mathcal{X} \times \{0,1\}$ is divided into a finite number of partitions. The noise model in each partition is an η-random classification noise with $\eta < \frac{1}{2}$ but η can be different in different partitions (see Definition 2.16).

Later, Ralaivola, Denis and Magnan [29] showed that any concept class that is $\mathrm{PAC}_{0,0}$-learnable with respect to random classification noise and $\mathcal{D}_{\mathcal{X}}$ is also $\mathrm{PAC}_{0,0}$-learnable with respect to CPCN and $\mathcal{D}_{\mathcal{X}}$ and vice versa.

Kearns [22] introduced another variable label noise model called η-variable classification noise. In this model, first, an infinite sequence of noise rates $\eta_1, \ldots, \eta_m, \ldots$ with $\eta_i \in [0,1]$ for all i is fixed by an adversary in advance. The only restriction on this sequence is that for any $m \in \mathbb{N}$, $\frac{1}{m}\sum_{i=1}^{m} \eta_i \leq \eta$ where $\eta < \frac{1}{2}$. Then, a sample with the number of examples requested by the learner is drawn from a noise free oracle and for the i^{th} example in the sample with probability η_i the label will be flipped before the example is returned to the learner.

Kearns [22] showed that any concept class that is $\mathrm{PAC}_{0,0}$-learnable with respect to η-random classification noise and $\mathcal{D}_{\mathcal{X}}$ is also $\mathrm{PAC}_{0,0}$-learnable with respect to η-variable classification noise and $\mathcal{D}_{\mathcal{X}}$ and vice versa.

It should be mentioned that all the label noise models introduced in this section can be represented by our model of label noise (see Definition 2.6) except the η-variable classification noise. The latter cannot be represented in our model of label noise because the noise function in the η-variable classification noise depends on the sample sequence instead of the instances themselves.

7.2.2 Other noise models in the PAC-learning framework

The main difference when the instances are subject to noise (regardless of whether the label is being flipped or not) compared to when the noise is only in the labels, is that the distribution of training instances is different than the distribution of testing instances.

One of the very first noise models that considers only distortion in instances (and not in the labels) was proposed by Goldman and Sloan [17]. This noise model, called the η-uniform random attribute noise, is designed for the case that the input space is a subset of $\{0,1\}^n$ for some $n \in \mathbb{N}$. An instance can therefore be regarded as a vector of n bits. Under uniform random attribute noise, with some fixed but unknown probability $\eta < \frac{1}{2}$ each bit of any instance will be flipped. Then the distorted instance along with the label of the undistorted instance is returned to the learner.

Goldman and Sloan [17] introduced an algorithm that sample-efficiently $\mathrm{PAC}_{0,0}$-learns the class of monomials over n variables with respect to η-uniform random attribute noise for any $\eta < \frac{1}{2}$.

They also introduced another instance noise model, called the η-product random attribute noise, for the case that the input space is a subset of $\{0,1\}^n$ for some $n \in \mathbb{N}$ [17]. In this model the noise rate is an unknown vector of size n, $(\eta_1, \ldots, \eta_n)$, where $\eta_i \leq \eta$ for all i. Under product random attribute noise, the i^{th} bit of the instance of any example is flipped with probability η_i. Then the distorted instance along with the label of the undistorted instance is returned to the learner. Therefore, uniform random attribute noise can be considered as a special case of product random attribute noise when all the elements of the noise rate vector are equal.

Goldman and Sloan [17] showed that any distinct[2] concept class over $\{0,1\}^n$ is $\mathrm{PAC}_{\frac{\eta}{2},0}$-learnable with respect to η-product random attribute noise and $\mathcal{D}_{\mathcal{X}}$.

In many other noise models, not only the instance but also the label of the examples will be flipped. The first such model is proposed by Valiant [38]. In this model, called the η-malicious

[2]A concept class $\mathcal{C}$ is distinct if there exist two concepts $c, c' \in \mathcal{C}$ and $x_1, x_2, x_3, x_4 \in \mathcal{X}$ such that $x_1 \in c, x_1 \in c', x_2 \notin c, x_2 \in c', x_3 \in c, x_3 \notin c', x_4 \notin c, x_4 \notin c'$, *i.e.*, there exist two concepts in the concept class that are not subset of each other, have non-empty intersection and their union is not the whole input space.

noise, with an unknown but fixed probability $\eta < \frac{1}{2}$, an example is returned to the learner about which no assumption whatsoever can be made. With probability $1 - \eta$ the learner will receive the correct example.

