

Improved Lower Bounds for Learning from Noisy Examples: an Information-Theoretic Approach

Claudio Gentile¹

DSI, Universita' di Milano, Via Comelico 39, 20135 Milano, Italy

E-mail: gentile@dsi.unimi.it

and

David P. Helmbold²

Computer Science Department, University of California, 95064 Santa Cruz, USA

E-mail: dph@cse.ucsc.edu

This paper presents a general information-theoretic approach for obtaining lower bounds on the number of examples needed to PAC learn in the presence of noise. This approach deals directly with the fundamental information quantities, avoiding a Bayesian analysis. The technique is applied to several different models, illustrating its generality and power. The resulting bounds add logarithmic factors to (or improve the constants in) previously known lower bounds.

Key Words: PAC learning; entropy; mutual information; lower bounds; noisy examples

1. INTRODUCTION

When labeled examples are scarce or expensive, one should employ a learning method that requires as few examples as possible. In order to determine this minimal number of examples, one must not only have good algorithms, but also good lower bounds. In this paper we present a unified information-theoretic approach for lower bounding the number of examples needed to learn in various models with noise. Not only does our approach allow the easy derivation of previously known bounds, but it also yields additional logarithmic factors in several cases.

¹Partially supported by MURST, Cofin99 (Project: Stochastic Processes with Spatial Structure), and by ESPRIT, Working Group EP 27150, Neural and Computational Learning II (NeuroColt II).

²Supported by NSF grants CCR 9700201 and CCR-9821087.

The models we consider are variants of the PAC model [35], where a domain of instances and a class of concepts (0-1 valued functions) on the domain are specified as part of the learning problem. An adversary (perhaps randomly) selects a target concept from the class and a distribution on the domain. The distribution on the domain is often used to generate examples (instances labeled by the target concept) for the learner, and the learner’s goal is to find a 0-1 valued hypothesis that, with high probability, closely approximates the target concept. Typically a noise process, which is random and/or adversarially controlled, corrupts some of the labeled examples so that the learner can see misleading and possibly contradictory examples.

Our basic approach emphasizes the amount of information that the algorithm must discover about the target concept. We use the PAC learning criterion to lower bound this amount of information. We also upper bound the information about the target contained in the sample as a function of the sample size. Since all the information about which concept from the class is the target comes from the sample, we can solve these two bounds to get a bound on the sample size required by the algorithm.

This approach deals directly with the fundamental information quantities, rather than bounding more abstract entities such as the Bayes risk. We believe that this directness is a major contributor to the clean results, simplicity of the proofs, and generality of the approach. Also, since we measure only the information learned about the target concept, our bounds hold even when the algorithm knows the probabilistic model.

We use a definition of learning that explicitly differentiates between the distribution of examples seen by the algorithm and the test distribution on which the algorithm’s hypothesis is evaluated. This enables us to easily apply our techniques to a wide variety of learning models including: malicious noise [26], classification noise [4], drifting distributions [8], and membership queries [1, 2].

When proving lower bounds, one must make assumptions about the complexity of the concept classes considered. Most of the bounds we obtain are of two forms: the first form uses a simple 2-concept class over a 2-element domain, while the second is based on (restricted) unions of intervals. Although bounds of the first form hold for any non-trivial concept class, they do not exploit the VC-dimension of the class. Bounds of the second form depend on the size of ϵ -covers for the concept class. These bounds are better by a $\log 1/\epsilon$ factor than previous bounds stated in terms of the VC-dimension for any concept class that can embed or simulate (restricted) unions of intervals. Some natural classes with this property include half spaces and axis-parallel hyper-rectangles [24].

For the malicious noise model of Kearns and Li [26] we use the simple 2-concept class to improve the lower bounds of Cesa-Bianchi et al. [13] by a $\log 1/\delta$ factor (Theorem 5.1) when the noise rate is bounded away from 0.

In the more benign classification noise model of Angluin and Laird [4], we have a bound using unions of intervals that improves by a $\log 1/\epsilon$ factor the previous bound proven by Simon [31] and Apolloni and Gentile [6] when the noise rate is bounded away from zero. Our result shows that the sample size required to learn natural classes has different limiting behavior as $\epsilon \rightarrow 0$ in the noise-free case (where

it is $\Omega(d_{VC}/\epsilon)$ [17]) and when the noise rate is a positive constant (where it grows as $\Omega(\frac{d_{VC}}{\epsilon} \log \frac{1}{\epsilon})$, Theorem 5.2).

Another model we consider is a membership query model [1, 2] augmented with classification noise. Here we obtain a new sample size bound for arbitrary concept classes that generalizes the results of Turàn [34] in two ways. First, it adds a factor indicating the dependence on the noise rate. Second, our bound depends on the size of ϵ -covers rather than the VC-dimension. This allows us to add a $\log 1/\epsilon$ factor to the bound for many natural concept classes (Theorem 5.5).

The final model we apply our techniques to is a batch version of Bartlett’s drifting distribution model [8] with classification noise. Here we bound the allowable rate of drift in the noisy case, generalizing results of Bartlett [8], Aslam and Decatur [7], Simon [31] and Apolloni and Gentile [6]. The drifting and membership query models illustrate the benefit of explicitly differentiating between the distribution from which the examples are drawn and the test distribution.

It is remarkable that the same basic techniques yield simple yet strong sample size bounds for such a wide variety of learning models. Furthermore, we shall see in Section 5 that one of our bounds is strong enough to imply a bound on the tails of certain binomial distributions that we have not seen in the literature.

The next section contains the main definitions used throughout the paper. Section 3 describes our general methodology. Lower bounds on the information required for learning are given in Section 4. Section 5 uses upper bounds on the information in a sample (combined with the bounds in Section 4) to get sample size bounds for various learning models.

2. PRELIMINARY DEFINITIONS AND NOTATION

This section defines the learning framework, as well as some of the notation used throughout the paper. Table 1 at the end of this section summarizes our notational conventions.

When X is a random variable, we use \mathcal{P}_X to denote its distribution (or density) function, and will sometimes drop the subscript when X is clear from the surrounding context. If f is a deterministic and measurable function, then $E_{\mathcal{P}_X}[f(X)]$ denotes the expected value of f . For two random variables X and Y their joint distribution is denoted by \mathcal{P}_{XY} , and the conditional distribution of X given $Y = y$ is denoted by $\mathcal{P}_{X|y}$. In general we will abuse the notation and write either $\mathcal{P}(x)$ or $\Pr_{\mathcal{P}}(x)$ (rather than $\Pr_{\mathcal{P}}(\{x\})$).

For simplicity of exposition, we assume throughout the paper that every random variable either takes values in a countable set or is continuous (so all densities are w.r.t. the counting measure or the Lebesgue measure).

For a random variable X over a domain \mathcal{X} and a random variable Y over a domain \mathcal{Y} the (Shannon) entropy $H(X)$ of X is defined by (here and throughout “log” means “log₂” while “ln” is the natural logarithm. As usual, $0 \log 0 = 0 \log \infty = 0$.)

$$H(X) = E_{\mathcal{P}_X}[-\log \mathcal{P}_X(X)]$$

and the joint entropy, $H(X, Y)$, of X and Y is defined as

$$H(X, Y) = E_{\mathcal{P}_{XY}}[-\log \mathcal{P}_{XY}(X, Y)].$$

The conditional entropy of X given $Y = y$ is denoted by $H(X | Y = y)$ and is defined by replacing \mathcal{P}_X in the definition of $H(X)$ above by the conditional distribution $\mathcal{P}_{X|y}$. As usual,

$$H(X | Y) = E_{\mathcal{P}_Y}[H(X | Y = y)].$$

We define the mutual information $I(X; Y)$ between X and Y by

$$I(X; Y) = \sup \sum_{i=1}^n \Pr_{\mathcal{P}_{XY}}(B_i) \log \frac{\Pr_{\mathcal{P}_{XY}}(B_i)}{\Pr_{\mathcal{P}_X}(B_{i|\mathcal{X}}) \Pr_{\mathcal{P}_Y}(B_{i|\mathcal{Y}})} \quad (1)$$

where the supremum is taken over all finite partitions $\{B_1, \dots, B_n\}$ of $\mathcal{X} \times \mathcal{Y}$ into Borel sets and $B_{i|\mathcal{X}}, B_{i|\mathcal{Y}}$ are the projections of B_i onto \mathcal{X} and \mathcal{Y} , respectively. It can be shown [15] that when X and Y have densities \mathcal{P}_X and \mathcal{P}_Y then the definition above is equivalent to

$$I(X; Y) = E_{\mathcal{P}_{XY}} \left[\log \frac{\mathcal{P}_{XY}(X, Y)}{\mathcal{P}_X(X) \mathcal{P}_Y(Y)} \right],$$

when “ $+\infty$ ” is considered equivalent to “ $+\infty$ ” if the sup in (1) is not finite.

If Z has distribution \mathcal{P}_Z we denote the conditional mutual information of X and Y given $Z = z$ by $I(X; Y | Z = z)$, defined by replacing \mathcal{P}_{XY} by $\mathcal{P}_{XY|z}$, \mathcal{P}_X by $\mathcal{P}_{X|z}$ and \mathcal{P}_Y by $\mathcal{P}_{Y|z}$ in (1). Again,

$$I(X; Y | Z) = E_{\mathcal{P}_Z}[I(X; Y | Z = z)].$$

Throughout the paper we will use the following well-known relationships (e.g., [14, 25]) among the quantities we have just defined. All of them hold when the entropies and the mutual informations involved are finite. For random variables X , Y , Z and T we have the following facts:

Fact F1. $I(X; Y | Z) = I(Y; X | Z)$ (symmetry of mutual information);

Fact F2. $I(X; Y | Z) = H(X | Z) - H(X | Y, Z)$;

Fact F3. $I(X, Y; Z | T) = I(X; Z | T) + I(Y; Z | X, T)$ (additivity of mutual information);

Fact F4. if A is any (deterministic and measurable) function of Y we have:

$$I(X; A(Y)) \leq I(X; Y).$$

For a discrete random variable with density (p_1, \dots, p_m) we sometimes denote its entropy by $\mathcal{H}[p_1, \dots, p_m]$. When $m = 2$ we abbreviate $\mathcal{H}[p, 1 - p]$ by the binary entropy function

$$\mathcal{H}(p) = -p \log p - (1 - p) \log(1 - p).$$

Throughout the domain \mathcal{X} is a fixed set which we assume to be either finite, countable, or \mathcal{R}^n for some $n \geq 1$, \mathcal{B} is an algebra of Borel sets over \mathcal{X} , and \mathcal{P} is a probability distribution on \mathcal{X} . We use $X^m = (X_1, \dots, X_m)$ and $L^m = (L_1, \dots, L_m)$ to denote an \mathcal{X}^m -valued and a $\{0, 1\}^m$ -valued random vector, respectively. Their realizations will be denoted in lower case, as $x^m = (x_1, \dots, x_m)$ and $l^m = (l_1, \dots, l_m)$, respectively.

A *concept* c on $(\mathcal{X}, \mathcal{B})$ is an element of \mathcal{B} and a *concept class* \mathcal{C} on $(\mathcal{X}, \mathcal{B})$ is a subset of \mathcal{B} . We will also find it useful to view a concept as a random variable with distribution \mathcal{D} over \mathcal{C} . In such a case we denote the random variable with the small capital ‘‘c’’. A (*labeled*) *example* is a pair $(x, l) \in \mathcal{X} \times \{0, 1\}$. A (*labeled*) *sample* $S_c(x^m)$ for concept c is a pair $(x^m, L_c^m(x^m))$ where $x^m = (x_1, \dots, x_m) \in \mathcal{X}^m$, and (in the absence of noise, see below) $L_c^m(x^m) = (I_c(x_1), \dots, I_c(x_m)) \in \{0, 1\}^m$ where I_c is the indicator function for concept c . The random variable R (with realization r) is a finite sequence of unbiased random bits representing the randomization available to the learning function described in Definition 2.1 below.

A *noise model* for the examples mathematically defines the way in which a sample $S_c(x^m)$ for a concept c is corrupted by the noise. This can be viewed as a process that takes as input $S_c(x^m)$ and r , and outputs a corrupted sample $\hat{S}_c(x^m) = (\hat{x}^m, \hat{l}^m)$ where $\hat{x}^m = (\hat{x}_1, \dots, \hat{x}_m) \in \mathcal{X}^m$ and $\hat{l}^m = (\hat{l}_1, \dots, \hat{l}_m) \in \{0, 1\}^m$.

A key distribution we will consider is the joint distribution between the information seen by the algorithm (the instances, their labels, and the randomization) and the target concept chosen by the adversary. We denote this joint distribution as \mathcal{M} over $\mathcal{X}^m \times \{0, 1\}^m \times \{0, 1\}^* \times \mathcal{C}$ (an m -indexing for \mathcal{M} is understood), and assume that \mathcal{M} factors as $\mathcal{M}(\cdot, \cdot, \cdot, c) = \mathcal{D}(c)\mathcal{M}(\cdot, \cdot, \cdot | c)$. Here \mathcal{D} is the distribution over the concept class mentioned above, while $\mathcal{M}(\cdot, \cdot, \cdot | c)$ is the conditional distribution given c on the corrupted sample and the random bits under the adopted noise model.

