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

**Claudio Gentile**
DSI, Universita' di Milano,
Via Comelico 39,
20135 Milano, Italy
gentile@dsi.unimi.it

**David P. Helmbold***
Computer Science Department,
University of California,
95064 Santa Cruz, USA
dph@cse.ucsc.edu

## Abstract

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.

## 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.

The models we consider are variants of the PAC model [34], 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 [25], classification noise [4], drifting distributions [7], 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 $\epsilon$-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 [23].

For the malicious noise model of Kearns and Li [25] we use the simple 2-concept class to improve the lower bounds of Cesa-Bianchi et al. [12] by a $\log 1/\delta$ factor (Theorem 4) 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

*Supported by NSF grant CCR 9700201.


COLT 98 Madison WI USA

improves by a $\log 1/\epsilon$ factor the previous bound proven by Simon [30] and Apolloni and Gentile [5] 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 $\epsilon \to 0$ in the noise-free case (where it is $\Omega(d_{VC}/\epsilon)$ [16]) and when the noise rate is a positive constant (where it grows as $\Omega(\frac{d_{VC}}{\epsilon}\log\frac{1}{\epsilon})$, Theorem 6).

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 [33] in two ways. First, it adds a factor indicating the dependence on the noise rate. Second, our bound depends on the size of $\epsilon$-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 8, part 1).

The final model we apply our techniques to is a batch version of Bartlett's drifting distribution model [7] with classification noise. Here we bound the allowable rate of drift in the noisy case, generalizing results of Bartlett [7], Aslam and Decatur [6], Simon [30] and Apolloni and Gentile [5]. 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.

The next section contains the major 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.

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[1] $H(X) = E_{\mathcal{P}_X}[-\log \mathcal{P}_X(X)]$ 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 $H(X \mid Y = y)$ 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, $H(X \mid Y) = E_{\mathcal{P}_Y}[H(X \mid Y = y)]$. We define the mutual information $I(X;Y)$ between $X$ and $Y$ by

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

where the supremum is taken over all finite partitions $\{B_1, ..., B_n\}$ 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 [14].

If $Z \sim \mathcal{P}_Z$ we denote by $I(X;Y \mid Z = z)$ the *conditional mutual information of $X$ and $Y$ given $Z = z$, 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, $I(X;Y \mid Z) = E_{\mathcal{P}_Z}[I(X;Y \mid Z = z)]$.

For a discrete random variable with density $(p_1, ..., p_m)$ we sometimes denote its entropy by $\mathcal{H}[p_1, ..., p_m]$. 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 $\mathcal{X}$ is a fixed set which we assume to be either finite, countable, or $\mathbb{R}^n$ 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 $X^m = (X_1, ..., X_m)$ and $L^m = (L_1, ..., L_m)$ 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 $x^m = (x_1, ..., x_m)$ and $l^m = (l_1, ..., l_m)$, 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". 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 1 below. 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 $x^m = (x_1, ..., x_m) \in \mathcal{X}^m$, and (in the absence of noise, see below) $L_c^m(x^m) = (I_c(x_1), ..., I_c(x_m)) \in \{0,1\}^m$ where $I_c$ is the indicator function for concept $c$.

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 $\hat{x}^m = (\hat{x}_1, ..., \hat{x}_m) \in \mathcal{X}^m$ and $\hat{l}^m = (\hat{l}_1, ..., \hat{l}_m) \in \{0,1\}^m$.

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 $\mathcal{M}(.,.,.,c) = \mathcal{D}(c)\mathcal{M}(.,.,. \mid c)$, where $\mathcal{M}(.,.,.|c)$ is the conditional distribution of the first three arguments, given that the target concept is $c$. Here $\mathcal{D}$ is the distribution over the concept class mentioned above, while $\mathcal{M}(.,.,.|c)$ is the conditional distribution induced on the sample and the random bits by 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

[1] Here and throughout "log" means "$\log_2$" while "ln" is the natural logarithm. As usual, $0\log 0 = 0\log\infty = 0$.