In that paper, Valiant proposed an algorithm for learning Boolean formulas in conjunctive normal form. He then showed that his algorithm can be modified to tolerate η-malicious noise for very small η compared to the maximum error that the concept returned by the learning algorithm is allowed to have.

Later, Kearns and Li [23] studied the η-malicious noise model in more detail. They showed that any concept class of finite VC-dimension is $\mathrm{PAC}_{\frac{\eta}{1-\eta},0}$-learnable with respect to η-malicious noise and $\mathcal{D}_{\mathcal{X}}$.

A more powerful adversarial noise model was later proposed by Bshouty *et al.* [8]. In this model, called the η-nasty noise model, an adversary gets to see the whole sample and then chooses to distort E examples, where E is a random variable distributed by the binomial distribution with parameters η and the size of the sample. As in malicious noise, about these E distorted examples no assumption whatsoever can be made. In their paper, Bshouty *et al.* [8] argued that the η-nasty noise model not only generalized some previous noise models, including random classification noise and CPCN, but also, it could model real world situations better than those models.

Bshouty *et al.* [8] showed that any concept class of finite VC-dimension is $\mathrm{PAC}_{2\eta,0}$-learnable with respect to η-nasty noise and $\mathcal{D}_{\mathcal{X}}$.

More complex noise models can be produced by combining two different noise models. These noise models are usually called hybrid noise models. One such hybrid noise model is proposed by Decatur [13]. In this model, called (η, η')-classification and malicious (CAM) noise each example is exposed to random classification noise with probability η and malicious noise with probability η'. Therefore, for each example with probability of $1 - \eta - \eta'$ the undistorted example will be returned to the learner.

Decatur [13] showed that the algorithm of Angluin and Laird [2] for random classification noise can be used to learn a finite concept class with respect to (η, η')-CAM noise when[3] $\eta < \frac{1}{2}$ and $\eta' < \frac{\epsilon}{4}(\frac{1}{2} - \eta)$. He then showed that learning algorithms in the statistical query model can further improve the maximum malicious noise rate η' in CAM to $\eta' < \tau(\frac{1}{2} - \eta)$ where τ is the tolerance of the learning algorithm.

A different kind of noise in PAC-learning model, known as distribution noise, was first introduced by Bartlett [4]. He proposed a distribution noise model in which the training instances are drawn from a distribution while the testing instances may be drawn from a different distribution [4]. However, the learner always receives examples with the correct label. Decatur [12] also introduced three different distribution noise models in which the distribution that the training and testing instances are drawn from may change during sampling. The reader is referred to their papers [4, 12] for more details.

7.3 Noise outside the PAC-learning framework

Learning from a noisy oracle has been considered in other learning frameworks as well. One such framework, which has become popular recently, is the active learning framework [1, 11]. Active learning can be considered as a variant of PAC-learning in which the learner can interactively choose instances from the input space that then will be labeled by the oracle. The term passive learning is then used to describe the normal PAC-style learning models based on $i.i.d.$ sampling. The goal of active learning is to (exponentially) decrease the sample complexity of passive learning by this different sampling technique. The first active learning method, called selective sampling, was introduced by Cohn, Atlas and Ladner [11]. Selective sampling can be considered as an extension of minimum pn-disagreement strategies in the absence of noise. Later, the effect of noise in active learning was studied by Kääriäinen [21] who found lower bounds on the sample complexity

[3] With the additional assumption that at least one $\frac{\epsilon}{2}$-good concept exists in the concept class.

of active learning under random classification noise. He proved that active learning can result in an exponential reduction in sample complexity compared to passive learners in the presence of random classification noise if the noise is not-persistent[4]. Later, Balcan, Beygelzimer and Langford [3] proposed the first active learning algorithm that is robust to arbitrary types of noise. They showed that their algorithm, called A^2, never uses more examples than passive learners. A^2 can be considered as a robust version of selective sampling. Finally, recently Hanneke [19] proposed the first lower and upper bounds on the sample complexity of active learning with persistent random classification noise.

