PAC-Learning with General Class Noise Models

Shahin Jabbari[1], Robert C. Holte[2], and Sandra Zilles[3]

[1] University of California, Los Angeles, shahin@cs.ucla.edu
[2] University of Alberta, holte@cs.ualberta.ca
[3] University of Regina, zilles@cs.uregina.ca

Abstract. We introduce a framework for class noise, in which most of the known class noise models for the PAC setting can be formulated. Within this framework, we study properties of noise models that enable learning of concept classes of finite VC-dimension with the Empirical Risk Minimization (ERM) strategy. We introduce simple noise models for which classical ERM is not successful. Aiming at a more general-purpose algorithm for learning under noise, we generalize ERM to a more powerful strategy. Finally, we study general characteristics of noise models that enable learning of concept classes of finite VC-dimension with this new strategy.

1 Introduction

Modeling noise in learning is a problem that has been widely addressed in the literature. Specific noise models have been formalized and studied with respect to their effect on learnability. Unfortunately, often noise models with strong positive learnability results are rather unrealistic models, whereas more realistic noise models leave little room for positive results. This trade-off has not been studied systematically—almost every previous study focuses on a specific noise model and produces results only for that model. To address this shortcoming, this paper provides a formal framework in which we can reason about a broad class of noise models, and presents quite general conditions on noise models in this class under which learnability in the PAC model [16] can be guaranteed.

The focus of this paper is on *class noise* (*e.g.*, [1]), which allows the labels of the examples given to the learner to be altered by noise, but not the instances themselves to be altered (in contrast to other types of noise, *e.g.*, [7]). In the class noise setting, for an instance x from input space $\mathcal{X}$, a distribution D over $\mathcal{X}$, and a target concept c, the *noise rate* of x given D and c is the probability that the wrong label for x is observed, given that x is sampled with respect to D.

Classical noise models, such as random classification noise [1], malicious classification noise [14], and constant partition classification noise (CPCN) [6], are rather restrictive. Random classification noise assumes that every instance x has the same noise rate, the latter being independent of D and c. Malicious classification noise allows different instances to have different noise rates but assumes a common upper bound on all the noise rates, which is independent of D and c. CPCN loosens these constraints by allowing the noise rate to depend on c as

well as x, but not on D. This allows one to model the type of noise that arises in many natural settings when instances closer to the decision boundary have a larger noise rate than instances far away from the decision boundary. However, in CPCN the transition between these noise rates is not smooth, since noise rates are determined by a finite partitioning of the set of all possible labeled examples.

The literature studies these noise models separately. Though the statistical query model [9] gave a unified account of the learnability results of various noise models, it does not permit the definition of new noise models that overcome the limitations of the classical ones or to study general properties of noise that enable PAC-learning of certain concept classes under specific classes of distributions.

We introduce a formal definition of "class noise model" in which many classical models can be formulated. Our flexible framework allows noise rates to depend arbitrarily on x, D, and c. We then focus on the question of what makes learning under some noise models harder than learning under others, and try to gain insight into why all known noise models that produce general positive learnability results are rather unrealistic. We address this question by proposing formal properties *on noise models* under which PAC-learning is possible. Empirical Risk Minimization (ERM) strategies [17], which were previously used to prove that every PAC-learnable class is PAC-learnable under random classification noise, simply return a concept c' that minimizes the number of observed examples whose labels disagree with those of c'. In a noisy setting, this kind of strategy might not be generally successful, since the noise model might obfuscate the differences between concepts, *i.e.*, two dissimilar concepts might look very similar after applying noise, and vice versa. Therefore we generalize ERM to a strategy that picks a concept whose *expected behavior after applying noise* minimizes the number of disagreements with the sample. Under some additional assumptions on the noise model, we show that similar properties as in the classical ERM case are sufficient for PAC-learning with this generalized strategy.

To sum up: As opposed to the research on *agnostic learning*, we study the problem of finding a concept that approximates the underlying noise-free target concept c, instead of approximating the observed (noisy) data. Our results suggest that no *realistic* noise model will lead to a *general* solution to this problem in the distribution-free setting. Our goal is *not* to show that approximating c under severe noise is possible in general, but to study conditions on the noise models under which this is possible. The main contributions of this work are: *(i)* a formal basis for the design and study of new noise models as well as for classes of distributions that ease learning. *(ii)* formal conditions under which ERM still works; *(iii)* a generalization of ERM including conditions under which it solves the learning problem we propose.

2 Preliminaries

We denote by $\mathcal{X}$ a set called the *input space*. For most of this paper, $\mathcal{X} = \mathbb{R}^n$ for some $n \in \mathbb{N}$. A *concept* c is a subset of $\mathcal{X}$ or, equivalently, a binary-valued function on $\mathcal{X}$. A *concept class*, $\mathcal{C}$, is a set of concepts. A *probabilistic concept*

(or a *noisy concept*) $c : \mathcal{X} \to [0,1]$ is a real-valued function that assigns to each element of $\mathcal{X}$ a value in the closed interval $[0,1]$. Hence, a concept can be considered as a special case of a probabilistic concept. Let D denote a probability *distribution* over $\mathcal{X}$ and $\mathcal{D}_{\mathcal{X}}$ denote the set of all distributions over $\mathcal{X}$. For a distribution D and probabilistic concept c, 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 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. If c is a concept, then for every (x, y) returned by $\mathrm{EX}(c, D)$, $y = c(x)$. In any example (x, y), x is called the *instance* and y is called the *label*. Every multi-set $\mathcal{S}$ of examples is called a *sample*. We study learning in the framework of PAC-learning [16].