Thus, for every fixed $c \in \mathcal{C}$, the noise model *induces* the distribution $\mathcal{M}(\hat{x}^m, \hat{l}^m, r | c)$ over the set of corrupted samples and the random bits. This distribution is generally a function of the underlying probabilistic model for generating the initial sample $S_c(x^m)$ and the random bits r . Although the noise we consider is usually i.i.d., in principle our techniques could be applied to other kinds of noise models.

Below we give two relevant examples that we will be using in the application section of the paper.

In the *classification noise model* of Angluin and Laird [4] the noise affects only the labels and $\mathcal{M}(x^m, \hat{l}^m, r | c)$ factors as

$$\mathcal{M}(x^m, \hat{l}^m, r | c) = \mathcal{P}^m(x^m) \mathcal{M}(\hat{l}^m | x^m, c) \mathcal{M}(r),$$

where \mathcal{P}^m denotes the m -fold \mathcal{P} -probability product and $\mathcal{M}(\hat{l}^m | x^m, c)$ describes the i.i.d. noisy labeling process. In particular

$$\mathcal{M}(\hat{l}^m | x^m, c) = \prod_{i=1}^m \mathcal{M}(\hat{l}_i | x_i, c),$$

where $\mathcal{M}(I_c(x_i) | x_i, c) = 1 - \eta$ (no noise) and the label is flipped with probability η .

We will also exhibit an application to learning with slowly *drifting* distributions (Bartlett [8]). Here the underlying marginal distribution $\mathcal{M}(x^m)$ is a product distribution: $\mathcal{M}(x^m) = \prod_{i=1}^m \mathcal{P}_i(x_i)$ where the \mathcal{P}_i 's are slowly changing, according to a suitable definition of distance between distributions that will be specified in Section 5.3.

Throughout the remainder of this paper, all functions are assumed to be *deterministic* (w.l.o.g.) and *measurable*.

Let $c, h \in \mathcal{B}$. When \mathcal{P} is understood from the context we say that c is ϵ -close to h if $\Pr_{\mathcal{P}}(c\Delta h) < \epsilon$, where $c\Delta h = \{x \in \mathcal{X} : I_c(x) \neq I_h(x)\}$, and c is ϵ -far from h otherwise.

DEFINITION 2.1. Concept class \mathcal{C} on $(\mathcal{X}, \mathcal{B})$ is *PAC-learnable* w.r.t. distributions \mathcal{P} and $\mathcal{M}(\hat{x}^m, \hat{I}^m, r | c)$ ($(\mathcal{P}, \mathcal{M})$ -learnable for short) if there exist functions for the *sample complexity* $m = m(\epsilon, \delta)$ and number of random bits $b = b(\epsilon, \delta)$ required by a *learning function*³ $A : \mathcal{X}^m \times \{0, 1\}^m \times \{0, 1\}^b \rightarrow \mathcal{B}$ such that for every $\epsilon, \delta > 0$ and $c \in \mathcal{C}$,

$$\Pr_{\mathcal{M}(\hat{x}^m, \hat{I}^m, r | c)}(\{\hat{S}_c(x^m), r : \Pr_{\mathcal{P}}(c\Delta A(\hat{S}_c(x^m), r)) < \epsilon\}) > 1 - \delta.$$

Here r is a sequence of b random bits, c is the *target concept* (or simply the *target*) and $A(\hat{S}_c(x^m), r)$ is the actual *hypothesis* generated by the learning function A . No assumptions are made on this hypothesis (other than its membership in \mathcal{B}).

This definition of learning has some interesting properties. In contrast to most PAC models, the learning function “knows” (or can be specialized for) the test distribution \mathcal{P} and the distribution of samples $\mathcal{M}(\hat{x}^m, \hat{I}^m, r | c)$. However, the learning constraint is for all possible targets $c \in \mathcal{C}$. Note that an equivalent definition results when the quantification “for all $c \in \mathcal{C}$ ” is replaced by “for all \mathcal{D} over \mathcal{C} ” and the “ $\Pr_{\mathcal{M}(\hat{x}^m, \hat{I}^m, r | c)}$ ” is replaced by “ $\Pr_{\mathcal{M}(\hat{x}^m, \hat{I}^m, r, c)}$ ” (with the interpretation that c is drawn according to \mathcal{D}).

For a concept class \mathcal{C} on $(\mathcal{X}, \mathcal{B})$ and a probability distribution \mathcal{P} on \mathcal{X} , the subclass $\mathcal{C}' \subseteq \mathcal{C}$ is called an ϵ -cover of \mathcal{C} w.r.t. \mathcal{P} if for every $c \in \mathcal{C}$ there is a $c' \in \mathcal{C}'$ such that c' is ϵ -close to c [11], [27], [36, pp.149-151]. We denote by $N(\mathcal{C}, \epsilon, \mathcal{P})$ the cardinality of a smallest ϵ -cover of \mathcal{C} w.r.t. \mathcal{P} .

In Section 5.4 we consider learning with queries. There we show that the finite coverability of \mathcal{C} w.r.t. \mathcal{P} (i.e., $N(\mathcal{C}, \epsilon, \mathcal{P}) < \infty$ for each $\epsilon > 0$) is necessary for the $(\mathcal{P}, \mathcal{M})$ -learnability of \mathcal{C} regardless of the query model \mathcal{M} .

We introduce the following key definition that is tailored for our lower bound purposes.

DEFINITION 2.2. For a concept class \mathcal{C} on $(\mathcal{X}, \mathcal{B})$ and a probability distribution \mathcal{P} on \mathcal{X} , the subclass $\mathcal{C}_\epsilon \subseteq \mathcal{C}$ is an ϵ -well-separated subclass of \mathcal{C} w.r.t. \mathcal{P} if for all $h \in \mathcal{B}$ there is at most one $c \in \mathcal{C}_\epsilon$ that is ϵ -close to h . We drop “w.r.t. \mathcal{P} ” when distribution \mathcal{P} is clear from the context.

³The learning function A can be considered deterministic once its internal randomization has been fixed. We use the term “learning function” to emphasize that A need not be computable.

Although we define $N(\mathcal{C}, \epsilon, \mathcal{P})$ to be the cardinality of proper ϵ -covers, the definition of ϵ -well-separated has a different flavor. A subclass \mathcal{C}_ϵ is ϵ -well-separated if and only if every subset of the domain is ϵ -close to at most one element of \mathcal{C}_ϵ .

If the concepts in subclass \mathcal{C}_ϵ are mutually 2ϵ -far from each other, then the triangle inequality implies that \mathcal{C}_ϵ is an ϵ -well-separated subclass. Therefore, we can use the following lemma to show the existence of large ϵ -well-separated subclasses.

LEMMA 2.1 ([27],[11]). *Let \mathcal{C} be a concept class on $(\mathcal{X}, \mathcal{B})$ with $N(\mathcal{C}, 2\epsilon, \mathcal{P}) = N$. Then there exists a finite subset of \mathcal{C} of cardinality at least N whose elements are mutually 2ϵ -far.*

Remark. Many papers in the lower bound literature (e.g. [11] and [19]) use mutual separation (as in the consequence of Lemma 2.1) rather than ϵ -well-separated subclasses to measure the complexity of a concept class. The cardinality of a largest subset $\mathcal{C}' \subseteq \mathcal{C}$ whose elements are mutually ϵ -far is usually called the ϵ -packing number of \mathcal{C} w.r.t. \mathcal{P} [27]. Let $M(\mathcal{C}, \epsilon, \mathcal{P})$ be this quantity. We point out that if \mathcal{C}_ϵ is ϵ -well-separated then its members are mutually ϵ -far from each other, so $|\mathcal{C}_\epsilon| \leq M(\mathcal{C}, \epsilon, \mathcal{P})$. As indicated above, the triangle inequality implies that if \mathcal{C}_ϵ is a maximum ϵ -well-separated set then $M(\mathcal{C}, 2\epsilon, \mathcal{P}) \leq |\mathcal{C}_\epsilon|$. Therefore $|\mathcal{C}_\epsilon|$ is sandwiched by two ϵ -packing numbers.

A simple but relevant example of ϵ -well-separation that we will use several times in the subsequent sections is the following. For $x_1, x_2 \in \mathcal{X}$, let \mathcal{P} be the distribution on \mathcal{X} defined by $\mathcal{P}(x_1) = 1 - \epsilon$, $\mathcal{P}(x_2) = \epsilon$, and $\mathcal{P}(x) = 0$ elsewhere. Let \mathcal{C} be the class on $(\mathcal{X}, \mathcal{B})$ defined by $\mathcal{C} = \{c_1, c_2\}$, where $c_1 = \{x_1, x_2\}$, $c_2 = \{x_1\}$. If $1 - \epsilon \geq \epsilon$ (i.e., $\epsilon \leq 1/2$) then \mathcal{C} itself is an ϵ -well-separated subclass (however, the concepts in \mathcal{C} are not 2ϵ -far). We will refer to this pair of \mathcal{C} and \mathcal{P} as an ϵ -binary pair on $(\mathcal{X}, \mathcal{B})$.

Finally, the VC-dimension of a concept class \mathcal{C} , denoted by $d_{VC}(\mathcal{C})$, is the cardinality of a largest subset of the domain *shattered* by \mathcal{C} (see Vapnik [37, p. 53] and Blumer et al. [12]).

3. MUTUAL INFORMATION AND THE METHOD OF INDUCED DISTRIBUTIONS

This section outlines the method we use for lower bounding the sample size required for learning. Let \mathcal{C} be a concept class on $(\mathcal{X}, \mathcal{B})$ and consider the mutual information $I(\mathbf{c} ; \hat{X}^m, \hat{L}^m, R)$ between the \mathcal{C} -valued random variable \mathbf{c} and the joint variable $(\hat{X}^m, \hat{L}^m, R)$ formed by the (corrupted) sample (\hat{X}^m, \hat{L}^m) for \mathcal{C} and the random bits R . Consider the distribution \mathcal{M} introduced in the last section. We recall that $(\hat{X}^m, \hat{L}^m, R)$ has distribution $\mathcal{M}(\hat{x}^m, \hat{l}^m, r)$, the marginal of \mathcal{M} w.r.t. \mathbf{c} . Let $\mathbf{H} = A(\hat{X}^m, \hat{L}^m, R)$ be the value of the learning function A for \mathcal{C} on arguments \hat{X}^m, \hat{L}^m, R . Since \mathbf{H} is a function of \hat{X}^m, \hat{L}^m and R , by Fact F4 we have

$$I(\mathbf{c} ; \mathbf{H}) \leq I(\mathbf{c} ; \hat{X}^m, \hat{L}^m, R).$$

TABLE 1
Summary of the main notation used in the paper

<i>symbol</i>	<i>meaning</i>
\mathcal{X}	domain
\mathcal{B}	algebra of Borel sets
\mathcal{P}	distribution over \mathcal{X}
\mathcal{P}_X	distribution or density of the random variable X
dist	distance function between two distributions
H	entropy
I, \hat{I}	mutual information and partitional mutual information, respectively
$\mathcal{H}[\dots], \mathcal{H}(\cdot)$	entropy function and binary entropy function, respectively
$c, \mathcal{C}, \mathcal{C}$	concept, concept class and \mathcal{C} -valued random variable, respectively
I_c	indicator function of c
$S_c(x^m)$	labeled sample for c
$\hat{S}_c(x^m) = (\hat{x}^m, \hat{l}^m)$	corrupted labeled sample for c
η	noise rate
r, R	sequence of random bits and $\{r\}$ -valued random variable, respectively
\mathcal{D}	distribution over \mathcal{C} (or over \mathcal{C}_ϵ)
d	$\mathcal{D}(c_1)$ when $\mathcal{C}_\epsilon = \{c_1, c_2\}$
\mathcal{M}	distribution over $\mathcal{X}^m \times \{0, 1\}^m \times \{0, 1\}^* \times \mathcal{C}$
A	learning function
$\mathbb{H} = A(\hat{S}_c(X^m), R)$	random variable denoting the hypothesis of A
m	sample complexity (of A)
\mathcal{C}_ϵ	ϵ -well-separated subclass of \mathcal{C}
N	(lower bound on the) cardinality of a largest \mathcal{C}_ϵ
$N(\mathcal{C}, \epsilon, \mathcal{P})$	cardinality of a smallest ϵ -cover of \mathcal{C} w.r.t. \mathcal{P}
$d_{VC}(\mathcal{C})$	VC-dimension of \mathcal{C}

If c is *independent* of the internal randomization R of A then $I(R ; c) = 0$. By the symmetry and the additivity of I we get

$$I(c ; \hat{X}^m, \hat{L}^m, R) = I(\hat{X}^m, \hat{L}^m, R ; c) = I(R ; c) + I(\hat{X}^m, \hat{L}^m ; c | R).$$

Hence we obtain the general inequality

$$I(c ; H) \leq I(\hat{X}^m, \hat{L}^m ; c | R), \tag{2}$$

which holds for every $(\mathcal{P}, \mathcal{M})$ -learning function A for \mathcal{C} , no matter which distribution \mathcal{D} over \mathcal{C} is chosen, as long as the target c is independent of the random bits used by A . From now on we assume c and R are independent.