Tsybakov [36] proposed a noise condition to make learnability easier. His condition bounds the probability mass of the areas in the input space which that noise rates very close to the information-theoretic bound of a half. Tsybakov showed that under this condition, the empirical risk converges surprisingly quickly to the risk of the concept that minimizes the risk in the concept class. His noise condition has been widely used (see *e.g.,* [5, 10, 33]).

Noise has been also studied in exact learning models. In exact learning, as opposed to models like PAC-learning, the learning algorithm is required to identify the target concept, rather than approximating it with high probability [19]. Grabowski [18] and Stephan [34] considered many variants of noise in the recursion-theoretic framework of Gold's [16] model of identification in the limit. Within the grammatical inference framework of identification in the limit, noise was modeled and analyzed by Tantini, de la Higuera and Janodet [35]. Both these exact learning frameworks are beyond the scope of this thesis.

Finally, the effect of noise in data has been studied in practical settings as well. The reader is referred to [26, 28, 41] for more details.

[4]In not-persistent random classification noise, the learner can potentially receive different labels by requesting the label of an instance twice.

Chapter 8

Conclusions

8.1 Summary

In this thesis, we considered the problem of PAC-learning in the presence of label noise. We introduced a framework for label noise in which any label noise model that depends on a specific instance, specific target concept in the concept class, and a distribution over the input space can be modeled. We also proposed a generalized definition of learning in the presence of label noise. We showed that almost all previously studied label noise models could be modeled in this framework and their learnability results could be interpreted using our learnability definition.

We also proposed four new label noise models. We claim that these models are a step towards more realistic noise models compared to the previously studied noise models. One way to interpret this is by considering the simple fact that in any classification task some instances are harder to label than other instances. Intuitively, we expect instances closer to the decision boundary to be harder to classify than instances that are further from the decision boundary. Our ball noise models constitute one way of capturing this property.

However, like any other problem, there is a trade-off between the complexity of the noise model and the ease of mathematically analyzing it. Not surprisingly, when the model is getting closer to real world scenarios, it is more difficult to provide mathematical support for it. In spite of this, we showed some learnability results for our noise models for the simple concept class of one-dimensional threshold functions. We also showed that learnability may still be difficult even for concept classes with high similarity to the class of one-dimensional threshold functions, like the class of two-dimensional axis-parallel threshold functions.

The minimum pn-disagreement strategy is the algorithm that has been used widely in the past for proving learnability results in the presence of noise (and even in the noise-free case). As has previously been shown in the literature ($e.g.$, in the article by Angluin and Laird [2]), there exists a combination of noise model, concept class and distribution in which the concept returned by the minimum pn-disagreement strategy has arbitrarily large error. One situation in which this can happen is when the noise rates for some of the instances are greater than or equal to a half for a specific combination of concept class and distribution. Therefore, many previously studied noise models limit themselves to situations in which the noise rate is strictly less than a half for all instances. We briefly looked at the effect of having instances with noise rate greater than or equal to a half in the context of finite input spaces in Chapter 4, providing new general learning results (Propositions 4.3 and 4.4).

The inability of minimum pn-disagreement to deal with such problems brings up the problem of finding more general strategies. We proposed the minimum nn-disagreement strategy as such a generalization. Instead of computing the disagreement between the sample and a concept, the minimum nn-disagreement strategy computes the disagreement between the sample and the noisy version of the concept. One of the advantages of this technique is that under this new setting the

target concept always minimizes the nn-disagreement in contrast to pn-disagreement where the target is not necessarily a minimizer.

We introduced a simple example in which PAC-learnability is possible using minimum nn-disagreement strategies but not using minimum pn-disagreement strategies. However, unfortunately, estimating the nn-disagreement is not as straightforward as estimating the pn-disagreement.

Finally, we investigated the general characteristics of a noise model under which minimum pn-disagreement or minimum nn-disagreement strategies can be used to learn specific concept classes. In each of these cases, we showed that it is sufficient if the noise model always guarantees that the target concept has the smallest disagreement value among all concepts in the concept class, with an additional assumption that there is a separation between the disagreement value of any ϵ-good concept and any ϵ-bad concept for different values of ϵ.