Definition 1. *[16] A concept class $\mathcal{C}$ is probably approximately correctly learnable (*PAC*-learnable), if there exists a learning algorithm $\mathcal{L}$ and a polynomial $m : \mathbb{R}^2 \to \mathbb{R}$ such that: for any target concept $c \in \mathcal{C}$, for any $\epsilon, \delta \in (0, 1/2)$ 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 $\lceil m(1/\epsilon, 1/\delta) \rceil$ examples, $\mathcal{L}$ outputs a concept $c' \in \mathcal{C}$ satisfying $Pr_{x \sim D}[c'(x) \neq c(x)] \leq \epsilon$.*[4]

One criticism of the PAC model is the unrealistic assumption that the oracle always provides examples according to the true underlying distribution D and the true target concept c. Often in practice information sources are susceptible to noise. Several kinds of noise were proposed to remedy this problem. In our research we focus on class noise, *i.e.*, we assume in the examples returned by the noisy oracle, the instances x given to the learner are drawn with respect to D but with some probability η the labels may sometimes be flipped from $c(x)$ to $1 - c(x)$. η is called the *noise rate* and can vary with the instance, target concept and distribution. Previously studied class noise models were proven not to restrict PAC-learnability. Every PAC-learnable class is also PAC-learnable under a random classification noise oracle [1], a malicious classification noise oracle [14], or a CPCN oracle [13], as long as the noise rates are less than 1/2.

3 A general framework for modeling class noise

Random classification noise and malicious classification noise involve noise rates that do not depend on the sampled instance x or on the target concept. In practice, this is unrealistic, since one might expect examples closer to the decision boundary to be more susceptible to noise than examples farther away [4]. For example, in optical character recognition, training examples for a certain character are more likely to be mislabeled the more similar they are to another character. The CPCN model addresses this issue, but does not allow for a smooth transition between noise rates when traversing the instance space. Moreover, the CPCN model does not allow the noise to depend on the distribution.

[4] Run-time efficiency issues are out of the scope of this paper. Further, note that Definition 1 is only sensible under mild measurability conditions.

One approach could be to introduce new noise models and compare them to existing ones. However, learnability results would then concern only the particular chosen noise models and might not provide much insight into what makes learning under noise difficult in general. Therefore, we abstract from specific noise models and introduce a framework that *(i)* captures most of the class noise models studied in the literature (Section 3.1), *(ii)* allows us to formalize new class noise models (Section 3.2), and *(iii)* allows us to study general properties of noise models that are sufficient or necessary for learnability (Section 4).

3.1 Class noise models

Class noise can be considered as a procedure that converts a concept to a probabilistic concept, because the correct label of an instance may be flipped.

Definition 2. *A* (class) noise model *is a mapping* $\Phi : 2^{\mathcal{X}} \times \mathcal{D}_{\mathcal{X}} \times \mathcal{X} \to [0,1]$.

Thus, noise can depend on the sampled instance x, the target concept c, and the distribution D. For every c and D, each instance x has a defined *noise rate* $\eta_{c,D}(x)$, *i.e.*, a probability with which its label is flipped, namely $\eta_{c,D}(x) = |c(x) - \Phi(c, D, x)|$. For example, random classification noise [1] can be defined by $\Phi(c, D, x) = 1 - \eta$, if $c(x) = 1$, and $\Phi(c, D, x) = \eta$, if $c(x) = 0$ where $\eta \in [0, 1/2)$ is the noise rate. As another example, CPCN [6] can be defined as follows. If $\eta = (\eta_1, \ldots, \eta_k) \in [0, 1/2)^k$, and $\pi = (\pi_1, \ldots, \pi_k) \subseteq (\mathcal{X} \times \{0,1\})^k$ is a k-tuple of pairwise disjoint sets such that $\pi_1 \cup \cdots \cup \pi_k = \mathcal{X} \times \{0,1\}$, then, for $(x, c(x)) \in \pi_i$, $\Phi(c, D, x) = 1 - \eta_i$, if $c(x) = 1$, and $\Phi(c, D, x) = \eta_i$, if $c(x) = 0$.[5]

Sampling according to c and D (via $\mathrm{EX}(c, D)$), followed by applying the noise model Φ, is defined as sampling from the noisy concept $\Phi(c, D, \cdot)$. We then say that a class $\mathcal{C}$ is *learnable w.r.t.* Φ if $\mathcal{C}$ is PAC-learnable as in Definition 1, where the oracle $\mathrm{EX}(c, D)$ is replaced by sampling from the noisy concept $\Phi(c, D, \cdot)$.

PAC-learning is distribution-free, *i.e.*, it requires the learner to be successful for any combination of target concept and underlying distribution. In the presence of noise, distribution-free learning may be difficult, and even impossible for many simple classes (see Proposition 1). Therefore, we sometimes restrict the class of distributions when dealing with noise. For any $\mathcal{D} \subseteq \mathcal{D}_{\mathcal{X}}$, we say $\mathcal{C}$ is learnable w.r.t. Φ and $\mathcal{D}$, if we require the learner to be successful only for distributions in $\mathcal{D}$, not for any distribution in $\mathcal{D}_{\mathcal{X}}$.