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 $\mathcal{M}(x^m, \hat{l}^m, r \mid c)$ factors as $\mathcal{M}(x^m, \hat{l}^m, r \mid c) = \mathcal{P}^m(x^m)\, \mathcal{M}(\hat{l}^m \mid x^m, c)\, \mathcal{M}(r)$, where $\mathcal{P}^m$ denotes the $m$-fold $\mathcal{P}$-probability product and $\mathcal{M}(\hat{l}^m \mid x^m, c)$ describes the i.i.d. noisy labeling process. In particular $\mathcal{M}(\hat{l}^m \mid x^m, c) = \prod_{i=1}^m \mathcal{M}(\hat{l}_i \mid x_i, c)$, where $\mathcal{M}(I_c(x_i) \mid x_i, c) = 1 - \eta$ (no noise) and the label is flipped with probability $\eta$.

We will also exhibit an application to learning with slowly *drifting* distributions (Bartlett [7]). 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.

Let $c, h \in \mathcal{B}$. When $\mathcal{P}$ is understood from the context we say that $c$ is *$\epsilon$-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 *$\epsilon$-far* from $h$ otherwise.

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

**Definition 1** *Concept class $\mathcal{C}$ on $(\mathcal{X}, \mathcal{B})$ is* PAC-learnable *w.r.t. distributions $\mathcal{P}$ and $\mathcal{M}(\hat{x}^m, \hat{l}^m, r \mid c)$ ($(\mathcal{P}, \mathcal{M})$-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[2] *$A : \mathcal{X}^m \times \{0,1\}^m \times \{0,1\}^b \to \mathcal{B}$ such that for every $\epsilon, \delta > 0$ and $c \in \mathcal{C}$,*

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

*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 $\mathcal{M}(\hat{x}^m, \hat{l}^m, r \mid c)$. However, the learning constraint is for all possible targets $c \in \mathcal{C}$. Note that an equivalent definition results when the quantification "for all $c \in C$" is replaced by "for all $\mathcal{D}$ over $\mathcal{C}$" and the "$\Pr_{\mathcal{M}(\hat{x}^m, \hat{l}^m, r \mid c)}$" is replaced by "$\Pr_{\mathcal{M}(\hat{x}^m, \hat{l}^m, r, c)}$" (so 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 $\epsilon$-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 $\epsilon$-close to $c$ [10], [26], [35, pp.149-151]. We denote by $N(\mathcal{C}, \epsilon, \mathcal{P})$ the cardinality of a smallest $\epsilon$-cover of $\mathcal{C}$ w.r.t. $\mathcal{P}$.

[2]The 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.

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** *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* $\epsilon$-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 $\epsilon$-close to $h$. We drop "w.r.t. $\mathcal{P}$" when $\mathcal{P}$ is clear from the context.*

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

**Lemma 1** *[10], [26] 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 $2\epsilon$-far.* □

**Remark 1** *Many papers in the lower bound literature, including [10, 18], use mutual separation (as in the consequence of Lemma 1) rather than $\epsilon$-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 $\epsilon$-far is usually called the* $\epsilon$-packing number of $\mathcal{C}$ w.r.t. $\mathcal{P}$ *[26]. Denote this quantity by $M(\mathcal{C}, \epsilon, \mathcal{P})$. We point out that if $\mathcal{C}_\epsilon$ is $\epsilon$-well-separated then its members are mutually $\epsilon$-far from each other, so $|\mathcal{C}_\epsilon| \leq M(\mathcal{C}, \epsilon, \mathcal{P})$. As indicated above, the triangle inequality implies that if $\mathcal{C}_\epsilon$ is a maximum $\epsilon$-well-separated set then $M(\mathcal{C}, 2\epsilon, \mathcal{P}) \leq |\mathcal{C}_\epsilon|$. Therefore $|\mathcal{C}_\epsilon|$ is sandwiched by two $\epsilon$-packing numbers.*

A simple but relevant example of $\epsilon$-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 $\epsilon$-well-separated subclass (however, the concepts in $\mathcal{C}$ are not $2\epsilon$-far). We will refer to this pair of $\mathcal{C}$ and $\mathcal{P}$ as an *$\epsilon$-binary pair* on $(\mathcal{X}, \mathcal{B})$.

Finally, we define the VC-dimension of a concept class $\mathcal{C}$, denoted $d_{VC}(\mathcal{C})$, as the cardinality of a largest subset of the domain *shattered* by $\mathcal{C}$ (see [11], [36, p. 53]).