8.2 Open problems and future work

As we previously specified, there are some open problems directly mentioned in this thesis. First of all, the question whether the class of one-dimensional threshold functions is $\mathrm{PAC}_{\underline{\epsilon},0}$ learnable with respect to ω-weight random classification noise is open for $\underline{\epsilon} \in (0, \omega)$. Second, in Chapter 5, we conjectured that pn-unambiguity and pn-monotonicity are not equivalent properties for a noise model (the same for nn-unambiguity and nn-monotonicity in Chapter 6). Finally, we showed that in the case of random classification noise, the minimum pn-disagreement strategy and the minimum nn-disagreement strategy are equivalent and we also proposed an example in which the minimum pn-disagreement strategy will fail in learning although the minimum nn-disagreement strategy will succeed. Based on these observations, we believe that any concept class that is learnable using a minimum pn-disagreement strategy with respect to a noise model Φ is also $\mathrm{PAC}_{0,0}$-learnable with respect to Φ using a minimum nn-disagreement strategy but a proof of this claim is still open.

In addition to the open problems already mentioned in the thesis, we suggest further extensions of this work. First, our attempt to design more realistic noise models did not yield strong PAC-learnability results. We showed there exists a combination of concept class and distribution in which the concept class is not PAC-learnable with respect to our noise models. However, the general question of which specific combinations of distribution and concept class guarantee PAC-learnablity with respect to our noise models is both an interesting and a reasonable question. Additionally, one should study the question of which changes can be applied to our noise models so that learning of interesting concept classes would be possible.

Second, we can study the applications of our new label noise models in other learning models like the model of active learning. We expect that the sampling technique of active learning, by focusing on the more interesting parts of the input space, will help in the learning process, especially in combination with our deterministic noise models. Also as mentioned in the related work chapter, the statistical query model of learning can be used to prove PAC-learnability results in the presence of noise. The question of studying our label noise model (or even any noise model that can be modeled in our label noise framework) in a statistical query setting seems like a potentially insightful direction of extending our work.

Finally, we can investigate other properties of noise models that provide sufficient learnability conditions for specific concept classes when a general learning strategy is being used. Furthermore, the question of whether the properties mentioned in this thesis for the noise models (like pn-unambiguity, nn-unambiguity etc.) are also necessary conditions for learning specific concept classes in the PAC model is open.

8.3 Final word

Finally, we briefly list the most important contributions of this thesis as *(i)* introducing a framework for unifying different label noise models, *(ii)* proposing four new label noise models that are more

similar to noise in real world applications and providing initial learnability results for them and finally *(iii)* suggesting a new general leaning algorithm for PAC-learning with respect to label noise along with studying the desirable characteristic of label noise models.

Bibliography

[1] Dana Angluin. Queries and concept learning. *Machine Learning*, 2(4):319–342, 1988.

[2] Dana Angluin and Philip D. Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

[3] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning (ICML'06)*, pages 65–72. ACM, 2006.

[4] Peter L. Bartlett. Learning with a slowly changing distribution. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT'92)*, pages 243–252. ACM, 1992.

[5] Bartlett, Peter L. and Jordan, Michael I. and McAuliffe, Jon D. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, March 2006.

[6] Gyora M. Benedek and Alon Itai. Learnability with respect to fixed distributions. *Theoretical Computer Science*, 86(2):377–389, 1991.

[7] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Classifying learnable geometric concepts with the Vapnik-Chervonenkis dimension. In *Proceedings of the 18th Annual ACM Symposium on Theory of Computing (STOC'86)*, pages 273–282, 1986.

[8] Nader H. Bshouty, Nadav Eiron, and Eyal Kushilevitz. PAC learning with nasty noise. *Theoretical Computer Science*, 288(2):255–275, 2002.

[9] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–509, 1952.

[10] Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and scoring using empirical risk minimization. In *Proceedings of the 18th Annual Conference on Computational Learning Theory (COLT'05)*, pages 1–15, 2005.

[11] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

[12] Scott E. Decatur. Statistical queries and faulty PAC oracles. In *Proceedings of the 6th Annual Conference on Computational Learning Theory (COLT'93)*, pages 262–268, New York, NY, USA, 1993. ACM.