It is quite instructive to interpret both sides of (2). Given the knowledge of \mathcal{C} , \mathcal{P} and \mathcal{M} , the LHS of (2) is roughly the number of bits of information required for A to perform the learning task at hand: A must identify a target concept inside a known concept class, up to ϵ error with confidence δ . The RHS refers to the average information content of the corrupted sample seen by the learning function. This information content is measured by the mutual information $I(\hat{X}^m, \hat{L}^m ; c | R)$ which represents the “degree of dependence” between target concept c and sample (\hat{X}^m, \hat{L}^m) , given the function’s randomization R . When the labeled sample is error free this mutual information is always a strictly increasing function of m for every \mathcal{D} (disregarding degenerate cases). This suggests that, as long as the LHS of (2) is finite, a sample of suitable (finite) size is sufficient to learn. On the other hand, if the sample is corrupted by a noise process then the sample might contain no information about the actual target. Such a situation is desirable when we are designing adversarial noise strategies to make the learning process as hard as possible. The method of *induced distributions* introduced in Kearns and Li [26] (but see also Angluin and Laird [4], Sloan [32], Cesa-Bianchi et al. [13]) can thus be reinterpreted by means of this mutual information argument. The adversary tries to make $I(\hat{X}^m, \hat{L}^m ; c | R)$ as small as possible. If for every m the adversary can make c and (\hat{X}^m, \hat{L}^m) statistically independent (given R), then it can prevent PAC learning, irrespective of the number of examples used and the computational resources of the learning function. Our approach is similar in spirit, but working from inequality (2) allows us to *quantify* the hardness of the learning task as $I(\hat{X}^m, \hat{L}^m ; c | R) \rightarrow 0$.

To quantify this hardness we need to perform two further steps.

1) Find a suitable *lower* bound on the LHS of (2) by exploiting the fact that A is a learning function for \mathcal{C} .

2) Find a suitable *upper* bound on the RHS of (2) by exploiting the data of the problem, i.e., the concept class \mathcal{C} and the distribution \mathcal{M} induced by the underlying distribution law over \mathcal{X}^m and the actual noise process.

Both of the above bounds depend on the distribution \mathcal{D} over the concept class. Our lower bound (step 1) will not depend on the sample size m , while our upper bound (step 2) will. Relating these bounds through inequality (2) yields a bound on the sample size. Moreover, examining the relationship between the two sides can guide the selection of \mathcal{D} , as we will see later.

Consider the case when the examples (\hat{X}_i, \hat{L}_i) in the sample (\hat{X}^m, \hat{L}^m) have the same distribution and are conditionally independent given R and conditionally independent given both C and R . Thus, once the function’s randomization and the target concept are fixed, the examples are i.i.d. Under these assumptions we can easily describe how the RHS of (2) depends on m . In particular $I(\hat{X}^m, \hat{L}^m ; C | R)$ factors as $mI(\hat{X}_i, \hat{L}_i ; C | R)$, and solving (2) for m results in a lower bound of the form

$$m(\mathcal{D}) \geq \frac{I(C ; H)}{I(\hat{X}_i, \hat{L}_i ; C | R)} \quad (3)$$

that holds for every learning function and every distribution \mathcal{D} over the concept class. One can compute the supremum over all possible distributions \mathcal{D} of the RHS of (3) to make the bound as tight as possible.

We wish to stress one subtlety: while the testing distribution \mathcal{P} *only* affects the LHS of (2), which can be bounded in terms of the (pseudo)-metric properties of \mathcal{C} w.r.t. \mathcal{P} (as we will see in Section 4), the distribution \mathcal{M} governing the sample *only* affects the RHS of (2). As a consequence we are able to clearly separate the roles of these two distributions. Varying the distribution \mathcal{M} allows us to treat several kinds of noise models, distributions that change over time, learning with membership queries, and various combinations of these.

The role in this argument played by the learning function’s randomization R is quite marginal, since in the LHS of (2) R is plugged into H and in the RHS of (2) it is a conditioning quantity. If, as is often assumed, the internal randomization of A is statistically independent of the relevant quantities that A is inferring, then R cannot provide any information about them, and we can drop the R -conditioning in all of the entropies.

When the observed examples (\hat{X}_i, \hat{L}_i) are i.i.d. given C , we could adopt Bayesian terminology and say that \mathcal{D} is a *prior* over a parameter space \mathcal{C} and that $I(C ; \hat{X}^m, \hat{L}^m)$ is the *Bayes risk* of the optimal (Bayesian) on-line estimator for the common density of (\hat{X}_i, \hat{L}_i) under log loss. From this Bayes risk one can obtain a lower bound on the *minimax risk*, which is essentially the capacity of the channel mapping concepts to samples described by the conditional distributions $\mathcal{M}(x^m, l^m | c) = \prod_{i=1}^m \mathcal{M}(x_i, l_i | c)$.

There is a large amount of literature related to the problem of finding upper and lower bounds on the mutual information I between a parameter and a set of m observations (see, e.g., Haussler and Barron [20], Haussler, Kearns and Schapire [21], Haussler and Opper [23], Yu [38] and the references therein). We emphasize that the present paper has a different concern. We do not regard I as a function of m ; we are instead interested in finding conditions that m must satisfy in order to meet the PAC-learnability requirements. Furthermore, the interpretation of \mathcal{D} as a prior over \mathcal{C} is somewhat misleading in this paper. We prefer to view it as a *free parameter* to be optimally tuned in order to obtain the tightest bounds. Since we are restricting ourselves to the PAC framework, we are able to get practical bounds on m which are not asymptotic in nature.

The idea of using information-theoretic tools to prove sample size lower bounds for PAC-learning is taken from Apolloni and Gentile [6]. In that paper the authors adopt an Algorithmic Complexity formalism and point out the ability of their

method to treat the testing distribution \mathcal{P} and the sampling distribution $\mathcal{M}(\cdot, \cdot, \cdot | c)$ as completely unrelated parameters of the learning problem.

The idea of using the symmetry of the mutual information I between the target concept and the training sample and then to compare the two alternative expressions for I is stressed in [19]. The generalization of this framework to arbitrary noise models and the method of maximizing the ratio of the two expressions for I over the \mathcal{D} 's is, to the best of our knowledge, a new one. By this method we are able to get new meaningful lower bounds in a clean and almost automatic way.

4. BOUNDING $I(\mathbb{C} ; \mathbb{H})$ AS A FUNCTION OF \mathcal{D}

In this section we will bound the information, $I(\mathbb{C} ; \mathbb{H})$, required for learning in terms of the cardinality of the ϵ -well-separated subclasses (as defined in Section 2). Here ϵ is intended to be the desired accuracy of the learning function under consideration.

Let \mathcal{C} be a concept class on $(\mathcal{X}, \mathcal{B})$ and A be a $(\mathcal{P}, \mathcal{M})$ -learning function for \mathcal{C} . Let N be the cardinality of a largest ϵ -well-separated subclass $\mathcal{C}_\epsilon = \{c_1, \dots, c_N\}$ of \mathcal{C} and \mathbb{C} be a random variable with distribution $\mathcal{D} = (d_1, \dots, d_N)$ over \mathcal{C}_ϵ . Now, set for brevity $\mathbb{H} = A(\hat{X}^m, \hat{L}^m, R)$. Since \mathcal{C}_ϵ is ϵ -well-separated, for any hypothesis h in the range of A (for some choice of its arguments) there exists at most one $c_k \in \mathcal{C}_\epsilon$ that is ϵ -close to h . On the other hand, since A is a learning function for \mathcal{C} , for any $c_k \in \mathcal{C}_\epsilon$ there exists at least one h in the range of A that is ϵ -close to c_k (with respect to distribution \mathcal{P}). Any h that is ϵ -far from all c_i will not have the desired accuracy. However, we do not know if the algorithm can produce such an h . Therefore we simplify the analysis by treating this “ ϵ -far from all c_i ” case as if h was ϵ -close to c_1 . Let

$$\text{Ran}(A) = \{h : h \text{ is in the range of } A\}.$$

Define

$$\text{cl}(1) = \{h \in \text{Ran}(A) : h \text{ is } \epsilon\text{-close to } c_1 \text{ or } \forall c_k \in \mathcal{C}_\epsilon \text{ } h \text{ is } \epsilon\text{-far from } c_k\}$$

and, for $k = 2, \dots, N$,

$$\text{cl}(k) = \{h \in \text{Ran}(A) : h \text{ is } \epsilon\text{-close to } c_k\}.$$

The family of sets $\{\text{cl}(k), k = 1, \dots, N\}$ partition the hypotheses produced by A , and thus induce a partition of A 's arguments, $\mathcal{X}^m \times \{0, 1\}^m \times \{0, 1\}^b$. Therefore, the family of sets

$$\{B_{ik} = \text{cl}(k) \times \{c_i\}, i, k = 1, \dots, N\}$$

can be viewed as partitioning $\mathcal{X}^m \times \{0, 1\}^m \times \{0, 1\}^b \times \mathcal{C}_\epsilon$.

Focus now on $I(\mathbb{C} ; \mathbb{H})$. By the definition of mutual information in (1) we know that the partitional mutual information $\hat{I}(\mathbb{C} ; \mathbb{H})$, defined by

$$\hat{I}(\mathbb{C} ; \mathbb{H}) = \sum_{i,k=1}^N \Pr_{\mathcal{M}}(B_{ik}) \log \frac{\Pr_{\mathcal{M}}(B_{ik})}{\Pr_{\mathcal{M}}(\mathbb{H} \in \text{cl}(k)) \Pr_{\mathcal{M}}(c_i)}, \quad (4)$$

is a lower bound on $I(\mathcal{C}; \mathbb{H})$. respectively.) We examine $\widehat{I}(\mathcal{C}; \mathbb{H})$ to lower bound $I(\mathcal{C}; \mathbb{H})$.

Two relevant special cases in which we can obtain such lower bounds are provided by the following lemmas.

LEMMA 4.1. *If distribution \mathcal{D} is uniform over an ϵ -well-separated subclass of \mathcal{C} containing N elements then the hypotheses output by any $(\mathcal{P}, \mathcal{M})$ -learning function for \mathcal{C} have the property that*

$$I(\mathcal{C}; \mathbb{H}) \geq (1 - \delta)\log(N - 1) - 1 .$$

Proof. We can use the classical Fano's inequality (see, e.g., Cover and Thomas [14, Theorem 2.11.1]) in the discretized setting described above to lower bound the partitional mutual information $\widehat{I}(\mathcal{C}; \mathbb{H})$ in terms of the probability of an “error”,

$$\Pr_{\mathcal{M}}(\text{error}) = 1 - \sum_{i=1}^N \Pr_{\mathcal{M}}(\mathcal{C} = c_i \text{ and } \mathbb{H} \in \text{cl}(i)).$$

Fano's inequality states that for any random variable Y ,

$$H(\mathcal{C} | Y) \leq \mathcal{H}(\Pr_{\mathcal{M}}(\text{error})) + \Pr_{\mathcal{M}}(\text{error}) \log(|\mathcal{C}_\epsilon| - 1) .$$

In particular, we will let Y be that index i such that $h \in \text{cl}(i)$. Using fact F2 and the fact that \mathcal{D} is uniform we can write

$$\begin{aligned} \widehat{I}(\mathcal{C}; \mathbb{H}) &= H(\mathcal{C}) - H(\mathcal{C} | Y) \\ &\leq \log(N - 1) - \mathcal{H}(\Pr_{\mathcal{M}}(\text{error})) - \Pr_{\mathcal{M}}(\text{error}) \log(N - 1) \\ &\leq (1 - \Pr_{\mathcal{M}}(\text{error})) \log(N - 1) - 1. \end{aligned}$$

The probability of “error”, $\Pr_{\mathcal{M}}(\text{error})$, is a lower bound on the probability that the learning function produces a hypothesis that is ϵ -far from the target because $\Pr_{\mathcal{M}}(\text{error})$ does not include the probability that the target is c_1 and the hypothesis is ϵ -far from all c_i .

Therefore, $\Pr_{\mathcal{M}}(\text{error}) \leq \delta$ and

$$\widehat{I}(\mathcal{C}; \mathbb{H}) \geq (1 - \delta) \log(N - 1) - 1,$$

completing the proof. ■

LEMMA 4.2. *If $\delta < 1/2$ and the distribution is $\mathcal{D} = (d, 1 - d)$ over two ϵ -well-separated concepts in \mathcal{C} , then the hypotheses output by any $(\mathcal{P}, \mathcal{M})$ -learning function for \mathcal{C} have the property that*

$$I(\mathcal{C}; \mathbb{H}) \geq \mathcal{H}((1 - \delta)d + \delta(1 - d)) - \mathcal{H}(\delta)$$

Proof. We set for brevity $p_{12} = \Pr_{\mathcal{M}}(\mathbb{H} \in \text{cl}(2) | c_1)$ and $p_{21} = \Pr_{\mathcal{M}}(\mathbb{H} \in \text{cl}(1) | c_2)$.