In our model, the learner is required to produce a concept that is similar to the target concept before it is corrupted by noise. This is a different task than agnostic learning [11], which requires the learner to find a concept that best

[5] Malicious classification noise [14] cannot be modeled by Definition 2. This can be easily fixed by using a mapping $\Phi : 2^{\mathcal{X}} \times \mathcal{D}_{\mathcal{X}} \times \mathcal{X} \to 2^{[0,1]}$ to a *set of values between 0 and 1*. This generalization allows defining malicious noise in which the adversary has the option of picking the value of Φ from a subset of $[0,1]$ that depends on the instance, the target concept and the distribution. Due to space constraints, we do not discuss such models any further. However, even this generalization cannot model noise that depends on the sequence of examples itself, *e.g.*, [5, 9].

approximates the probabilistic (noisy) concept observed. An extra difficulty of our task arises from the fact that the noise process may generate two similar probabilistic concepts from two dissimilar concepts. In fact, unlike in the agnostic case, a necessary condition for PAC-learnability with any arbitrary error is that the noise model Φ does not "make two distinct concepts equal."

Lemma 1. *Let Φ be a noise model. Let $\mathcal{C}$ be a concept class, $c, c' \in \mathcal{C}$ with $c \neq c'$ and $\mathcal{D} \subseteq \mathcal{D}_{\mathcal{X}}$. If there is some $D \in \mathcal{D}$ such that $\Phi(c, D, x) = \Phi(c', D, x)$ for all $x \in \mathrm{supp}(D)$, then the learner cannot distinguish between c and c' regardless of the number of examples it receives.*

An immediate consequence of Lemma 1 is that it implies a lower error bound of $Pr_{x\sim D}[c(x) \neq c'(x)]/2$ for learning $\mathcal{C}$ w.r.t. Φ and $\mathcal{D}$.

3.2 Defining new noise models

To illustrate the flexibility of our definition of noise, we introduce examples of noise models in which the noise rate depends on the target concept, the instance, and sometimes on the distribution. The first noise model was suggested by Shai Ben-David (personal communication) and is based on the idea that noise is often more likely when an instance lies close to the decision boundary.

In this model, the noise rate for an example $(x, c(x))$ is given by the probability of an instance in the vicinity of x being labeled by $1 - c(x)$, where c is the target concept. In other words, the probability of x being labeled 1 by the oracle equals the probability mass of the set of positively labeled instances in a ball around x, relative to the mass of the whole ball around x. There are different ways of defining the ball around an instance, *e.g.*, the *distance ball* around x is defined as $\mathrm{DB}_\rho(x) = \{x' \in \mathcal{X} \mid \mathrm{dist}(x, x') < \rho\}$ for some metric dist.

Definition 3. *Let $\rho \geq 0$. The ρ-distance random classification noise model, $\Phi^{\mathrm{dr}(\rho)}$, is defined by*

$$\Phi^{\mathrm{dr}(\rho)}(c, D, x) = Pr_{x'\sim D}[c(x') = 1 \mid x' \in \mathrm{DB}_\rho(x)],$$

for $x \in \mathrm{supp}(D)$. $\Phi^{\mathrm{dr}(\rho)}(c, D, x) = 0$ for $x \notin \mathrm{supp}(D)$.

To gain some intuition about this new noise model, we show that the class of linear separators in $\mathbb{R}$ is learnable with respect to $\Phi^{\mathrm{dr}(\rho)}$, where the metric in the definition of the distance ball is the Euclidean distance.

Theorem 1. *Let $\mathcal{X} = \mathbb{R}$ and $\rho \geq 0$. Let $\mathcal{C}$ be the class of linear separators in $\mathbb{R}$. $\mathcal{C}$ is learnable w.r.t. $\Phi^{\mathrm{dr}(\rho)}$.*

Theorem 1 is proven by showing that the noisy concepts $\Phi^{\mathrm{dr}(\rho)}$ are all non-decreasing functions, *i.e.*, the probability of the label for x being 1 never decreases as x increases. Such probabilistic concepts can be approximated, with high probability, in a sample-efficient way [10], which helps to reconstruct the target concept approximately.

The second noise model follows a similar idea about the origin of noise but uses a different definition for the ball around an instance. The *weight ball*, $\mathrm{WB}_\omega(x)$, around an instance x is the largest distance ball that has the mass of at most ω with respect to the distribution *i.e.*, $\mathrm{WB}_\omega(x) = \mathrm{DB}_\rho$ where $\rho = \sup\ \{\rho' \mid Pr_{x' \sim D}[x' \in \mathrm{DB}_{\rho'}(x)] \leq \omega\}$.

Definition 4. *Let $\omega \in [0,1]$. The ω-weight random classification noise model, $\Phi^{\mathrm{wr}(\omega)}$, is defined by*

$$\Phi^{\mathrm{wr}(\omega)}(c, D, x) = \Pr_{x' \sim D}[c(x') = 1 \mid x' \in \mathrm{WB}_\omega(x)],$$

for $x \in \mathrm{supp}(D)$. $\Phi^{\mathrm{wr}(\omega)}(c, D, x) = 0$ for $x \notin \mathrm{supp}(D)$.

The idea behind the weight ball is that the expertise of the expert labeling the examples is built based on the same distribution with respect to which learning takes place. If x is close to the decision boundary, but in a dense area, the expert has more experience in the area around x and is thus less likely to make mistakes than in the case where the area around x is sparse.