[13] Scott E. Decatur. Learning in hybrid noise environments using statistical queries. In *Learning from Data: Artificial Intelligence and Statistics*, pages 175–185. Springer Verlag, 1995.

[14] Scott E. Decatur. PAC learning with constant-partition classification noise and applications to decision tree induction. In *Proceedings of the 14th International Conference on Machine Learning (ICML'97)*, pages 83–91. Morgan Kaufmann Publishers Inc., 1997.

[15] William Feller. *An Introduction to Probability Theory and Its Applications*. Wiley, 1968.

[16] Mark E. Gold. Language identification in the limit. *Information and Control*, 10:447–474, 1967.

[17] Sally Goldman and Robert H. Sloan. Can PAC learning algorithms tolerate random attribute noise? *Algorithmica*, 14:70–84, 1995.

[18] Jan Grabowski. Inductive inference of functions from noised observations. In *Proceedings of the International Workshop on Analogical and Inductive Inference (AII'97)*, volume 265 of *Lecture Notes in Computer Science*, pages 55–60. Springer, 1987.

[19] Steve Hanneke. Teaching dimension and the complexity of active learning. In *Proceedings of the 20th Annual Conference on Learning Theory (COLT'07)*, pages 66–81. Springer-Verlag, 2007.

[20] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[21] Matti Kääriäinen. Active learning in the non-realizable case. In *Proceedings of the 17th Annual Conference on Algorithmic Learning Theory (ALT'06)*, pages 63–77, 2006.

[22] Michael J. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998.

[23] Michael J. Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM Journal on Computing (SICOMP)*, 22:267–280, 1993.

[24] Michael J. Kearns and Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. In *Proceedings of the 31st Annual Symposium on Foundations of Computer Science (SFCS'90)*, pages 382–391. IEEE Computer Society, 1990.

[25] Michael J. Kearns and Umesh V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, 1994.

[26] Taghi M. Khoshgoftaar and Jason Van Hulse. Empirical case studies in attribute noise detection. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 39(4):379–388, 2009.

[27] Philip D. Laird. *Learning from Good and Bad Data*. Kluwer Academic Publishers, 1988.

[28] David F. Nettleton, Albert Orriols-Puig, and Albert Fornells. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial Intelligence Review*, 33(4):275–306, 2010.

[29] Liva Ralaivola, François Denis, and Christophe N. Magnan. CN = CPCN. In *Proceedings of the 23rd International Conference on Machine Learning (ICML'06)*, pages 721–728. ACM, 2006.

[30] Walter Rudin. *Functional Analysis*. McGraw-Hill, 1973.

[31] Robert H. Sloan. *Computational Learning Theory: New Models and Algorithms*. PhD thesis, Massachusetts Institute of Technology, 1989.

[32] Robert H. Sloan. Four types of noise in data for PAC learning. *Information Processing Letters*, 54(3):157–162, 1995.

[33] Ingo Steinwart and Clint Scovel. Fast rates for support vector machines. In *Proceedings of the 18th Annual Conference on Learning Theory (COLT'05)*, volume 3559 of *Lecture Notes in Computer Science*, pages 279–294. Springer, 2005.

[34] Frank Stephan. Noisy inference and oracles. *Theoretical Computer Science*, 185:129–157, 1997.

[35] Frédéric Tantini, Colin de la Higuera, and Jean-Christophe Janodet. Identification in the limit of systematic-noisy languages. In *Proceedings of the 8th International Colloquium on Grammatical Inference: Algorithms and Applications (ICGI'06)*, volume 4201 of *Lecture Notes in Computer Science*, pages 19–31. Springer, 2006.

[36] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32:135–166, 2004.

[37] Leslie G. Valiant. A theory of the learnable. In *Proceedings of the 16th Annual ACM Symposium on Theory of Computing (STOC'84)*, pages 436–445. ACM, 1984.

[38] Leslie G. Valiant. Learning disjunctions of conjunctions. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence (IJCAI'85)*, pages 560–566. Morgan Kaufmann Publishers Inc., 1985.

[39] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[40] Vladimir N. Vapnik and Alexey. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971.