## 3 The symmetry of mutual information and the method of induced distributions

The following 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 $I(C\,;\, \hat{X}^m, \hat{L}^m, R)$ between the $\mathcal{C}$-valued random variable $C$ and the joint variable $(\hat{X}^m, \hat{L}^m, R)$ formed by the (corrupted) sample $(\hat{X}^m, \hat{L}^m)$ for $C$ and the random bits $R$. Consider the distribution $\mathcal{M}$ introduced in the last section. We recall that $(\hat{X}^m, \hat{L}^m, R) \sim \mathcal{M}(\hat{x}^m, \hat{l}^m, r)$, the marginal of

$\mathcal{M}$ w.r.t. C. Let $\mathrm{H} = A(\hat{X}^m, \hat{L}^m, R)$ be the value of the learning function $A$ for $\mathcal{C}$ on arguments $\hat{X}^m, \hat{L}^m, R$. Since H is a function of $\hat{X}^m$, $\hat{L}^m$ and $R$, we have $I(\mathrm{C}\ ;\ \mathrm{H}) \leq I(\mathrm{C}\ ;\ \hat{X}^m, \hat{L}^m, R)$ (see, e.g., [24]). If C is *independent* of the internal randomization $R$ of $A$ then $I(R\ ;\ \mathrm{C}) = 0$. By the symmetry and the additivity of $I$ we get $I(\mathrm{C}\ ;\ \hat{X}^m, \hat{L}^m, R) = I(\hat{X}^m, \hat{L}^m, R\ ;\ \mathrm{C}) = I(R\ ;\ \mathrm{C}) + I(\hat{X}^m, \hat{L}^m\ ;\ \mathrm{C} \mid R)$. Hence we obtain the general inequality

$$I(\mathrm{C}\ ;\ \mathrm{H}) \leq I(\hat{X}^m, \hat{L}^m\ ;\ \mathrm{C} \mid R), \tag{2}$$

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[3] used by $A$.

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 $\epsilon$ error with confidence $\delta$. 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 $I(\hat{X}^m, \hat{L}^m\ ;\ \mathrm{C} \mid R)$ 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 [25] (but see also Angluin and Laird [4], Sloan [31], Cesa-Bianchi et al. [12]) can thus be reinterpreted by means of this mutual information argument. The adversary tries to make $I(\hat{X}^m, \hat{L}^m\ ;\ \mathrm{C} \mid R)$ as small as possible. If 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. On the other hand, working from inequality (2) has the advantage of being able to *quantify* the hardness of the learning task as $I(\hat{X}^m, \hat{L}^m\ ;\ \mathrm{C} \mid R) \to 0$.

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 $I(\hat{X}^m, \hat{L}^m\ ;\ \mathrm{C} \mid R)$ factors as $mI(\hat{X}_i, \hat{L}_i\ ;\ \mathrm{C} \mid R)$. Solving (2) for $m$ results in a lower bound of the form

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

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 the above.

The role played by the learning function's randomization $R$ in this argument in 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 the entropy formulas.

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 $I(\mathrm{C}\ ;\ \hat{X}^m, \hat{L}^m)$ 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 $\mathcal{M}(x^m, l^m \mid c) = \prod_{i=1}^m \mathcal{M}(x_i, l_i \mid c)$.

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 [19], Haussler, Kearns and Schapire [20], Haussler and Opper [22], Yu [37] 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

[3] From now on we assume C and $R$ are independent.

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 [5]. 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 $\mathcal{M}(.,.,.|c)$ 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 than to compare the two alternative expressions for $I$ is stressed in [18]. 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 $I(\mathrm{C}\ ;\ \mathrm{H})$ as a function of $\mathcal{D}$

In this section we will fully exploit the existence of large $\epsilon$-well-separated subclasses (as defined in Section 2) to bound the information $I(\mathrm{C}\ ;\ \mathrm{H})$ required for learning. Here $\epsilon$ 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 $\epsilon$-well-separated subclass $\mathcal{C}_\epsilon = \{c_1, ..., c_N\}$ of $\mathcal{C}$ and C be a random variable with distribution $\mathcal{D} = (d_1, ..., d_N)$ over $\mathcal{C}_\epsilon$. Now, set for brevity $\mathrm{H} = A(\hat{X}^m, \hat{L}^m, R)$. Since $\mathcal{C}_\epsilon$ is $\epsilon$-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 $\epsilon$-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 $\epsilon$-close to $c_k$. Let

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

Define

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

and, for $k = 2, ..., N$,

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

The family of sets $\{\mathrm{cl}(k),\ k = 1, ..., N\}$ 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