In general, the new noise models introduced in this section are restrictive. The proof is based on Lemma 1 and is omitted due to space constraints.

Proposition 1. *For any of the noise models Φ introduced in Section 3.2, there exists a concept class $\mathcal{C}$ of finite* VC*-dimension that is not learnable w.r.t. Φ.*

The criteria for distribution-free learning seem too restrictive though for realistic settings; for example, often the distribution depends on the target concept. Thus, in cases where distribution-free learning is not possible, we have to ask ourselves whether the unrealistic requirements concerning unrestricted distributions are the actual reason for the negative learnability result.

One idea for limiting the distributions was recently proposed [2]. Recall that $f : \mathcal{X} \to \mathbb{R}$ is Lipschitz if $|f(x) - f(x')| \leq \gamma . \mathrm{dist}(x, x')$ for all $x, x' \in \mathcal{X}$, given a fixed $\gamma > 0$ and some metric dist. If f is a concept, the Lipschitz condition would make f constant. Relaxing the definition by requiring the Lipschitz condition to hold with some high probability, we can model situations in which no clear margin between instances with different labels around the boundary exists.

Definition 5. *[2] Let $\psi : \mathbb{R} \to [0,1]$. A function $f : \mathcal{X} \to \mathbb{R}$ is ψ-Lipschitz with respect to a distribution D if for all $\gamma > 0$*

$$Pr_{x \sim D}[\exists\ x' : |f(x) - f(x')| > \gamma . \mathrm{dist}(x, x')] \leq \psi(\gamma).$$

This gives us positive learnability results for classes that are not learnable if we do not limit the distributions, if we don't require that an arbitrarily low error can be achieved.

Theorem 2. *Let $\mathcal{X} = \mathbb{R}^n$ for some $n \in \mathbb{N}$ and $\rho \geq 0$ ($\omega \in [0,1]$). Let $\mathcal{C}$ be the class of linear separators in $\mathbb{R}^n$. Let $\mathcal{D} \subset \mathcal{D}_\mathcal{X}$ such that for all $c \in \mathcal{C}$ and $D \in \mathcal{D}$, c is ψ-Lipschitz with respect to D. Then for all $\gamma > 0$, $\mathcal{C}$ is learnable w.r.t. $\Phi^{\mathrm{dr}(\rho)}$ ($\Phi^{\mathrm{wr}(\omega)}$) and $\mathcal{D}$, with a lower bound of $\psi(\gamma)$ on the error bound.*

3.3 Noise rates different from 1/2

The positive results on learning in the classical noise models discussed above (random classification noise, CPCN, malicious classification noise) assume that the noise rate for any instance is always less than 1/2 unless the noise rates for *all* the instances are always greater than 1/2. (The latter case can be reduced to the former by flipping all the labels.)

The models introduced in Section 3.2 typically do not have this property. Noise rates can be greater than 1/2 for some instance x and less than 1/2 for another instance x', given the same distribution and target concept, or they can be greater than 1/2 for some instance x given a particular distribution D, and less than 1/2 for x under some other distribution $D' \neq D$. However, for finite instance spaces, learning under such noise models is still possible, namely if only the instance determines whether the noise rate is above or below 1/2.

Theorem 3. *Let $\mathcal{X}$ be finite. Let $\mathcal{C}$ be a concept class over $\mathcal{X}$ and Φ a noise model such that $\eta_{c,D}(x) \neq 1/2$ for all $c \in \mathcal{C}$, $D \in \mathcal{D}_{\mathcal{X}}$, and $x \in \mathcal{X}$. If $[\eta_{c,D}(x) > 1/2 \iff \eta_{c',D'}(x) > 1/2]$ for all $c, c' \in \mathcal{C}$, $D, D' \in \mathcal{D}_{\mathcal{X}}$, and $x \in \mathcal{X}$, then $\mathcal{C}$ is learnable w.r.t. Φ.*

The idea behind the proof is that the probabilistic concepts generated by the noise model can be learned by repeatedly sampling a set of instances that contain an arbitrarily large portion of the distribution mass. The assumption that the noise rates are not equal to 1/2 can be relaxed (at the cost of error values no longer approaching zero) if we assume the weight of the area with noise rate close to 1/2 is bounded (*e.g.*, by applying Tsybakov's noise condition [15]).

4 Minimum disagreement strategies

ERM [17] refers to learning algorithms that pick a concept $c' \in \mathcal{C}$ that minimizes the number of examples in the given sample $\mathcal{S} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ that are labeled differently than c' would label them. In the absence of noise, $y_i = c(x_i)$ where c is the target concept. This means ERM picks a $c' \in \mathcal{C}$ that minimizes the empirical error, $1/m \sum_{i=1}^{m} |y_i - c'(x_i)|$. When the sample size grows, this corresponds to minimizing $\mathrm{err}_D(c', c) = Pr_{x \sim D}[c'(x) \neq c(x)]$, *i.e.*, the expected error of c', which is supposed to be kept small in PAC-learning. We call a learning algorithm that uses the ERM principle a *minimum disagreement strategy*. When c and D are clear from the context, we use $\mathrm{err}(c')$ instead of $\mathrm{err}_D(c', c)$ for brevity.