Recall that $\text{cl}(1)$ includes all h that are neither ϵ -close to c_1 nor ϵ -close to c_2 , and thus p_{12} may underestimate the probability of error when the target is c_1 . We have

$$\Pr_{\mathcal{M}}(\mathbb{H} \in \text{cl}(1)) = (1 - p_{12})d + p_{21}(1 - d)$$

and the partitional mutual information

$$\widehat{I}(\mathcal{C}; \mathbb{H}) = \mathcal{H}((1 - p_{12})d + p_{21}(1 - d)) - d\mathcal{H}(p_{12}) - (1 - d)\mathcal{H}(p_{21}). \quad (5)$$

Now, a derivative argument shows that for any fixed d and $p_{21} \in [0, 1]$, $\widehat{I}(\mathcal{C}; \mathbb{H})$ is non-increasing when $p_{12} \in [0, 1 - p_{21}]$. Similarly, for any fixed d and $p_{12} \in [0, 1]$, $\widehat{I}(\mathcal{C}; \mathbb{H})$ is non-increasing when $p_{21} \in [0, 1 - p_{12}]$. Since the PAC-learning constraints require that both p_{12} and $p_{21} < \delta$ (which is less than $1/2$), we obtain a lower bound on $\widehat{I}(\mathcal{C}; \mathbb{H})$ by substituting δ for p_{12} and p_{21} in (5). This yields the bound of the lemma. ■

Let “is- ϵ -far” be the event that the learning algorithm outputs a hypothesis that is ϵ -far from every element in the ϵ -well-separated set. If this event has positive probability then an additional $(1 - \delta)\Pr_{\mathcal{M}}(\text{is-}\epsilon\text{-far})$ can be added to the bounds of Lemmas 4.1 and 4.2. As mentioned above, we cannot assume that $\Pr_{\mathcal{M}}(\text{is-}\epsilon\text{-far})$ is positive, and thus we state the lemmas in their simpler form.

The reader should observe the different nature of the lower bounds contained in the two previous lemmas. Lemma 4.1 measures the complexity of a class \mathcal{C} through the size of an ϵ -cover for \mathcal{C} , but it applies only to a uniform \mathcal{D} . Lemma 4.2, on the other hand, does not exploit the complexity of \mathcal{C} . It holds for any non-trivial \mathcal{C} and it allows some (though small) flexibility in choosing \mathcal{D} .

Remark. The preceding lemmas could also have been obtained through the following more general approach for lower bounding $\widehat{I}(\mathcal{C}; \mathbb{H})$. The difficulty is that $\widehat{I}(\mathcal{C}; \mathbb{H})$ depends not only on \mathcal{D} , but also on the probabilities $\Pr_{\mathcal{M}}(\mathbb{H} \in \text{cl}(k) | c_i)$ which depend on the learning function. In order to get a general bound on $\widehat{I}(\mathcal{C}; \mathbb{H})$ we examine an optimization problem based on how the learning function can affect these probabilities.

Set for brevity $p_{ik} = \Pr_{\mathcal{M}}(\mathbb{H} \in \text{cl}(k) | c_i)$, so that $\Pr_{\mathcal{M}}(B_{ik}) = p_{ik}d_i$ and $\Pr_{\mathcal{M}}(\mathbb{H} \in \text{cl}(k)) = \sum_{i=1}^N p_{ik}d_i$.

Since, with probability $1 - \delta$, the learning function must produce a hypothesis that is ϵ -close to the target, we can write $\sum_{k=1, k \neq i}^N p_{ik} \leq \delta$. Adding the probability constraints that $\sum_{k=1}^N p_{ik} = 1$ and $p_{ik} \geq 0$ gives us the following minimization problem.

$$\min_{\Omega} \widehat{I}(\mathcal{C}; \mathbb{H}) \quad (6)$$

$$\Omega = \left\{ [p_{ik}]_{i,k=1}^{N,N} : \sum_{k=1, k \neq i}^N p_{ik} \leq \delta, \sum_{k=1}^N p_{ik} = 1, p_{ik} \geq 0 \right\}$$

As the p_{ik} values associated with any learning function represent a feasible solution, the optimal value of the minimization problem gives a general lower bound on $\widehat{I}(c ; \mathbb{H})$ as a function of \mathcal{D} .

Since $\widehat{I}(c ; \mathbb{H})$ is *convex* on Ω for any \mathcal{D} (see, e.g., Cover and Thomas [14, p. 31]) and the constraints defining Ω are linear, a standard Kuhn-Tucker analysis can be attempted. Unfortunately, the solution to (6) appears to have an easy form only for the two cases covered by the above lemmas.

5. APPLICATIONS

This section presents several applications of the method outlined in Section 3. It is aimed at revisiting well-known lower bounds as well as at showing new ones. We feel that this section illustrates the main point of the paper: all these sample size lower bounds have the same underlying structure.

5.1. Malicious Noise

The malicious noise model was introduced by Kearns and Li [26] as a way to formalize the worst possible kind of noise in the examples. This noise model starts with an error-free sample $S_c(x^m)$ of the target c , where x^m is drawn from the underlying distribution \mathcal{P}^m . For each example in the sample an independent coin with probability η of heads is tossed. If the coin for example (x, l) comes up heads, then (x, l) is replaced by a corrupted example (\hat{x}, \hat{l}) about which no assumptions can be made. Otherwise, the example is left unchanged.

In particular, the corrupted examples (\hat{x}, \hat{l}) can be maliciously chosen by an adversary⁴ that knows $\epsilon, \delta, c, \mathcal{P}, \mathcal{M}$ and the internal state of a device computing A . Hence the factorization of $\mathcal{M}(\hat{x}^m, \hat{l}^m, r \mid c)$ depends on the specific noise process. As a short-hand, we call this model “ $(\mathcal{P}, \mathcal{M})$ -learning in the malicious noise model”, leaving the details of \mathcal{M} as a separate issue.

Results similar to Theorem 5.1 below have been shown by Kearns and Li [26] and by Cesa-Bianchi et al. [13]. In the former paper no sample size lower bounds are proved, while in the latter paper the authors prove a looser bound by a Bayesian argument that involves a subtle study of the properties of the binomial distribution (see Fact 3.2 therein). We prove a tighter bound for this model using a more direct mutual information argument. Our bound adds a $\log(1/\delta)$ factor to the bound of [13], and illustrates how the RHS of (3) can guide the choice of \mathcal{D} .

THEOREM 5.1. *Let \mathcal{C} and \mathcal{P} be an ϵ -binary pair on $(\mathcal{X}, \mathcal{B})$ (defined in Section 2). If A is a $(\mathcal{P}, \mathcal{M})$ -learning function for \mathcal{C} in the malicious noise model with rate $0 \leq \eta < \frac{\epsilon}{1+\epsilon}$, $\Delta = \frac{\epsilon}{1+\epsilon} - \eta$, $\epsilon \leq 1/2$ and $\delta < 1/2$, then the following relation on m*

⁴There is a subtlety here that is worth mentioning: we may distinguish whether the action of the adversary for the i 'th example only depends on the previous $i - 1$ examples or it is allowed to depend also on the “future” $m - i$ examples. However the difference between the two models seems to be relevant [13] only when one is proving lower bounds for *special classes* of learning functions (e.g., disagreement minimization). Thus we will not be concerned with this subtlety as we are not restricting the behavior of A .

must hold

$$m \geq \frac{\eta(1-2\delta) \ln \frac{1-\delta}{\delta}}{\Delta^2(1+\epsilon)^2} = \Omega\left(\frac{\eta}{\Delta^2} \ln \frac{1}{\delta}\right).$$

Proof. Since \mathcal{C} and \mathcal{P} are an ϵ -binary pair, we can assume that $\mathcal{X} = \{x_1, x_2\}$ and $\mathcal{P}(x_1) = 1 - \epsilon$, so $\mathcal{C}_\epsilon = \mathcal{C}$ is an ϵ -well-separated set. For any $\mathcal{D} = (d, 1 - d)$, applying Lemma 4.2 gives

$$I(\mathcal{C} ; \mathbb{H}) \geq \mathcal{H}((1 - \delta)d + \delta(1 - d)) - \mathcal{H}(\delta).$$

The malicious adversary behaves as follows [26, 13]: if noise occurs and c is the target, it replaces the current example (x, l) by $(x_2, 1 - I_c(x_2))$. The following induced distribution results: $\mathcal{M}(\hat{x}^m, \hat{l}^m | c, r) = \prod_{i=1}^m \mathcal{M}(\hat{x}_i, \hat{l}_i | c, r)$, where (independent of r)

$$\begin{aligned} \mathcal{M}(x_1, 0 | c_1, r) &= 0, & \mathcal{M}(x_1, 0 | c_2, r) &= 0, \\ \mathcal{M}(x_1, 1 | c_1, r) &= (1 - \eta)(1 - \epsilon), & \mathcal{M}(x_1, 1 | c_2, r) &= (1 - \eta)(1 - \epsilon), \\ \mathcal{M}(x_2, 0 | c_1, r) &= (1 - \eta)\epsilon, & \mathcal{M}(x_2, 0 | c_2, r) &= \eta, \\ \mathcal{M}(x_2, 1 | c_1, r) &= \eta, & \mathcal{M}(x_2, 1 | c_2, r) &= (1 - \eta)\epsilon. \end{aligned}$$

Also, by Fact F2, $I(\hat{X}_i, \hat{L}_i ; \mathcal{C} | R) = H(\hat{X}_i, \hat{L}_i | R) - H(\hat{X}_i, \hat{L}_i | \mathcal{C}, R)$, and for the \mathcal{M} shown above it is easy to verify that

$$H(\hat{X}_i, \hat{L}_i | R) = \mathcal{H}[(1 - \eta)(1 - \epsilon), d(1 - \eta)\epsilon + (1 - d)\eta, d\eta + (1 - d)(1 - \eta)\epsilon]$$

and

$$H(\hat{X}_i, \hat{L}_i | \mathcal{C}, R) = \mathcal{H}[(1 - \eta)(1 - \epsilon), (1 - \eta)\epsilon, \eta].$$

The difference $H(\hat{X}_i, \hat{L}_i | R) - H(\hat{X}_i, \hat{L}_i | \mathcal{C}, R)$ is the concern of the following fact, whose proof is in the appendix.

Fact F5. If $0 < \epsilon \leq 1$, $0 \leq \eta < \frac{\epsilon}{1+\epsilon}$, and $0 \leq d \leq 1$ then

$$\begin{aligned} & \mathcal{H}[(1 - \eta)(1 - \epsilon), d(1 - \eta)\epsilon + (1 - d)\eta, d\eta + (1 - d)(1 - \eta)\epsilon] - \mathcal{H}[(1 - \eta)(1 - \epsilon), (1 - \eta)\epsilon, \eta] \\ &= (\eta + (1 - \eta)\epsilon) \left[\mathcal{H}\left(\frac{d\eta + (1 - d)(1 - \eta)\epsilon}{\eta + (1 - \eta)\epsilon}\right) - \mathcal{H}\left(\frac{\eta}{\eta + (1 - \eta)\epsilon}\right) \right]. \end{aligned}$$

With these expressions for $I(\mathcal{C} ; \mathbb{H})$ and $I(\hat{X}_i, \hat{L}_i ; \mathcal{C} | R)$, we can apply (3) to see that $m(d)$ is at least

$$\frac{\mathcal{H}((1 - \delta)d + \delta(1 - d)) - \mathcal{H}(\delta)}{(\eta + (1 - \eta)\epsilon) \left[\mathcal{H}\left(\frac{d\eta + (1 - d)(1 - \eta)\epsilon}{\eta + (1 - \eta)\epsilon}\right) - \mathcal{H}\left(\frac{\eta}{\eta + (1 - \eta)\epsilon}\right) \right]}. \quad (7)$$

Note that this bound is a function of d and $\sup_{d \in [0, 1]} m(d)$ is at least $\lim_{d \rightarrow 0^+} m(d)$. Since both the numerator and the denominator of (7) vanish as

$d \rightarrow 0$ we use De l'Hôpital's rule to evaluate this limit. Some simple algebra leads to

$$\lim_{d \rightarrow 0^+} m(d) \geq \frac{(1 - 2\delta) \ln \frac{1-\delta}{\epsilon \eta d \alpha}}{(\epsilon - \eta(1 + \epsilon)) \ln \frac{\epsilon(1-\eta)}{\eta}}. \quad (8)$$

Since $\epsilon - \eta(1 + \epsilon) = \Delta(1 + \epsilon)$ and $\frac{\epsilon(1-\eta)}{\eta} = 1 + \frac{\Delta(1+\epsilon)}{\eta}$ we rewrite the denominator as

$$\Delta(1 + \epsilon) \ln \left(1 + \frac{\Delta(1 + \epsilon)}{\eta} \right) \leq \frac{\Delta^2(1 + \epsilon)^2}{\eta}$$

with the inequality following from the well-known fact $\ln(1+x) \leq x$. This concludes the proof. ■

The previous bound is meaningful only when η is close to the information-theoretic limit $\frac{\epsilon}{1+\epsilon}$, but has the advantage of *diverging* as $\delta \rightarrow 0$. By selecting $d = 1/2$ (instead of $d = 0$), a completely analogous argument proves the bound $m \geq O\left(\frac{\eta+\Delta}{\Delta^2}\right)$, which does not vanish when $\eta = 0$. In fact, any choice of $d \in [0, 1]$ gives a bound on m .

The same trade-off (a bound that vanishes for $\eta = 0$ but diverges as $\delta \rightarrow 0$ versus a bound which does not vanish for $\eta = 0$ but does not contain a dependence on δ) will recur. Both Theorem 5.4 and Theorem 5.6 are phrased to emphasize the dependence on δ . For these theorems also, if a more moderate value is used for d then the resulting bounds are meaningful when $\eta = 0$ (but lose their dependence on δ).