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

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

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

$$\widehat{I}(\mathrm{C}\ ;\ \mathrm{H}) = \sum_{i,k=1}^{N} \mathrm{Pr}_{\mathcal{M}}(B_{ik}) \log \frac{\mathrm{Pr}_{\mathcal{M}}(B_{ik})}{\mathrm{Pr}_{\mathcal{M}}(\mathrm{H} \in \mathrm{cl}(k))\ \mathrm{Pr}_{\mathcal{M}}(c_i)}, \tag{4}$$

is a lower bound on $I(\mathrm{C}\ ;\ \mathrm{H})$. We lower bound $I(\mathrm{C}\ ;\ \mathrm{H})$ by lower bounding $\widehat{I}(\mathrm{C}\ ;\ \mathrm{H})$.

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

**Lemma 2** *If $\mathcal{D}$ is uniform over $\mathcal{C}_\epsilon$, i.e., each $d_i = 1/N$, then*

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

*Proof.* In the discretized scenario we described thus far we apply the classical Fano's inequality (see, e.g., [13, Theorem 2.11.2]) that lower bounds the partitional mutual information $\widehat{I}(\mathrm{C}\ ;\ \mathrm{H})$ in terms of the probability of an "error",

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

Since $\mathcal{D}$ is uniform over $\mathcal{C}_\epsilon$, Fano's inequality yields

$$\begin{aligned}\widehat{I}(\mathrm{C}\ ;\ \mathrm{H}) &\geq \log N - \mathrm{Pr}_{\mathcal{M}}(error)\ \log(N-1) - 1 \\ &\geq (1 - \mathrm{Pr}_{\mathcal{M}}(error))\ \log(N-1) - 1.\end{aligned}$$

The probability of error, $\mathrm{Pr}_{\mathcal{M}}(error)$, is a lower bound on the probability that the learning function produces a hypothesis that is $\epsilon$-far from the target. Therefore, $\mathrm{Pr}_{\mathcal{M}}(error) \leq \delta$ and

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

completing the proof. □

**Lemma 3** *If $N = 2$, $\mathcal{D} = (d, 1-d)$ and $\delta < 1/2$ then*

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

*Proof sketch.* We set for brevity $p_{12} = \mathrm{Pr}_{\mathcal{M}}(\mathrm{H} \in \mathrm{cl}(2) \,|\, c_1)$ and $p_{21} = \mathrm{Pr}_{\mathcal{M}}(\mathrm{H} \in \mathrm{cl}(1) \,|\, c_2)$. We have $\mathrm{Pr}_{\mathcal{M}}(\mathrm{H} \in \mathrm{cl}(1)) = (1-p_{12})d + p_{21}(1-d)$ and the partitional mutual information

$$\begin{aligned}\widehat{I}(\mathrm{C}\ ;\ \mathrm{H}) \quad = \quad & \mathcal{H}\big((1-p_{12})d + p_{21}(1-d)\big) \\ & - d\mathcal{H}(p_{12}) - (1-d)\,\mathcal{H}(p_{21}).\end{aligned} \tag{5}$$

Now, a derivative argument shows that for any fixed $d$ and $p_{21} \in [0,1]$, $\widehat{I}(\mathrm{C}\ ;\ \mathrm{H})$ is non-increasing when $p_{12} \in [0, 1 - p_{21}]$. Similarly, for any fixed $d$ and $p_{12} \in [0,1]$, $\widehat{I}(\mathrm{C}\ ;\ \mathrm{H})$ 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 $\widehat{I}(\mathrm{C}\ ;\ \mathrm{H})$ by substituting $\delta$ for $p_{12}$ and $p_{21}$ in Equation (5). This yields the bound of the lemma. □

**Remark 2** *The preceding lemmas could also have been obtained through the following more general approach for lower bounding $\widehat{I}(\mathrm{C}\ ;\ \mathrm{H})$. The difficulty is that $\widehat{I}(\mathrm{C}\ ;\ \mathrm{H})$ depends not only on $\mathcal{D}$, but also on the probabilities $\mathrm{Pr}_{\mathcal{M}}(\mathrm{H} \in cl(k) \mid c_i)$ which depend on the learning function. In order to get a general bound on $\widehat{I}(\mathrm{C}\ ;\ \mathrm{H})$ we must consider how the learning function can affect these probabilities.*

*Set for brevity $p_{ik} = \mathrm{Pr}_{\mathcal{M}}(\mathrm{H} \in cl(k) \mid c_i)$, so that in (4) $\mathrm{Pr}_{\mathcal{M}}(B_{ik}) = p_{ik}d_i$ and $\mathrm{Pr}_{\mathcal{M}}(\mathrm{H} \in cl(k)) = \sum_{i=1}^{N} p_{ik}d_i$.*

*Since, with probability $1-\delta$, the learning function must produce a hypothesis that is $\epsilon$-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.*

$$\min_{\Omega} \widehat{I}(\mathrm{C}\ ;\ \mathrm{H}) \tag{6}$$

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

*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* $\widehat{I}(\mathrm{C}\ ;\ \mathrm{H})$ *as a function of* $\mathcal{D}$.