If $\mathcal{C}$ is infinite, it is in general impossible to compute a minimum disagreement strategy. Then an approximation strategy typically reduces $\mathcal{C}$ to a finite set $\mathcal{C}' \subset \mathcal{C}$ such that, for any target concept $c \in \mathcal{C}$, at least one concept $c' \in \mathcal{C}'$ differs from c by at most ϵ, and then applies the minimum disagreement strategy over $\mathcal{C}'$. If the target concept is the unique minimizer of the empirical error, every such approximation strategy is called a minimum disagreement strategy as well. This is used implicitly in the proofs of Theorems 4 and 6.

Given noise, a minimum disagreement strategy (with growing sample size) minimizes the difference between the concept c' and the noisy (probabilistic)

concept $\Phi(c, D, x)$ resulting from the target c when applying the underlying noise model Φ, *i.e.*, $\mathrm{err}_D(c', \Phi(c, D, .)) = E[|c'(x) - \Phi(c, D, x)|]$. When c and D are clear from the context, we use $\mathrm{err}(c', \Phi)$ instead of $\mathrm{err}_D(c', \Phi(c, D, .))$.

Minimum disagreement strategies, in the noise-free PAC case, are always successful for classes of finite VC-dimension [3]. This result carries over to learning from random classification noise [1]. The latter means that finding a concept with low error is accomplished by finding a concept that looks most similar to the noisy version of the target concept *i.e.*, the minimizer of $\mathrm{err}(c, \Phi)$. Obviously, this is not possible in general (see Proposition 2). But if the noise model fulfills some advantageous properties, minimum disagreement strategies still work.

In the following subsection, we analyze properties of class noise models under which minimum disagreement strategies are successful. Since a minimum disagreement strategy in the presence of noise returns the same concept as an agnostic learner, these are properties under which the concept returned by an agnostic learner satisfies the learning criteria in our framework.

4.1 Disagreement between concepts and noisy samples

One desirable property of a noise model is that it won't let two concepts $c, c' \in \mathcal{C}$ appear almost "equally similar" to the noisy version of the target concept, if c is "much more similar" to the target concept than c' is.

Definition 6. *Let $\mathcal{C}$ be a concept class, $\mathcal{D} \subseteq \mathcal{D}_{\mathcal{X}}$ a class of distributions and Φ a noise model. Φ is distinctive with respect to $\mathcal{C}$ and $\mathcal{D}$ if there exist polynomial functions $f : (0, 1/2) \to (0, 1/2)$ and $g : (0, 1/2) \to (0, 1)$ such that for any target concept $c \in \mathcal{C}$, for any $c', \bar{c} \in \mathcal{C}$, $D \in \mathcal{D}$ and $\epsilon \in (0, 1/2)$*

$$\mathrm{err}(c') < f(\epsilon) \wedge \mathrm{err}(\bar{c}) > \epsilon \Rightarrow \mathrm{err}(\bar{c}, \Phi) - \mathrm{err}(c', \Phi) \geq g(\epsilon).$$

An example of a distinctive noise model is random classification noise for any noise rate $\eta < 1/2$: Note that, in this model, $\mathrm{err}(c', \Phi) = \eta + (1 - 2\eta)\,\mathrm{err}(c')$ for all $c' \in \mathcal{C}$ [1]. Then $f(\epsilon) = \epsilon/2$ and $g(\epsilon) = \epsilon(1 - 2\eta)/2$ yield, as soon as $\mathrm{err}(c') < f(\epsilon)$ and $\mathrm{err}(\bar{c}) > \epsilon$, that $\mathrm{err}(\bar{c}, \Phi) - \mathrm{err}(c', \Phi) = (1 - 2\eta)(\mathrm{err}(\bar{c}) - \mathrm{err}(c')) \geq \epsilon(1 - 2\eta)/2 = g(\epsilon)$.

Distinctiveness guarantees learnability of classes of finite VC-dimension (of course, sample bounds are higher in the noisy setting).

Theorem 4. *Let $\mathcal{C}$ be a concept class of finite* VC*-dimension d and Φ a noise model. If Φ is distinctive with respect to $\mathcal{C}$ and $\mathcal{D}_{\mathcal{X}}$ then $\mathcal{C}$ is learnable w.r.t. Φ using a minimum disagreement strategy.*

Proof. A minimum disagreement strategy, $\mathcal{L}$, can learn any concept class of finite VC-dimension in the agnostic setting when the examples are drawn from any joint distribution over $\mathcal{X} \times \{0, 1\}$ [8]. Fix the target concept c, D, and $\delta, \epsilon \in (0, 1/2)$. Let $m(g(\epsilon)/2, \delta, d)$ and c' be the sample complexity and concept returned by $\mathcal{L}$ resp., when the examples are drawn from Φ. By the definition of agnostic learning, $\mathrm{err}(c', \Phi) \leq \min_{\bar{c} \in \mathcal{C}} \ \mathrm{err}(\bar{c}, \Phi) + g(\epsilon)/2$ with probability $\geq 1 - \delta$.