5.2. Classification noise

The classification noise model was introduced by Angluin and Laird [4] as a way to model the mildest kind of error in the examples. Here each example (x, l) of the error-free sample $S_c(x^m)$ is processed by a noise process that independently with probability $1 - \eta$ leaves it unchanged, and with probability η flips the label l into $\hat{l} = 1 - l$. As a short-hand, we call this “ $(\mathcal{P}, \mathcal{P}^m)$ -learning in the classification noise model”, to be understood as $(\mathcal{P}, \mathcal{M})$ -learning in which \mathcal{M} factors as described in Section 2.

When η is bounded away from 0, say $\eta \geq 1/100$, the following theorem adds a logarithmic factor to a bound proved in Simon [31] and Apolloni and Gentile [6], by considering a very natural family of concept classes over the unit interval. In order to prove it we make use of the following technical lemma.

LEMMA 5.1. *Let $f(\alpha, \eta) = \mathcal{H}(\eta + \alpha(1 - 2\eta)) - \mathcal{H}(\eta)$.*

1. *If $\alpha \in [0, 1/2]$ and $\eta \in (0, 1)$ then*

$$f(\alpha, \eta) \leq \frac{2\alpha(1 - 2\eta)^2}{(\ln 2)(1 - 2\alpha)(1 - (1 - 2\eta)^2)}.$$

2. *If $\alpha, \eta > 0$ and $\eta + \alpha \leq 1/2$ then $f(\alpha, \eta) \leq \alpha \log(1/\eta)$.*

3. *$f(\alpha, 0) = \mathcal{H}(\alpha) \leq \alpha \log(e/\alpha)$ (where e is the base of \ln).*

Proof. 1. See [6]; 2. In appendix; 3. Easily derived from the fact $-(1 - \alpha) \ln(1 - \alpha) \leq \alpha$, for all $\alpha \in [0, 1]$. ■

THEOREM 5.2. *Let \mathcal{C}_k be the class of unions of $k \geq 1$ intervals on the unit interval $\mathcal{X} = [0, 1]$ and \mathcal{P} be uniform over \mathcal{X} . If A is a $(\mathcal{P}, \mathcal{P}^m)$ -learning function for \mathcal{C}_k in the classification noise model with rate $\eta \neq 1/2$, $\epsilon < 1/16$ and $\delta < 1$, then:*

1. *If $0 < \eta_0 \leq \eta \leq 1 - \eta_0$ then*

$$m = \Omega \left(\frac{d_{VC}(\mathcal{C}_k)}{\epsilon(1 - 2\eta)^2} \log \frac{1}{\epsilon} \right)$$

where the hidden constant in this Ω -expression depends on η_0 .

2. *If $\log(1/\eta) = o(\log(1/\epsilon))$ when $\epsilon \rightarrow 0$ and $\eta \rightarrow 0$ then*

$$m = \omega \left(\frac{d_{VC}(\mathcal{C}_k)}{\epsilon} \right).$$

3. *If $\eta = 0$ then*

$$m = \Omega \left(\frac{d_{VC}(\mathcal{C}_k)}{\epsilon} \right).$$

Proof. Let us consider the subclass $\hat{\mathcal{C}}_k$ of \mathcal{C}_k defined as follows. Set $T = \frac{1}{4\epsilon}$. In order to simplify the notation, we assume that T is an integer. We split $[0, 1]$ into k intervals, I_1 through I_k , each of length $1/k$. We then split each interval I_i into T sub-intervals, I_{i1} through I_{iT} , of length $1/Tk$. Thus for $i = 1, \dots, k$ and $j = 1, \dots, T$, interval $I_i = [\frac{i-1}{k}, \frac{i}{k}]$ and sub-interval $I_{ij} = [\frac{i-1}{k} + \frac{j-1}{Tk}, \frac{i-1}{k} + \frac{j}{Tk}]$. Define

$$\hat{\mathcal{C}}_k = \{I_{1j_1} \cup I_{2j_2} \cup \dots \cup I_{kj_k} : j_1, j_2, \dots, j_k \in \{1, \dots, T\}\}$$

so that each $c \in \hat{\mathcal{C}}_k$ is the union of k sub-intervals, one from each different interval. The class $\hat{\mathcal{C}}_k$ is defined as a function of the accuracy ϵ . Since $\hat{\mathcal{C}}_k$ is a subset of \mathcal{C}_k for every ϵ , we are bounding the difficulty of learning the *larger* class \mathcal{C}_k by considering, for every value of ϵ , a hard subclass $\hat{\mathcal{C}}_k$ of \mathcal{C}_k .

We will lower bound the cardinality of a largest ϵ -well-separated subclass of $\hat{\mathcal{C}}_k$ by underestimating $N(\hat{\mathcal{C}}_k, 2\epsilon, \mathcal{P})$ (Lemma 2.1). Two concepts in $\hat{\mathcal{C}}_k$, are 2ϵ -close if and only if they share the same sub-intervals in more than $k/2$ of the intervals. For any given $c_0 \in \hat{\mathcal{C}}_k$, the number of concepts in $\hat{\mathcal{C}}_k$ that are 2ϵ -close to c_0 is thus

$$\sum_{l=0}^{\lceil k/2 \rceil - 1} \binom{k}{l} (T-1)^l \leq 2^k \sum_{l=0}^{\lceil k/2 \rceil - 1} T^l \leq 2^k T^{\lceil k/2 \rceil}.$$

Since $|\hat{\mathcal{C}}_k| = T^k$, every 2ϵ -cover of $\hat{\mathcal{C}}_k$ contains at least $T^k / 2^k T^{\lceil k/2 \rceil} \geq \frac{1}{2} (T/4)^{\lceil k/2 \rceil}$ concepts. Now Lemma 2.1 implies that the cardinality of a largest ϵ -well-separated subclass $\hat{\mathcal{C}}_{k\epsilon}$ of $\hat{\mathcal{C}}_k$ satisfies

$$|\hat{\mathcal{C}}_{k\epsilon}| \geq \frac{1}{2} \left(\frac{T}{4} \right)^{\lceil k/2 \rceil} = \frac{1}{2} \left(\frac{1}{16\epsilon} \right)^{\lceil k/2 \rceil}. \quad (9)$$

Let \mathcal{D} be uniform over $\hat{\mathcal{C}}_{k\epsilon}$. In this noise model X^m , c and R are independent, so by the additivity of mutual information and Fact F2 we have

$$I(X^m, \hat{L}^m ; c | R) = I(\hat{L}^m ; c | X^m) = H(\hat{L}^m | X^m) - H(\hat{L}^m | X^m, c).$$

We overestimate $H(\hat{L}^m | X^m)$ by $H(\hat{L}^m)$ and note that $H(\hat{L}^m | X^m, c) = m\mathcal{H}(\eta)$ for the classification noise model. Applying (2) and Lemma 4.1 now results in

$$(1 - \delta)\log(|\hat{\mathcal{C}}_{k\epsilon}| - 1) - 1 \leq H(\hat{L}^m) - m\mathcal{H}(\eta). \quad (10)$$

But $H(\hat{L}^m) = m\mathcal{H}(\Pr_{\mathcal{P}}(\hat{L}_i = 1)) = m\mathcal{H}(\eta + 4\epsilon(1 - 2\eta))$ for any \mathcal{D} , since $\Pr_{\mathcal{P}}(c) = 4\epsilon$ for every $c \in \hat{\mathcal{C}}_k$.

Combining (10), (9), the fact that $d_{VC} = d_{VC}(\mathcal{C}_k) = 2k$, and underestimating the logarithm yields

$$(1 - \delta)\left(\frac{d_{VC}}{4} - 1\right)\log\left(\frac{1}{16\epsilon} - 1\right) - 1 \leq m\mathcal{H}(\eta + 4\epsilon(1 - 2\eta)) - m\mathcal{H}(\eta) \quad (11)$$

(The assumption $\epsilon < 1/16$ prevents a non-positive argument in the log.) Part 1 of the theorem now follows from bound 1 of Lemma 5.1, after noting that the denominator of bound 1 is at least a constant. Part 2 follows from (11), bound 2 of Lemma 5.1, and a comparison of the factors $\log(1/\epsilon)$ and $\log(1/\eta)$ which occur in opposite sides of the inequality. Part 3 follows from (11) and bound 3 of Lemma 5.1. ■

The lower bounds of Theorem 5.2 actually hold for every pair of \mathcal{C} and \mathcal{P} such that \mathcal{C} contains a subclass $\hat{\mathcal{C}}$ of concepts each with \mathcal{P} -measure $O(\epsilon)$ and where $\log N(\hat{\mathcal{C}}, 2\epsilon, \mathcal{P}) = \Omega(d_{VC}(\mathcal{C})\log(1/\epsilon))$. For instance, by the embedding technique⁵ of Helmbold et al. [24], the same lower bound holds for various common geometric concept classes such as axis-parallel rectangles in \mathcal{R}^n and half-spaces.

Moreover, we remark that such a lower bound is the best possible when η is bounded away from 0. Indeed a matching information-theoretic upper bound on the sample size required to $(\mathcal{P}, \mathcal{P}^m)$ -learn any concept class \mathcal{C} of finite VC-dimension d_{VC} in the classification noise model is provided by the analysis in Laird [28, p. 190]. This analysis of disagreement minimization is valid only for finite size classes. Disregarding the dependence on δ , the sample complexity there is $O\left(\frac{\log|\mathcal{C}|}{\epsilon(1-2\eta)^2}\right)$. Since we are assuming that the learning function *knows* the distribution \mathcal{P} , it knows in principle a smallest ϵ -cover of \mathcal{C} w.r.t. \mathcal{P} . Dudley contains a proof [16, Theorem 9.3.1] that smallest ϵ -covers contain at most $K\left(\frac{1}{\epsilon}\right)^{O(d_{VC})}$ elements. Here K is a constant that depends on the concept class but neither on \mathcal{P} nor on ϵ . Using Laird's analysis of minimizing disagreements on a smallest $\epsilon/2$ -cover of \mathcal{C} , one could obtain the sample size bound $m = O\left(\frac{\log K + d_{VC}(\mathcal{C})\log\frac{1}{\epsilon}}{\epsilon(1-2\eta)^2}\right)$, which matches our lower bound as $\epsilon \rightarrow 0$.

When $\eta = 0$, the 1-inclusion graph algorithm of Haussler, Littlestone, and Warmuth [22] can be converted into an algorithm requiring $O\left(\frac{d_{VC}}{\epsilon} \log\frac{1}{\delta}\right)$ examples. Therefore the $\log(1/\epsilon)$ factor cannot occur in noise-free lower bounds.

⁵The proof of Theorem 5.2 requires that the entropy of the observed labels be small. Therefore, embedding techniques based on initial segments (like those of Haussler et al. [22]) are more difficult to apply in this context.

It is natural to expect that the ideas behind Theorem 5.2 can lead to a similar improvement in the bounds for learning in the malicious noise model of Section 5.1. However, this remains an open problem.

5.3. Classification noise with drifting distributions

Here we adopt the terminology of the drifting distribution model introduced by Bartlett [8] and further explored by Bartlett and Helmbold [9] and Barve and Long [10]. Notice, however, that we are still considering a *batch* learning setting.

Following Bartlett [8], we define a distance between probability distributions \mathcal{P}_1 and \mathcal{P}_2 on \mathcal{X} as follows:

$$\text{dist}(\mathcal{P}_1, \mathcal{P}_2) = \sup_{A \in \mathcal{B}} |\Pr_{\mathcal{P}_1}(A) - \Pr_{\mathcal{P}_2}(A)|.$$

A sequence of probability distributions $\{\mathcal{P}_j\}_{j=1 \dots m}$ is called γ -*admissible* if $\text{dist}(\mathcal{P}_j, \mathcal{P}_{j+1}) \leq \gamma$, for $j = 1, \dots, m - 1$. In the drifting distribution model, \mathcal{M} factors as

$$\mathcal{M}(x^m, \hat{l}^m, r \mid c) = \prod_{j=1}^m \mathcal{P}_j(x_j) \mathcal{M}(\hat{l}^m \mid x^m, c) \mathcal{M}(r),$$

and $\mathcal{M}(\hat{l}^m \mid x^m, c)$ factors as for the classification noise model. For brevity, we call this model “ $(\mathcal{P}_m, \prod_j \mathcal{P}_j)$ -learning in the classification noise model”. Here we assume that the testing distribution \mathcal{P}_m is the last distribution in the sequence.

In this subsection we give two lower bound theorems. The first is a generalization of results in [8, 31, 6]. The second is a generalization of a result proved by Aslam and Decatur [7]. The method we employ provides particularly clean proofs and yields far better constants when specialized to $(\mathcal{P}, \mathcal{P}^m)$ -learning in the classification noise model (without distribution drift). The bound related to Theorem 5.3 was proved in [31] by appealing to the central limit theorem and in [6] with worse constants and only for η bounded away from 0. Theorem 5.3 can also be derived by the Bayesian argument in Barve and Long [10, Theorem 18], but the constants therein are exceedingly large.