*Since* $\widehat{I}(\mathrm{C}\ ;\ \mathrm{H})$ *is* convex *on* $\Omega$ *for any* $\mathcal{D}$ *(see, e.g., [13, p. 31]) and the constraints defining* $\Omega$ *are linear, a standard Kuhn-Tucker analysis (6) 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 [25] 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 $\eta$ 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 adversary that knows $\epsilon, \delta, c, \mathcal{P}, \mathcal{M}$ and the internal state of a device computing $A$ [4]. 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 4 below have been shown by Kearns and Li [25] and by Cesa-Bianchi et al. [12]. 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 [12], and illustrates how the RHS of (3) can guide the choice of $\mathcal{D}$.

**Theorem 4** *Let* $\mathcal{C}$ *and* $\mathcal{P}$ *be an* $\epsilon$*-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* $m$ *must hold*

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

*Proof.* Since $\mathcal{C}$ and $\mathcal{P}$ are an $\epsilon$-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 $\epsilon$-well-separated set. For any $\mathcal{D} = (d, 1-d)$, applying Lemma 3 gives $I(\mathrm{C}\ ;\ \mathrm{H}) \geq \mathcal{H}\big((1-\delta)d + \delta(1-d)\big) - \mathcal{H}(\delta)$.

The malicious adversary behaves as follows [25, 12]: if noise occurs and $c$ is the target, it replaces the current example $(x, l)$ by $(x_2, 1 - l_c(x_2))$. 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 $r$)

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

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

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

and

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

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

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

$$\mathcal{H}\big[(1-\eta)(1-\epsilon), d(1-\eta)\epsilon + (1-d)\eta, d\eta + (1-d)(1-\eta)\epsilon\big] - \mathcal{H}\big[(1-\eta)(1-\epsilon), (1-\eta)\epsilon, \eta\big] =$$

$$(\eta + (1-\eta)\epsilon)\Big[\mathcal{H}\Big(\tfrac{d\eta + (1-d)(1-\eta)\epsilon}{\eta + (1-\eta)\epsilon}\Big) - \mathcal{H}\Big(\tfrac{\eta}{\eta + (1-\eta)\epsilon}\Big)\Big].$$

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

$$\frac{\mathcal{H}\big((1-\delta)d + \delta(1-d)\big) - \mathcal{H}(\delta)}{(\eta + (1-\eta)\epsilon)\Big[\mathcal{H}\Big(\tfrac{d\eta + (1-d)(1-\eta)\epsilon}{\eta + (1-\eta)\epsilon}\Big) - \mathcal{H}\Big(\tfrac{\eta}{\eta + (1-\eta)\epsilon}\Big)\Big]}. \tag{7}$$

Note that this bound is a function of $d$ and $\sup_{d\in[0,1]} m(d)$ is at least $\lim_{d\to 0+} m(d)$. Since both the numerator and the denominator of (7) vanish as $d \to 0$ we use De l'Hôspital's rule to evaluate this limit. Some simple algebra leads to

$$\lim_{d\to 0^+} m(d) = \frac{(1-2\delta)\ln\frac{1-\delta}{\delta}}{(\epsilon - \eta(1+\epsilon))\ln\frac{\epsilon(1-\eta)}{\eta}}. \tag{8}$$

[4] There 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 (Cesa-Bianchi et al. [12]) 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 $A$.