By distinctiveness, $\{c\} = \arg\min_{\bar{c} \in \mathcal{C}} \ \mathrm{err}(\bar{c}, \Phi)$. Thus, $\mathrm{err}(c', \Phi) \leq \mathrm{err}(c, \Phi) + g(\epsilon)/2$. Hence, $\mathrm{err}(c') \leq \epsilon$ because otherwise $\mathrm{err}(c', \Phi) \geq \mathrm{err}(c, \Phi) + g(\epsilon)$, due to distinctiveness. Therefore, learning in the presence of noise is equivalent to agnostic learning under the assumptions of Theorem 4. □

If both the concept class and the collection of distributions are finite, a weaker property can be proven to be sufficient for learning. It simply requires the target concept to always be the unique minimizer of $\mathrm{err}(c', \Phi)$, among all $c' \in \mathcal{C}$. This property is necessary for learning with minimum disagreement strategies, since otherwise, for small enough ϵ, picking the minimizer of the disagreement could result in choosing a concept whose error is larger than ϵ, with high probability.

Definition 7. *Let $\mathcal{C}$ be a concept class, $\mathcal{D} \subseteq \mathcal{D}_{\mathcal{X}}$, and Φ a noise model. Φ is monotonic with respect to $\mathcal{C}$ and $\mathcal{D}$ if for any target concept $c \in \mathcal{C}$, for any $D \in \mathcal{D}$ and for any $c' \in \mathcal{C}$:* $\mathrm{err}(c') > 0 \Rightarrow \mathrm{err}(c', \Phi) > \mathrm{err}(c, \Phi)$.

Monotonicity is implied by distinctiveness, since $g(\epsilon) > 0$ for all ϵ in the definition of distinctiveness. The sufficiency result mentioned above can be formulated as follows. The proof is omitted due to space constraints.

Theorem 5. *Let $\mathcal{C}$ be a finite concept class, $\mathcal{D} \subseteq \mathcal{D}_{\mathcal{X}}$ finite, and Φ a noise model. $\mathcal{C}$ is learnable w.r.t. Φ and $\mathcal{D}$ using a minimum disagreement strategy iff Φ is monotonic w.r.t. $\mathcal{C}$ and $\mathcal{D}$.*

For random classification noise, minimum disagreement strategies are universal, *i.e.*, they are successful for every concept class that is PAC-learnable by *any* other learning algorithm [1]. This is not true for all noise models as stated in Proposition 2. (This result is due to [1], but we give our own proof).

Proposition 2. *There exists a concept class $\mathcal{C}$, a distribution D, and a noise model Φ such that $\mathcal{C}$ is learnable w.r.t. Φ and $\{D\}$, but no minimum disagreement strategy can learn $\mathcal{C}$ w.r.t. Φ and $\{D\}$.*

Proof. Let $\mathcal{X} = \{x_1, x_2\}$, $\mathcal{C} = \{c_1, c_2, c_3\}$ where $c_1 = \{x_1, x_2\}$, $c_2 = \{x_2\}$, and $c_3 = \{x_1\}$. Let $D \in \mathcal{D}_{\mathcal{X}}$ be defined by $Pr_{x \sim D}[x = x_1] = 0.25$ and $Pr_{x \sim D}[x = x_2] = 0.75$. Let Φ be a noise model with $\Phi(c, D, x_1) = |c(x_1) - 0.75|$ and $\Phi(c, D, x_2) = |c(x_2) - 0.25|$ for any $c \in \mathcal{C}$ and suppose c_2 is the target concept. Then $\Phi(c_2, D, x_1) = \Phi(c_2, D, x_2) = 0.75$, $\mathrm{err}(c_1) = 0.25$, $\mathrm{err}(c_3) = 1$, $\mathrm{err}(c_1, \Phi) = 0.25$, $\mathrm{err}(c_2, \Phi) = 0.375$, and $\mathrm{err}(c_3, \Phi) = 0.625$. Since $c_2 \notin \arg\min_{c \in \mathcal{C}} \mathrm{err}(c, \Phi)$ ($\mathrm{err}(c_1, \Phi) = 0.25$ while $\mathrm{err}(c_2, \Phi) = 0.375$), Φ is not monotonic with respect to $\mathcal{C}$ and $\{D\}$ (Φ is not distinctive with respect to $\mathcal{C}$ and $\{D\}$ either.) By Theorem 5, no minimum disagreement strategy can PAC-learn $\mathcal{C}$ w.r.t. Φ and $\{D\}$. □

This proof relies on the noise rates exceeding 1/2, which might well happen in realistic noise models. The noise models defined in Section 3.2 can also yield noise rates greater than 1/2 on parts of the instance space. So far, for noise rates exceeding 1/2, we only dealt with strategies for special cases on finite $\mathcal{X}$ (Theorem 3). The following subsection deals with general strategies for learning under noise in cases where minimum disagreement strategies might fail.

4.2 Disagreement between noisy concepts and noisy samples

Minimum disagreement strategies return a concept c' that minimizes the disagreement with the sample. Thus they ideally minimize $\mathrm{err}(c', \Phi)$, *i.e.*, the difference between c' and the noisy target concept. However, our goal is to return a concept that minimizes $E[|\Phi(c', D, x) - \Phi(c, D, x)|]$, *i.e.*, whose *noisy version* is similar to the noisy target concept. When the target concept and the distribution are clear from the context, with a slight abuse of notation, we use $\mathrm{err}(\Phi(c'), \Phi)$ to denote $E[|\Phi(c', D, x) - \Phi(c, D, x)|]$.[6]

Note that the target concept, c, always minimizes $\mathrm{err}(\Phi(c'), \Phi)$ among all $c' \in \mathcal{C}$, since $\mathrm{err}(\Phi(c), \Phi) = E[|\Phi(c, D, x) - \Phi(c, D, x)|] = 0$. This is not the case for $\mathrm{err}(c', \Phi)$ (see the proof of Proposition 2).