THEOREM 5.3. *Let \mathcal{C} be a concept class on $(\mathcal{X}, \mathcal{B})$, $d_{VC}(\mathcal{C}) = d_{VC} \geq 68$, $\epsilon < 1/16$ and $\delta \leq 1/40$. Then for every m there exists a γ -admissible sequence of distributions $\{\mathcal{P}_j\}_{j=1 \dots m}$ on \mathcal{X} , with $\gamma = \frac{320\epsilon^2(1-\mathcal{H}(\eta))}{d_{VC}-1}$, such that $(\mathcal{P}_m, \prod_j \mathcal{P}_j)$ -learning in the classification noise model is impossible.*

Proof. We restrict our attention to a shattered set $\mathcal{X} = \{x_1, \dots, x_{d_{VC}}\}$ and the sub-class $\mathcal{C}' = \{c \in 2^{\{x_1, \dots, x_{d_{VC}}\}} : I_c(x_{d_{VC}}) = 0\}$, i.e., those concepts that label the instance $x_{d_{VC}}$ with zero. To simplify the notation we assume that $t = \frac{d_{VC}-1}{20\epsilon(1-\mathcal{H}(\eta))}$ is an integer. Define $\{\mathcal{P}_j\}_{j=1 \dots m}$ to be the following γ -admissible sequence of dis-

tributions on \mathcal{X} .

$$\begin{aligned} \text{for } j = 1, \dots, m-t, \quad \mathcal{P}_j(x_i) &= \begin{cases} 0 & \text{for } i = 1, \dots, d_{VC} - 1 \\ 1 & \text{for } i = d_{VC} \end{cases} \\ \text{for } j = m-t+1, \dots, m, \quad \mathcal{P}_j(x_i) &= \begin{cases} \frac{\gamma(j-m+t)}{d_{VC}-1} & \text{for } i = 1, \dots, d_{VC} - 1 \\ 1 - \gamma(j-m+t) & \text{for } i = d_{VC} . \end{cases} \end{aligned}$$

The sequence $\{\mathcal{P}_j\}_{j=1\dots m}$ is γ -admissible, and the choices of γ and t ensure that the probability of $x_{d_{VC}}$ is always at least $1 - 16\epsilon$. In fact, the testing distribution \mathcal{P}_m is actually:

$$\mathcal{P}_m(x_i) = \begin{cases} \frac{16\epsilon}{d_{VC}-1} & \text{for } i = 1, \dots, d_{VC} - 1 \\ 1 - 16\epsilon & \text{for } i = d_{VC}. \end{cases}$$

Let \mathcal{C}'_ϵ be a largest ϵ -well-separated subclass of \mathcal{C}' (with respect to \mathcal{P}_m), $N = |\mathcal{C}'_\epsilon|$, and \mathcal{D} be uniform over \mathcal{C}'_ϵ . Recall that for this noise model, $I(X^m, \hat{L}^m; c | R) = I(\hat{L}^m; c | X^m) = H(\hat{L}^m | X^m) - H(\hat{L}^m | X^m, c) = H(\hat{L}^m | X^m) - m\mathcal{H}(\eta)$. Combining this with (2) and Lemma 4.1 gives the following necessary condition on m

$$(1 - \delta)\log(N - 1) - 1 \leq H(\hat{L}^m | X^m) - m\mathcal{H}(\eta) . \quad (12)$$

From computations like those in the proof of Theorem 5.2 (see Corollary 1 in [6] for a very similar result), it follows that

$$\log(N - 1) > \frac{11}{25}(d_{VC} - 2) \quad (13)$$

for $d_{VC} \geq 7$ and $\epsilon < 1/16$.

We now upper bound the entropy $H(\hat{L}^m | X^m)$, where $\hat{L}^m = (\hat{L}_1, \dots, \hat{L}_m)$ and $X^m = (X_1, \dots, X_m)$. Since reducing the conditioning only increases the entropy, we have

$$\begin{aligned} H(\hat{L}^m | X^m) &\leq \sum_{j=1}^m H(\hat{L}_j | X_j) \\ &= \sum_{j=1}^m E_{\mathcal{P}_j} \left[\mathcal{H}(\text{Pr}_{\mathcal{M}}(\hat{L}_j = 1 | X_j = x)) \right] \end{aligned} \quad (14)$$

The first $m-t$ instances will be $x_{d_{VC}}$ since it has unit probability. Furthermore, $I_c(x_{d_{VC}}) = 0$ for all of the concepts $c \in \mathcal{C}'$. Therefore \hat{L}_j is 1 (for $1 \leq j \leq m-t$) if and only if the label \hat{L}_j is noisy, so $E_{\mathcal{P}_j} \left[\mathcal{H}(\text{Pr}_{\mathcal{M}}(\hat{L}_j = 1 | X_j = x)) \right] = \mathcal{H}(\eta)$ for these examples.

Similarly, when $j > m-t$ the instance $X_j = x_{d_{VC}}$ with probability $1 - \gamma(j-m+t)$. When X_j is not $x_{d_{VC}}$, then the entropy of \hat{L}_j is at most 1. Continuing from

inequality (14),

$$\begin{aligned}
H(\hat{L}^m | X^m) &\leq \sum_{j=1}^{m-t} \mathcal{H}(\eta) + \sum_{j=m-t+1}^m ((1 - \gamma(j - m + t))\mathcal{H}(\eta) + \gamma(j - m + t) \cdot 1) \\
&= m\mathcal{H}(\eta) + \gamma(1 - \mathcal{H}(\eta)) \sum_{j=m-t+1}^m (j - m + t) \\
&= m\mathcal{H}(\eta) + \gamma(1 - \mathcal{H}(\eta)) \frac{t(t+1)}{2} \\
&\leq m\mathcal{H}(\eta) + 8\epsilon t(1 - \mathcal{H}(\eta)) + 8\epsilon
\end{aligned}$$

where the last step used the fact that $\gamma t = 16\epsilon$.

Since $t = \frac{d_{VC}-1}{20\epsilon(1-\mathcal{H}(\eta))}$ and $\epsilon \leq 1/16$,

$$H(\hat{L}^m | X^m) - m\mathcal{H}(\eta) \leq \frac{2}{5}(d_{VC} - 1) + \frac{1}{2}.$$

Combining this with inequalities (12) and (13) gives

$$\frac{11}{25}(1 - \delta)(d_{VC} - 2) - 1 \leq \frac{2}{5}(d_{VC} - 1) + \frac{1}{2},$$

which cannot be satisfied for $\delta \leq 1/40$ and $d_{VC} \geq 68$. Thus for every m we have a sequence of distributions that prevents learning, thereby completing the proof. ■

The contradiction in the proof of Theorem 5.3 requires that δ is small while γ and d_{VC} are sufficiently large. The constants in the statement of the theorem are only one way of obtaining the desired contradiction. These constants can be traded off against each other, so that an analogous theorem can be proven for smaller d_{VC} (for example) if γ is made larger or δ smaller.

THEOREM 5.4. *Let \mathcal{C} and \mathcal{P}_m be an ϵ -binary pair on $(\mathcal{X}, \mathcal{B})$, with $\epsilon \leq 1/2$ and $\delta < 1/2$. Then for every m there exists a γ -admissible sequence of distributions $\{\mathcal{P}_j\}_{j=1\dots m}$ on \mathcal{X} , with $\gamma = O\left(\frac{\epsilon^2(1-2\eta)^2}{\eta \ln \frac{1}{\delta}}\right)$, such that $(\mathcal{P}_m, \prod_j \mathcal{P}_j)$ -learning in the classification noise model is impossible.*

Proof. As this proof is similar to that of Theorem 5.1, we will only sketch the main steps.

Consider the class $\mathcal{C} = \{c_1, c_2\}$ with $c_1 = \{x_1, x_2\}$, $c_2 = \{x_1\}$. Pick any $\gamma > \frac{\epsilon^2(1-2\eta)^2}{2\eta(1-2\delta)\ln \frac{1-\delta}{\delta} - \epsilon(1-2\eta)^2}$. Set $t = \epsilon/\gamma$, and we again assume that t is an integer. Let $\{\mathcal{P}_j\}_{j=1\dots m}$ be the following γ -admissible sequence of distributions on \mathcal{X} :

$$\begin{aligned}
\text{for } j = 1, \dots, m-t, \quad \mathcal{P}_j(x_i) &= \begin{cases} 1, & \text{for } i = 1, \\ 0 & \text{for } i = 2, \end{cases} \\
\text{for } j = m-t+1, \dots, m, \quad \mathcal{P}_j(x_i) &= \begin{cases} 1 - \gamma(j - m + t), & \text{for } i = 1, \\ \gamma(j - m + t), & \text{for } i = 2. \end{cases}
\end{aligned}$$

\mathcal{C} and \mathcal{P}_m are an ϵ -binary pair so that we can assume $\mathcal{C} = \mathcal{C}_\epsilon$. Let $\mathcal{D} = (d, 1 - d)$ be a distribution over \mathcal{C} . By the noise model it can be easily verified that

$$H(\hat{L}_j | X_j) = H(\hat{L}_j | X_j, \mathbf{c}) = \mathcal{H}(\eta), \quad \text{for } j = 1, \dots, m - t,$$

while

$$\begin{aligned} H(\hat{L}_j | X_j) &= \mathcal{H}(\eta) + \gamma(j - m + t)(\mathcal{H}((1 - \eta)d + \eta(1 - d)) - \mathcal{H}(\eta)), \\ H(\hat{L}_j | X_j, \mathbf{c}) &= \mathcal{H}(\eta), \quad \text{for } j = m - t + 1, \dots, m. \end{aligned}$$

Therefore

$$\begin{aligned} I(X^m, \hat{L}^m; \mathbf{c} | R) &= I(\hat{L}^m; \mathbf{c} | X^m) && \text{(by the additivity of } I) \\ &= H(\hat{L}^m | X^m) - H(\hat{L}^m | X^m, \mathbf{c}) && \text{(by Fact F2)} \\ &= \beta(\mathcal{H}((1 - \eta)d + \eta(1 - d)) - \mathcal{H}(\eta)), \end{aligned}$$

where $\beta = \sum_{j=1}^t \gamma j = \gamma t(t + 1)/2 = \frac{\epsilon}{2} \left(\frac{\epsilon}{\gamma} + 1 \right)$.

We can apply inequality (2) and Lemma 4.2, yielding

$$\mathcal{H}((1 - \delta)d + \delta(1 - d)) - \mathcal{H}(\delta) \leq \beta(\mathcal{H}((1 - \eta)d + \eta(1 - d)) - \mathcal{H}(\eta)).$$

Solving for β shows that β must be at least $\beta(d)$ defined by

$$\beta(d) = \frac{\mathcal{H}((1 - \delta)d + \delta(1 - d)) - \mathcal{H}(\delta)}{\mathcal{H}((1 - \eta)d + \eta(1 - d)) - \mathcal{H}(\eta)}.$$

Now, $\sup_{d \in [0, 1]} \beta(d) \geq \lim_{d \rightarrow 0^+} \beta(d)$, and employing De l'Hôpital's rule yields

$$\lim_{d \rightarrow 0^+} \beta(d) = \frac{(1 - 2\delta) \ln \frac{1 - \delta}{\delta}}{(1 - 2\eta) \ln \frac{1 - \eta}{\eta}}. \quad (15)$$

Since $\frac{1 - \eta}{\eta} = 1 + \frac{1 - 2\eta}{\eta}$ we have $\ln \frac{1 - \eta}{\eta} = \ln(1 + \frac{1 - 2\eta}{\eta}) \leq \frac{1 - 2\eta}{\eta}$. Plugging the last inequality into the denominator of (15) gets the bound

$$\beta \geq \frac{(1 - 2\delta) \eta \ln \frac{1 - \delta}{\delta}}{(1 - 2\eta)^2} \quad (16)$$

Since $\beta = \frac{\epsilon}{2} \left(\frac{\epsilon}{\gamma} + 1 \right)$, we can solve for γ to obtain the claimed bound. \blacksquare

The previous theorems show that the the noise rate η significantly affects the drifting constant γ .

The proof of Theorem 5.3 can be specialized to $(\mathcal{P}, \mathcal{P}^m)$ -learning in the classification noise model without distribution drift (i.e., with $\gamma = 0$). Thus, in the notation of the theorem's proof, $t = m$ and $\mathcal{P}_j = \mathcal{P}_m = \mathcal{P}$, $j = 1, \dots, m$. The resulting version of Theorem 5.3 shows that if A is a $(\mathcal{P}, \mathcal{P}^m)$ -learning function for \mathcal{C} then its sample complexity must satisfy (12). In the proof of Theorem 5.3 we showed that the LHS of (12) is $\Omega(d_{VC})$ and its RHS is $O(m\epsilon(1 - \mathcal{H}(\eta))) = O(m\epsilon(1 - 2\eta)^2)$. By

a more careful analysis of the constants (which is omitted from this paper) we can prove the bound

$$m \geq \frac{d_{VC} - 8}{38\epsilon(1 - \mathcal{H}(\eta))} = \Omega\left(\frac{d_{VC}}{\epsilon(1 - 2\eta)^2}\right)$$

in the non-drifting case.