[41] Xingquan Zhu and Xindong Wu. Class noise vs. attribute noise: a quantitative study of their impacts. *Artificial Intelligence Review*, 22(3):177–210, 2004.

Appendix A

Some Tools for Probabilistic Analysis

In this appendix, we list the definitions and lemmas that are used throughout this thesis without proofs. The reader is referred to the references mentioned at the beginning of each lemma for proof and more details.

The *Hoeffding inequalities* bound the probability that a random variable deviates from its mean.

Lemma A.1. ***(Hoeffding [20])*** *Let $X_1, \ldots, X_n$ be independent random variables with $E[X_i] \in [0,1]$ for $1 \leq i \leq n$. Let $S = \sum_{i=1}^{n} X_i$ and $s \geq 0$. Then*

$$Pr[\frac{S}{n} - \frac{E[S]}{n} \geq s] \leq e^{-2s^2 n},$$

and

$$Pr[\frac{S}{n} - \frac{E[S]}{n} \leq s] \leq e^{-2s^2 n}.$$

In particular, if for all $1 \leq i \leq n$, $E[X_i] = p$ with $p \in [0,1]$

$$Pr[S \geq (p+s)n] = Pr[S - pn \geq sn] = Pr[\frac{S}{n} - p \geq s] \leq e^{-2s^2 n},$$

and

$$Pr[S \leq (p-s)n] = Pr[S - pn \leq sn] = Pr[\frac{S}{n} - p \leq s] \leq e^{-2s^2 n}.$$

Sometimes, these inequalities are referred to by the name of *Chernoff bounds* because Chernoff [9] first discovered them.

Angluin and Laird introduced the following notation, which indicates the probability of observing at least (at most) a specific number of heads in m times flipping a biased coin with probability p of coming up heads on each flip.

Definition A.2. ***(Angluin and Laird [2])*** *Let C be a biased coin with probability $p \in [0,1]$ of heads. Let $q \in (0,1)$ and $m \in \mathbb{N}$. Let $\mathrm{GE}(p,m,q)$ denote the probability of observing at least $\lfloor qm \rfloor$ heads in m independent flips of C, $i.e.$,*

$$\mathrm{GE}(p,m,q) = \sum_{i=\lfloor qm \rfloor}^{m} \binom{m}{i} p^i (1-p)^{m-i}. \tag{A.1}$$

Similarly, let $\mathrm{LE}(m,p,q)$ denote the probability of observing at most $\lfloor qm \rfloor$ heads in m independent flips of C, $i.e.$,

$$\mathrm{LE}(p,m,q) = \sum_{i=0}^{\lfloor qm \rfloor} \binom{m}{i} p^i (1-p)^{m-i}. \tag{A.2}$$

The Hoeffding inequalities (Lemma A.1) can be used to bound the probabilities introduced in Definition A.2.

Lemma A.3. ***(Angluin and Laird [2])*** *Let $s \in (0,1)$ and $s = |q - p|$. Then*

$$\mathrm{GE}(p, m, p + s) \leq e^{-2s^2 m}, \tag{A.3}$$

and

$$\mathrm{LE}(p, m, p - s) \leq e^{-2s^2 m}. \tag{A.4}$$

Setting the right hand side of Equations A.3 and A.4 equal to δ and solving for m we get the following lemma.

Lemma A.4. ***(Angluin and Laird [2])*** *Let $p \in [0,1]$ and $s, \delta \in (0,1)$. If*

$$m \geq \frac{1}{2s^2} \ln(\frac{1}{\delta})$$

then

$$\mathrm{LE}(p, m, p - s) \leq \delta,$$

and

$$\mathrm{GE}(p, m, p + s) \leq \delta.$$

Finally, the coupon collector problem is a classic problem in probability theory that can be found in many textbooks of the field like the one by Feller [15]. We state the lemma from the article by Benedek and Itai [6].

Lemma A.5. ***(Coupon Collector Problem) (Benedek and Itai [6])*** *Let $A_1, \ldots, A_k$ be events with probability of greater than or equal to p. Then in a sequence of*

$$m = \frac{1}{p} \ln(\frac{k}{\delta}) \tag{A.5}$$

independent trials, the probability that every event occurrs at least once is greater than $1 - \delta$.

Index