A natural strategy for minimizing $\mathrm{err}(\Phi(c'), \Phi)$ is to pick a concept whose *noisy version* agrees best with the sample drawn from the noisy target concept.

Definition 8. *Let $\mathcal{C}$ be a concept class, $c \in \mathcal{C}$ the target concept, $D \in \mathcal{D}_{\mathcal{X}}$, and Φ a noise model. Let $\mathcal{S} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ be a sample of size m drawn from the noisy concept $\Phi(c, D, \cdot)$. For any $c' \in \mathcal{C}$, $\mathrm{err}(c', \Phi, \mathcal{S})$ is defined by*

$$\mathrm{err}(\Phi(c'), \Phi, S) = \frac{1}{m} \sum_{i=1}^{m} \left| \Phi(c', D, x_i) - \frac{\#^+(x_i, \mathcal{S})}{\#(x_i, \mathcal{S})} \right|$$

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

The term $\#^+(x_i, \mathcal{S})/\#(x_i, \mathcal{S})$ approximates $\Phi(c, D, x_i)$ for the target concept c. As sample size grows, $\#^+(x_i, \mathcal{S})/\#(x_i, \mathcal{S}) \to \Phi(c, D, x_i)$ and $\mathrm{err}(\Phi(c'), \Phi, S) \to \mathrm{err}(\Phi(c'), \Phi)$. Unfortunately, to compute $\mathrm{err}(\Phi(c'), \Phi, S)$ for some c', the learning algorithm would have to know $\Phi(c', D, x)$—a probabilistic concept that depends on the unknown distribution D. The best we could hope for is that $\Phi(c', D, x)$ can be approximated using knowledge about D obtained from sampling.

Definition 9. *For any sample $\mathcal{S} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ of size m a distribution $D(\mathcal{S})$ is defined by $Pr_{x' \sim D(\mathcal{S})}[x' = x] = \#(x, \mathcal{S}) \cdot \frac{1}{m}$ for all $x \in \mathcal{X}$, where $\#(x, \mathcal{S}) = |\{j \in \{1, \ldots, m\} \mid x = x_j\}|$.*

Replacing D by $D(\mathcal{S})$ in Definition 8 allows us to approximate $\mathrm{err}(\Phi(c'), \Phi, S)$.

Definition 10. *Let $\mathcal{C}$ be a concept class, $c \in \mathcal{C}$ the target concept, $D \in \mathcal{D}_{\mathcal{X}}$, and Φ a noise model. Let $\mathcal{S} = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ be a sample of size m drawn from the noisy concept $\Phi(c, D, \cdot)$. For any $c' \in \mathcal{C}$, $\mathrm{err}(\Phi(c'), \Phi, \mathcal{S})$ can be estimated as follows (with $\#^+(x_i, \mathcal{S})$ and $\#(x_i, \mathcal{S})$ as in Definition 8).*

$$\widehat{\mathrm{err}}(\Phi(c'), \Phi, \mathcal{S}) = \frac{1}{m} \sum_{i=1}^{m} \left| \Phi(c', D(\mathcal{S}), x_i) - \frac{\#^+(x_i, \mathcal{S})}{\#(x_i, \mathcal{S})} \right|$$

[6] This quantity was first introduced as *variational distance* [10].

We call any algorithm that returns a concept minimizing $\widehat{\mathrm{err}}(\Phi(c'), \Phi, \mathcal{S})$ a *noisy minimum disagreement strategy*. In essence, it is a form of maximum likelihood process. Since $\widehat{\mathrm{err}}(\Phi(c'), \Phi, \mathcal{S})$ approximates $\mathrm{err}(\Phi(c'), \Phi, \mathcal{S})$ (which itself approximates $\mathrm{err}(\Phi(c'), \Phi)$), a noisy minimum disagreement strategy is expected to be successful only if the $\widehat{\mathrm{err}}(\Phi(c'), \Phi, \mathcal{S})$ provides a good estimate of $\mathrm{err}(\Phi(c'), \Phi)$.

Definition 11. *Φ is smooth with respect to concept class $\mathcal{C}$ and a class of distributions $\mathcal{D}$ iff there is a function $M : (0, 1/2) \times (0, 1/2) \to \mathbb{N}$ such that (1) $M(\epsilon, \delta)$ is polynomial in $1/\epsilon$ and $1/\delta$, for $\epsilon, \delta \in (0, 1/2)$; and (2) For all $\epsilon, \delta \in (0, 1/2)$, 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 the noisy oracle then, with probability of at least $1 - \delta$, for all $c' \in \mathcal{C}$ we obtain $|\mathrm{err}(\Phi(c'), \Phi) - \widehat{\mathrm{err}}(\Phi(c'), \Phi, \mathcal{S})| < \epsilon$.*

Distinctiveness and monotonicity can be generalized to the new setting by replacing $\mathrm{err}(c, \Phi)$ with $\mathrm{err}(\Phi(c), \Phi)$, resulting in *noise-distinctiveness* and *noise-monotonicity*, resp. It is not hard to show that random classification noise is both noise-distinctive (with $f(\epsilon) = \epsilon/2$ and $g(\epsilon) = \epsilon(1 - 2\eta)/2$) and noise-monotonic.

Sufficiency of noise-distinctiveness for learning of classes of finite VC-dimension is guaranteed if the smoothness property is fulfilled.