The proof of Theorem 5.4 can also be specialized in a similar way. We can replace β by $m\epsilon$ and from (16) we obtain

$$m \geq \frac{(1 - 2\delta) \eta \ln \frac{1-\delta}{\delta}}{\epsilon(1 - 2\eta)^2} = \Omega\left(\frac{\eta}{\epsilon(1 - 2\eta)^2} \ln \frac{1}{\delta}\right).$$

5.4. Learning with membership queries

In this section we assume that the learning function A (which is now necessarily a learning *algorithm*) has the additional capability of making *membership queries*, so that A can request the label of any instance x in the domain \mathcal{X} . The algorithm can use an arbitrary (computable) strategy to determine which instances are queried. This strategy can depend not only on the algorithm’s randomization but also on the results of previous queries. We can still adopt the notion of learnability provided by Definition 2.1: by its choice of queries, an algorithm *induces a distribution* over \mathcal{X}^* , the set of all finite sequences on \mathcal{X} . As a short-hand we will speak of a “ $(\mathcal{P}, \mathcal{M})$ -learning algorithm that uses membership queries”, where it is understood that here \mathcal{M} depends on the specific behavior of A .

There is a vast literature related to the problem of PAC learning with membership queries. See, for instance, Angluin [1, 2], Maass and Turàn [29], Sakakibara [30], Angluin et al. [3] and the references in those papers. Perhaps the references closest to our work are Eisenberg and Rivest [18] and Turàn [34]. But in contrast to the former paper, here we are assuming that the learning algorithm *knows* the distribution \mathcal{P} . Compared to the latter paper, we emphasize random examples and membership queries (as a matter of fact, both theorems of this section can easily be extended to arbitrary Yes/No queries), but we are making more general statements in a unifying context. We also assume that the labels of both the random and the chosen examples are subject to classification noise, as described in Section 5.2.

Since the distribution \mathcal{M} depends on which queries the algorithm makes, little can be assumed about \mathcal{M} if we want to obtain a general lower bound. However, we will exploit the fact that each query instance is a (deterministic) function of the algorithm’s randomization and the past examples. We use the term “query model” for those \mathcal{M} having the property that every instance X_i can only depend on the target c through the past examples $(X_1, \hat{L}_1), \dots, (X_{i-1}, \hat{L}_{i-1})$. Therefore query models have the important property that

$$I(X_i; c | (X_1, \hat{L}_1), \dots, (X_{i-1}, \hat{L}_{i-1}), R) = 0. \tag{17}$$

This is equivalent to a “data processing inequality” (data processing inequalities are discussed in [14] and [25]).

THEOREM 5.5. *Let \mathcal{C} be a concept class on $(\mathcal{X}, \mathcal{B})$ and \mathcal{P} be a distribution on \mathcal{X} such that \mathcal{C} has an ϵ -well-separated subclass of cardinality $N \geq 2$. If A is a*

$(\mathcal{P}, \mathcal{M})$ -learning algorithm for \mathcal{C} that uses membership queries in the classification noise model with noise rate $\eta \neq 1/2$ then

$$m \geq \frac{(1 - \delta)\log(N - 1) - 1}{1 - \mathcal{H}(\eta)} = \Omega\left(\frac{\log N}{(1 - 2\eta)^2}\right).$$

Proof. Let \mathcal{D} be the uniform distribution over \mathcal{C}_ϵ . From (2) and Lemma 4.1 we obtain

$$(1 - \delta)\log(N - 1) - 1 \leq I(X^m, \hat{L}^m; \mathbf{c} \mid R).$$

We continue by upper bounding $I(X^m, \hat{L}^m; \mathbf{c} \mid R)$. By the additivity of I ,

$$\begin{aligned} I(X^m, \hat{L}^m; \mathbf{c} \mid R) &= \sum_{i=1}^m I(X_i, \hat{L}_i; \mathbf{c} \mid (X_1, \hat{L}_1), \dots, (X_{i-1}, \hat{L}_{i-1}), R) \\ &= \sum_{i=1}^m \left(I(X_i; \mathbf{c} \mid (X_1, \hat{L}_1), \dots, (X_{i-1}, \hat{L}_{i-1}), R) \right. \\ &\quad \left. + I(\hat{L}_i; \mathbf{c} \mid X_i, (X_1, \hat{L}_1), \dots, (X_{i-1}, \hat{L}_{i-1}), R) \right). \end{aligned}$$

The first term in the sum is zero by (17), we remove it and expand the last mutual information expression using Fact F2.

$$\begin{aligned} I(X^m, \hat{L}^m; \mathbf{c} \mid R) &= \sum_{i=1}^m \left(H(\hat{L}_i \mid X_i, (X_1, \hat{L}_1), \dots, (X_{i-1}, \hat{L}_{i-1}), R) \right. \\ &\quad \left. - H(\hat{L}_i \mid \mathbf{c}, X_i, (X_1, \hat{L}_1), \dots, (X_{i-1}, \hat{L}_{i-1}), R) \right) \\ &\leq m(1 - \mathcal{H}(\eta)). \end{aligned}$$

The last step uses the facts that $H(\hat{L}_i \mid \dots) \leq 1$ and $H(\hat{L}_i \mid \mathbf{c}, X_i, \dots) = \mathcal{H}(\eta)$ in the classification noise model. This concludes the proof. \blacksquare

THEOREM 5.6. *Let \mathcal{C} and \mathcal{P} be an ϵ -binary pair on $(\mathcal{X}, \mathcal{B})$. If A is a $(\mathcal{P}, \mathcal{M})$ -learning algorithm for \mathcal{C} , that uses membership queries in the classification noise model with noise rate $\eta < 1/2$, $\epsilon \leq 1/2$ and $\delta < 1/2$ then*

$$m \geq \frac{(1 - 2\delta)\eta \ln \frac{1-\delta}{\delta}}{(1 - 2\eta)^2} = \Omega\left(\frac{\eta}{(1 - 2\eta)^2} \ln \frac{1}{\delta}\right). \quad (18)$$

Proof. We can apply Lemma 4.2 after setting $\mathcal{D} = (d, 1 - d)$ and recalling that $\mathcal{C}_\epsilon = \mathcal{C}$. Along with the proof of Theorem 5.5 above this gives

$$\begin{aligned} &\mathcal{H}((1 - \delta)d + \delta(1 - d)) - \mathcal{H}(\delta) \\ &\leq \sum_{i=1}^m H(\hat{L}_i \mid X_i, (X_1, \hat{L}_1), \dots, (X_{i-1}, \hat{L}_{i-1}), R) - m\mathcal{H}(\eta). \quad (19) \end{aligned}$$

We now upper bound each $H(\hat{L}_i | X_i, (X_1, \hat{L}_1), \dots, (X_{i-1}, \hat{L}_{i-1}), R)$ by $H(\hat{L}_i | R)$ and notice that the weakest bound results when each $H(\hat{L}_i | R)$ is as large as possible.

By averaging over the target and noise (recall that X_i is the random variable for the i th instance in the sample),

$$\begin{aligned}
\Pr_{\mathcal{M}}(\hat{L}_i = 1|r) &= (1-\eta)d\Pr_{\mathcal{M}}(X_i = x_1 | c_1, r) \\
&\quad + (1-\eta)d(1-\Pr_{\mathcal{M}}(X_i = x_1 | c_1, r)) \\
&\quad + (1-\eta)(1-d)\Pr_{\mathcal{M}}(X_i = x_1 | c_2, r) \\
&\quad + \eta(1-d)(1-\Pr_{\mathcal{M}}(X_i = x_1 | c_2, r)) \\
&= (1-\eta)d + (1-d)((1-\eta)p_i + \eta(1-p_i)) \\
&= (1-\eta)d + \eta(1-d) + p_i(1-d)(1-2\eta),
\end{aligned}$$

where $p_i = \Pr_{\mathcal{M}}(X_i = x_1 | c_2, r)$ represents the probabilistic⁶ query strategy of the algorithm in this context. For any i and for any r ,

$$\begin{aligned}
H(\hat{L}_i | R = r) &= \mathcal{H}(\Pr_{\mathcal{M}}(\hat{L}_i = 1|r)) \\
&= \mathcal{H}((1-\eta)d + \eta(1-d) + p_i(1-d)(1-2\eta)) .
\end{aligned}$$

If $d \geq 1/2$ then both $(1-\eta)d + \eta(1-d) \geq 1/2$ and $(1-d)(1-2\eta) \geq 0$, since $\eta \leq 1/2$. Therefore the argument to the entropy is at least $1/2$, the entropy is maximized when $p_i = 0$, and

$$\sum_{i=1}^m H(\hat{L}_i | R) \leq m\mathcal{H}((1-\eta)d + \eta(1-d)).$$

Plugging this bound into (19) and solving for m gives

$$m(d) \geq \frac{\mathcal{H}((1-\delta)d + \delta(1-d)) - \mathcal{H}(\delta)}{\mathcal{H}((1-\eta)d + \eta(1-d)) - \mathcal{H}(\eta)}, \quad (20)$$

so

$$m \geq \lim_{d \rightarrow 1^-} m(d) \geq \frac{(1-2\delta)\ln\frac{1-\delta}{\delta}}{(1-2\eta)\ln\frac{1-\eta}{\eta}} \geq \frac{(1-2\delta)\eta \ln\frac{1-\delta}{\delta}}{(1-2\eta)^2}.$$

This concludes the proof. \blacksquare

Note that one can easily find a concept class \mathcal{C} and a distribution \mathcal{P} where $N(\mathcal{C}, \epsilon, \mathcal{P}) = \Theta(|\mathcal{C}_\epsilon|)$ and where a learning algorithm can use membership queries to perform a binary search in a smallest ϵ -cover of \mathcal{C} . One example is the class of initial segments of $[0,1]$ with the uniform distribution, as mentioned in Eisenberg

⁶Actually, since the random bits of A are given, the choices of A are deterministic. For the present argument, however, we can formally allow $\Pr_{\mathcal{M}}(X_i = x_1 | c_2, r)$ to be probabilities instead of being only 0 or 1.

and Rivest [18]. This shows that, at least in the $\eta = 0$ case, Theorem 5.5 is in some sense the best possible general lower bound.

The bound $(1 - \delta)\log(N - 1) - 1 \leq m(1 - \mathcal{H}(\eta))$ in Theorem 5.5 holds for any query model \mathcal{M} . By Lemma 2.1, this implies that if \mathcal{C} is not finitely coverable w.r.t. \mathcal{P} then N is unbounded and \mathcal{C} is not $(\mathcal{P}, \mathcal{M})$ -learnable, regardless of the query model \mathcal{M} .

Theorem 5.5 generalizes a lower bound by Turàn [34, Theorem 1] in two directions.

- It measures the descriptive complexity of \mathcal{C} w.r.t. \mathcal{P} using the size of a largest ϵ -well-separated subclass of \mathcal{C} . Thus if \mathcal{P} is the distribution over d_{VC} shattered points mentioned in Ehrenfeucht et al. [17], (so $N = \Omega(d_{VC})$, see Apolloni and Gentile [6]), then we immediately obtain the bound $m \geq \Omega(d_{VC}/(1 - 2\eta)^2)$ which holds for arbitrary \mathcal{C} with $d_{VC}(\mathcal{C}) = d_{VC}$. On the other hand, the generalization is proper, as applying Theorem 5.5 to the concept class and the distribution mentioned in Theorem 5.2 yields the tighter bound $m \geq \Omega\left(\frac{d_{VC}}{(1 - 2\eta)^2} \log \frac{1}{\epsilon}\right)$.

- Turàn’s result is specifically for the noise-free case, and our bound includes a dependence on the noise rate.

Although the proof of Theorem 5.5 holds for both membership queries and random examples, the information involved essentially comes only from the membership queries. This is reasonable since the learner knows \mathcal{P} ahead of time, so does not need random examples to learn the testing distribution.

Theorem 5.5 can be extended to the case where the learner can make only a bounded number of membership queries and any additional information it needs must be provided by random examples. This extension easily follows from the additivity of information and we omit the details.

On the other hand, the lower bound of Theorem 5.6 is due solely to the difficulty of learning with noise, as it takes only a single noise-free query to learn an ϵ -binary pair. Furthermore, Theorem 5.6 can be used to obtain a clean lower bound on the tail of the binomial distribution.

Since Theorem 5.6 applies to any learning algorithm that uses membership queries in the classification noise model, it also applies to the following (Bayes) algorithm A : We recall that for an ϵ -binary pair \mathcal{C} and \mathcal{P} we have $\mathcal{C} = \{c_1, c_2\}$ with $c_1 = \{x_1, x_2\}$, $c_2 = \{x_1\}$ and $\mathcal{P}(x_1) = 1 - \epsilon$, $\mathcal{P}(x_2) = \epsilon$. Algorithm A uses membership queries only, requesting the label of x_2 some odd m number of times. If most of the labels are 0 then A returns the hypothesis c_2 . Otherwise, A returns hypothesis c_1 . The probability of error is simply the probability of returning a hypothesis that misclassifies x_2 . It is easy to see that this probability is exactly

$$\delta = \sum_{i=0}^{(m-1)/2} \binom{m}{i} (1 - \eta)^i \eta^{m-i} . \quad (21)$$

We can interpret Theorem 5.6 as a bound on this δ . In the appendix we sketch the proof of the following fact.