Theorem 6. *Let $\mathcal{C}$ be a concept class of finite* VC*-dimension d and Φ a noise model. If Φ is both noise-distinctive and smooth with respect to $\mathcal{C}$ and $\mathcal{D}_\mathcal{X}$ then $\mathcal{C}$ is learnable w.r.t. Φ using a noisy minimum disagreement strategy.*

Proof. Let f and g witness the noise-distinctiveness of Φ w.r.t. $\mathcal{C}$ and $\mathcal{D}$, and let $\epsilon, \delta \in (0, 1/2)$. We show that the noisy minimum disagreement strategy, with a sample $\mathcal{S}$ of at least $m = \max(m_1, m_2, m_3)$ examples, learns $\mathcal{C}$ w.r.t. Φ, where

$$m_1 = \left\lceil \max\left(\frac{4}{f(\epsilon)} \ln\left(\frac{8}{\delta}\right), \frac{8d}{f(\epsilon)} \ln\left(\frac{8d}{f(\epsilon)}\right)\right)\right\rceil, \; m_2 = \left\lceil M\left(\frac{g(\epsilon)}{2}, \frac{\delta}{4}\right)\right\rceil, \; m_3 = \left\lceil \frac{8}{g(\epsilon)^2} \ln\left(\frac{3({m_1}^d + 1)}{\delta}\right)\right\rceil.$$

m_1 examples suffice to find a set $\mathcal{C}_N$ of $N \leq {m_1}^d + 1$ concepts in $\mathcal{C}$ among which at least one has an error $\leq f(\epsilon)$ with probability $\geq 1 - \frac{\delta}{4}$ [12]. We show that the noisy minimum disagreement strategy will return one of these N concepts.

Since Φ is smooth for $\mathcal{C}$ and $\mathcal{D}_\mathcal{X}$, m_2 examples are sufficient to satisfy Definition 11 with ϵ and δ replaced by $g(\epsilon)/2$ and $1 - \delta/4$, resp. Finally, m_3 examples are sufficient for a noisy minimum disagreement strategy to select a concept in $\mathcal{C}_N$ that has an error $\leq \epsilon$ with probability $\geq 1 - \delta/2$ (cf. proof of Theorem 4). □

In parallel to Theorem 5, it is not hard to show that noise-monotonicity is necessary for learning a finite concept class using a noisy minimum disagreement strategy when the class of distributions is finite.

Finally, we show that noisy minimum disagreement strategies are a proper generalization of minimum disagreement strategies.

Proposition 3. *There is a concept class $\mathcal{C}$ over a finite input space $\mathcal{X}$ and a noise model Φ such that $\mathcal{C}$ is learnable w.r.t. Φ using a noisy minimum disagreement strategy, but no minimum disagreement strategy learns $\mathcal{C}$ w.r.t. Φ.*

Proof. Let $\mathcal{C}$ and Φ be as in the proof of Proposition 2. Since $|\mathcal{X}| = 2$, each $D \in \mathcal{D}_\mathcal{X}$ is uniquely identified by the probability p with which x_1 is sampled. It

is then easy to prove that Φ is smooth and that $f(\epsilon) = \epsilon$ and $g(\epsilon) = \epsilon/2$ witness noise-distinctiveness of Φ w.r.t. $\mathcal{C}$ and $\mathcal{D}_X$. Theorem 6 then proves the claim. □

5 Conclusions

A high-level study of noise models, as our definition allows, gives insights into conditions under which learning under noise in general can be guaranteed. We hope that our formal framework and the insights gained from it will inspire the definition of new, potentially more realistic noise models and classes of distributions under which sample-efficient learning is possible.

Acknowledgements. This work was supported by the Alberta Innovates Centre for Machine Learning (AICML) and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2:343–370, 1988.
2. S. Ben-David, S. Shalev-Shwartz, and R. Urner. Domain adaptation–can quantity compensate for quality? In *ISAIM*, 2012.
3. A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Classifying learnable geometric concepts with the Vapnik-Chervonenkis dimension. In *STOC*, pages 273–282, 1986.
4. N. Bshouty, N. Eiron, and E. Kushilevitz. PAC learning with nasty noise. *Theoretical Computer Science*, 288:255–275, 2002.
5. K. Crammer, M. Kearns, and J. Wortman. Learning from data of variable quality. In *NIPS*, pages 219–226, 2005.
6. S. Decatur. PAC learning with constant-partition classification noise and applications to decision tree induction. In *ICML*, pages 83–91, 1997.
7. S. Goldman and R. Sloan. Can PAC learning algorithms tolerate random attribute noise? *Algorithmica*, 14:70–84, 1995.
8. D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.
9. M. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45:983–1006, 1998.
10. M. Kearns and R. Schapire. Efficient distribution-free learning of probabilistic concepts. In *SFCS*, pages 382–391, 1990.
11. M. Kearns, R. Schapire, and L. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994.
12. P. Laird. *Learning from Good and Bad Data*. Kluwer Academic Publishers, 1988.
13. L. Ralaivola, F. Denis, and C. Magnan. CN = CPCN. In *ICML*, pages 721–728, 2006.
14. R. Sloan. Four types of noise in data for PAC learning. *Information Processing Letters*, 54:157–162, 1995.
15. A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32:135–166, 2004.
16. L. Valiant. A theory of the learnable. In *STOC*, pages 436–445, 1984.
17. V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.