Fact F6. If $\delta \in [0, 1/2]$ then

$$(1 - 2\delta) \ln \left(\frac{1 - \delta}{\delta} \right) \geq \max \left\{ \ln \frac{5}{12\delta}, 8 \left(\frac{1}{2} - \delta \right)^2 \right\}.$$

We use these substitutions into (18) and solve for δ . This results in the following bound on the probability of error (21).

COROLLARY 5.1. *If $\eta \in [0, 1/2]$ and $m \geq 1$, m odd, then the following holds.*

$$\sum_{i=0}^{(m-1)/2} \binom{m}{i} (1 - \eta)^i \eta^{m-i} \geq \max \left\{ \frac{5}{12} e^{-\frac{m(1-2\eta)^2}{\eta}}, \frac{1}{2} - (1 - 2\eta) \sqrt{\frac{m}{8\eta}} \right\}. \quad (22)$$

For the sake of comparison, we point out that a bound on the RHS of (21) is also obtained by Barve and Long [10] by dealing directly with the binomial terms. Their bound can be written as

$$\sum_{i=0}^{(m-1)/2} \binom{m}{i} (1 - \eta)^i \eta^{m-i} \geq \frac{1 - 2e^{-2}}{2} (4\eta(1 - \eta))^m (2(1 - \eta))^{-2\sqrt{m}-3}. \quad (23)$$

One can easily show that for any m there exists an $\eta_0 \in (0, 1/2)$ such that for any $\eta \geq \eta_0$ and $\eta < 1/2$ the RHS of (22) is greater than the RHS of (23). This is evident from the fact that

$$\lim_{\eta \rightarrow 1/2^-} \text{RHS of (23)} = \frac{1 - 2e^{-2}}{2} < \frac{1}{2} = \lim_{\eta \rightarrow 1/2^-} \text{RHS of (22)}.$$

Observe that (18) derives from (20) through a limit that aims at emphasizing the behavior of m when both δ approaches 0 *and* η approaches 1/2. Therefore it is not surprising that (22) yields sharp bounds only in this specific, though interesting, case.

We finally compare our lower bound to the well-known Chernoff's bounds that bound the tail of the binomial distribution *from above* (see Angluin and Valiant [5]). In the case under consideration the Chernoff bound is

$$\sum_{i=0}^{(m-1)/2} \binom{m}{i} (1 - \eta)^i \eta^{m-i} \leq \exp \left(-\frac{(m(1 - 2\eta) + 1)^2}{8m(1 - \eta)} \right). \quad (24)$$

If we expand the square $(m(1 - 2\eta) + 1)^2$ in the RHS we see that

$$\exp \left(-\frac{(m(1 - 2\eta) + 1)^2}{8m(1 - \eta)} \right) = \exp \left(-\frac{m(1 - 2\eta)^2}{8(1 - \eta)} - \frac{(1 - 2\eta)}{4(1 - \eta)} - \frac{1}{8m(1 - \eta)} \right).$$

This allows one to see the close similarity between (24) and the first half of (22) when η is close to 1/2.

6. CONCLUSIONS AND OPEN PROBLEMS

We have presented a simple method for obtaining sample size lower bounds in various PAC-style learning models. This method provides analytical tools that

avoid a Bayesian interpretation of the learning process. In fact, similar results can be proved for other noise models, such as the attribute noise of Shackelford and Volper [33]. Our techniques are strong enough to produce theorems implying bounds on the tail of certain binomial distributions that we have not seen in the literature.

There are several directions in which this work can be extended. Theorem 5.2 adds a $\log 1/\epsilon$ factor to the sample size bounds for certain concept classes. We would like to see a simple characterization of the concept classes for which this $\log 1/\epsilon$ factor can be obtained. We would like to generalize Theorem 5.2 to other noise models, such as the malicious noise model. Finally, it might be possible to apply our technique to prove better bounds for *specific classes* of learning functions (such as those that minimize disagreements).

APPENDIX

A.1. PROOF OF FACT F5 USED IN THEOREM 5.1

We exploit a property of the entropy function $\mathcal{H}[p_1, \dots, p_n]$ that for $n = 3$ reduces to the following [25, p.10]: let p_1, p_2, p_3 be nonnegative numbers with $p_3 > 0$ and $p_1 + p_2 + p_3 = 1$. Then

$$\mathcal{H}[p_1, p_2, p_3] = \mathcal{H}(p_1) + (p_2 + p_3)\mathcal{H}\left(\frac{p_2}{p_2 + p_3}\right).$$

For $\mathcal{H}[(1 - \eta)(1 - \epsilon), d(1 - \eta)\epsilon + (1 - d)\eta, d\eta + (1 - d)(1 - \eta)\epsilon]$ we set

$$\begin{aligned} p_1 &= (1 - \eta)(1 - \epsilon) \\ p_2 &= d\eta + (1 - d)(1 - \eta)\epsilon \\ p_3 &= d(1 - \eta)\epsilon + (1 - d)\eta \end{aligned}$$

and for $\mathcal{H}[(1 - \eta)(1 - \epsilon), (1 - \eta)\epsilon, \eta]$ we set

$$\begin{aligned} p_1 &= (1 - \eta)(1 - \epsilon) \\ p_2 &= \eta \\ p_3 &= (1 - \eta)\epsilon \end{aligned}$$

Observing that in both cases $p_2 + p_3 = \eta + (1 - \eta)\epsilon$ we easily get the thesis.

A.2. PROOF OF LEMMA 5.1, PART 2

If $\alpha, \eta > 0$ and $\eta + \alpha \leq 1/2$ then, since $\eta + \alpha(1 - 2\eta) < \eta + \alpha$, we have $\mathcal{H}(\eta + \alpha(1 - 2\eta)) < \mathcal{H}(\eta + \alpha)$. Now, for any fixed η , $\mathcal{H}(\eta + \alpha)$ is obviously concave in α and therefore by a first-order Taylor expansion around $\alpha = 0$ we get

$$\mathcal{H}(\eta + \alpha) - \mathcal{H}(\eta) \leq \alpha \log\left(\frac{1 - \eta}{\eta}\right)$$

which is $\leq \alpha \log(1/\eta)$, namely the thesis.

A.3. PROOF SKETCH OF FACT F6 USED IN SECTION 5.4

The bound

$$(1 - 2\delta) \ln \left(\frac{1 - \delta}{\delta} \right) \geq 8 \left(\frac{1}{2} - \delta \right)^2$$

follows immediately by a Taylor expansion of $(1 - 2\delta) \ln \left(\frac{1 - \delta}{\delta} \right)$ around $\delta = 1/2$.

To complete the proof, we argue that the function

$$f(\delta) = (1 - 2\delta) \ln \left(\frac{1 - \delta}{\delta} \right) - \ln \frac{5}{12\delta}$$

is positive. An examination of the second derivative shows that f is convex in $[0, 1/2]$, so f is lower bounded by its tangents. Since $f(0.323) > 0.007$ and $0 > f'(0.323) > -0.003$, the tangent to f at $\delta = 0.323$ remains positive when $\delta \in [0, 1/2]$. This concludes the proof.

REFERENCES

1. Angluin, D. (1987), Learning regular sets from queries and counterexamples, *Information and Computation*, **75**, 2, 87-106.
2. Angluin, D. (1988), Queries and concept learning, *Machine Learning*, **2**, 4, 319-342.
3. Angluin D., Krikis, M., Sloan, R., and Turán, G. (1997), Malicious Omissions and Errors in Answers to Membership Queries, *Machine Learning*, **28**, 2/3, 211-255.
4. Angluin, D., and Laird, P. D. (1988), Learning from Noisy Examples, *Machine Learning*, **2**, 2, 343-370.
5. Angluin, D., and Valiant, L. (1979), Fast probabilistic algorithms for Hamiltonian circuits and matchings, *Journal of Computer and System Sciences*, **18**, 155-193.
6. Apolloni, B., and Gentile, C. (1998), Sample Size Lower Bounds in PAC Learning by Algorithmic Complexity Theory, *Theoretical Computer Science*, **209**, 141-162.
7. Aslam, J. A., and Decatur S. E. (1996), On the sample complexity of noise-tolerant learning, *Information Processing Letters*, **57**, 189-195.
8. Bartlett, P. (1992), Learning with a slowly changing distribution, in "Proceedings, 5th Workshop on Computational Learning Theory", pp. 243-252.
9. Bartlett, P., and Helmbold, D. (1996), manuscript.
10. Barve, R. D., and Long, P. M. (1997), On the complexity of learning from drifting distributions, *Information and Computation*, **138**, 2, 170-193.
11. Benedek, G., and Itai A. (1991), Learnability by Fixed Distributions, *Theoretical Computer Science*, **86**, 2, 377-389, 1991.
12. Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. (1989), Learnability and the Vapnik-Chervonenkis Dimension, *Journal of the ACM*, **36**, 929-965.
13. Cesa-Bianchi, N., Dichterman, E., Fischer, P., Shamir, E., and Simon, H. U. (1997), Sample-efficient Strategies for Learning in the Presence of Noise, *eCOLT Technical Report. 97-003*, WWW: <http://ecolt.informatik.uni-dortmund.de/>. Preliminary versions in "Proceedings, 28th Symposium on Theory of Computing", pp. 141-150 and in "Proceedings, 3rd European Conference on Computational Learning Theory", pp. 119-133.
14. Cover, T. M., and Thomas, J. A. (1991), *Elements of information theory*, John Wiley & Sons, Inc., New York.
15. Dobrushin, R. L. (1959), General formulation of Shannon's main theorem in information theory, *Uspekhi Mat. Nauk*, **14**, 3-104; translated in *American Mathematical Society Translations, Ser. 2*, **33**, 323-438, 1963.
16. Dudley, R. M. (1984), *A Course on Empirical Processes*, Lecture Notes in Mathematics, **1097**, Springer-Verlag, Berlin/New York.

17. Ehrenfeucht, A., Haussler D., Kearns, M., and Valiant L. (1989), A General Lower Bound on the Number of Examples Needed for Learning, *Information and Computation*, **82**, 3, 247-261.
18. Eisenberg, B., and Rivest, R. L. (1990), On the Sample Complexity of Pac-Learning using Random and Chosen Examples, in "Proceedings, 3th Workshop on Computational Learning Theory", pp. 154-162.
19. Gentile, C. (1997), A note on sample size lower bounds for PAC-learning, manuscript.
20. Haussler, D., and Barron, A. (1992), How well do Bayes methods work for on-line prediction of $\{-1,+1\}$ values?, in "Proceedings, 3rd NEC Symposium on Computation and Cognition", pp. 74-100.
21. Haussler, D., Kearns, M., and Schapire, R. (1994), Bounds on the Sample Complexity of Bayesian Learning Using Information Theory and the VC Dimension, *Machine Learning*, **14**, 84-114.
22. Haussler, D., Littlestone, N., and Warmuth, M. (1994), Predicting $\{0,1\}$ -functions on randomly drawn points, *Information and Computation*, **115**, 2, 248-293.
23. Haussler, D., and Opper, M. (1997), Mutual Information, Metric Entropy, and Cumulative Relative Entropy Risk, *Annals of Statistics*, **25**, 6, 2451-2492.
24. Helmbold, D., Littlestone, N., and Long, P. (2000), Apple Tasting, *Information and Computation* (to appear).
25. Ihara S. (1993), *Information theory for continuous systems*, World Scientific, River Edge, NJ.
26. Kearns, M., and Li, M. (1993), Learning in the presence of malicious errors, *SIAM Journal of Computing*, **22**, 807-837.
27. Kolmogorov, A. N., and Tihomirov, V. M. (1961), ϵ -entropy and ϵ -capacity of sets in functional spaces, *American Mathematical Society Translations (Ser. 2)*, **17**, 277-364.
28. Laird, P. (1988), *Learning from Good and Bad Data*, Kluwer International Series in Engineering and Computer Science, Kluwer Academic Publishers, Boston, MA.
29. Maass, W. and Turán, G. (1990), On the complexity of learning from counterexamples and membership queries, in "Proceedings, 31th Symposium on the Foundations of Computer Science", pp. 203-210.
30. Sakakibara, K. (1991), On learning from queries and counterexamples in the presence of noise, *Information Processing Letters*, **37**, 5, 279-284.
31. Simon, H. U. (1996), General Bounds on the Number of Examples Needed for Learning Probabilistic Concepts, *Journal of Computer and System Sciences*, **52**, 2, 239-254.
32. Sloan, R. (1995), Four types of noise in data for PAC learning, *Information Processing Letters*, **54**, 157-162.
33. Shackelford, G., and Volper, D. (1988), Learning k-DNF with noise in the attributes, in "Proceedings, 1988 Workshop on Computational Learning Theory", pp. 97-103.
34. Turán, G. (1993), Lower bounds for PAC learning with queries, in "Proceedings. 6th Workshop on Computational Learning Theory" pp. 384-391.
35. Valiant, L. (1984), A theory of the learnable, *Communication of ACM*, **27**, 11, 1134-1142.
36. Vapnik, V. N. (1982), "Estimation of dependencies based on empirical data". Springer Verlag, New York.
37. Vapnik, V. N. (1995), "The Nature of Statistical Learning Theory", Springer Verlag, New York.
38. Yu, B. (1996), Lower Bounds on Expected Redundancy for Nonparametric Classes, *IEEE Trans. on Information Theory*, **42**, 1, 272-275.