Since $\epsilon - \eta(1+\epsilon) = \Delta(1+\epsilon)$ and $\frac{\epsilon(1-\eta)}{\eta} = 1 + \frac{\Delta(1+\epsilon)}{\eta}$ we can use the inequality $\ln(1+x) \leq x$ to overestimate the denominator of (8) with $\frac{\Delta^2(1+\epsilon)^2}{\eta}$. This completes the proof. □

The previous bound is meaningful only when $\eta$ is close to the information-theoretic limit $\frac{\epsilon}{1+\epsilon}$, but has the advantage of *diverging* as $\delta \to 0$. By selecting $d = 1/2$ (instead of $d = 0$), a completely analogous argument proves the bound $m \geq O(\frac{\eta+\Delta}{\Delta^2})$, 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 $\delta \to 0$ versus a bound which does not vanish for $\eta = 0$ but does not contain a dependence on $\delta$) will recur. Both Theorem 7 part 2 and Theorem 8 part 2 are phrased to emphasize the dependence on $\delta$. 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 $\delta$).

### 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 $\eta$ 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 $\eta$ is bounded away from 0, say $\eta \geq 1/100$, the following theorem adds a logarithmic factor to a bound proved in Simon [30] and Apolloni and Gentile [5], 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** *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 [5].
2. In appendix.
3. Easily derived from the fact $-(1-\alpha)\ln(1-\alpha) \leq \alpha$, for all $\alpha \in [0,1]$. □

**Theorem 6** *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*

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

*where the hidden constant in this $\Omega$-expression depends on $\eta_0$.*

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

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

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

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

*Proof.* Let us consider the subclass $\hat{\mathcal{C}}_k$ of $\mathcal{C}_k$ defined as follows. Set [5] $T = \frac{1}{4\epsilon}$. We split $[0,1]$ into $k$ intervals, $I_1$ through $I_k$, each of length $1/k$. We then split each $I_i$ into $T$ sub-intervals, $I_{i1}$ through $I_{iT}$, of length $1/Tk$. Thus for $i = 1, ..., k$ and $j = 1, ..., T$, 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[6]

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

so that each $c \in \hat{\mathcal{C}}_k$ is the (disjoint) union of $k$ sub-intervals, one from each different interval. We will lower bound the cardinality of a largest $\epsilon$-well-separated subclass of $\hat{\mathcal{C}}_k$ by underestimating $N(\hat{\mathcal{C}}_k, 2\epsilon, \mathcal{P})$ (Lemma 1). Two concepts in $\hat{\mathcal{C}}_k$, are $2\epsilon$-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 $2\epsilon$-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 $2\epsilon$-cover of $\hat{\mathcal{C}}_k$ contains at least $T^k / 2^k T^{\lceil k/2 \rceil} \geq \frac{1}{2}(T/4)^{\lfloor k/2 \rfloor}$ concepts. Now Lemma 1 implies that the cardinality of a largest $\epsilon$-well-separated subclass $\hat{\mathcal{C}}_{k\epsilon}$ of $\hat{\mathcal{C}}_k$ satisfies

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

Let $\mathcal{D}$ be uniform over $\hat{\mathcal{C}}_{k\epsilon}$. Since in this noise model $X^m$, C and $R$ are independent, we have $I(X^m, \hat{L}^m \ ; \ \mathrm{C} \mid R) = I(\hat{L}^m \ ; \ \mathrm{C} \mid X^m) = H(\hat{L}^m \mid X^m) - H(\hat{L}^m \mid X^m, \mathrm{C})$. We overestimate $H(\hat{L}^m \mid X^m)$ by $H(\hat{L}^m)$ and note that $H(\hat{L}^m \mid X^m, \mathrm{C}) = m\mathcal{H}(\eta)$ for the classification noise model. Applying (2) and Lemma 2 now results in

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

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

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

[5] In order not to complicate the notation, in this proof we assume $T$ is an integer.

[6] The class $\hat{\mathcal{C}}_k$ is defined as a function of the accuracy $\epsilon$. Since $\hat{\mathcal{C}}_k$ is a subset of $\mathcal{C}_k$ for every $\epsilon$, we are bounding the difficulty of learning the *larger* class $\mathcal{C}_k$ by considering, for every value of $\epsilon$, a hard *subclass* of $\mathcal{C}_k$.

(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, after noting that the denominator of bound 1 is at least a constant. Part 2 follows from (11), bound 2 of Lemma 5, 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. □

The lower bounds of Theorem 6 actually hold for every pair of $\mathcal{C}$ and $\mathcal{P}$ such that $\mathcal{C}$ contains a subclass $\hat{C}$ of concepts each with $\mathcal{P}$-measure $O(\epsilon)$ and where $\log N(\hat{C}, 2\epsilon, \mathcal{P}) = \Omega(d_{VC}(C)\log(1/\epsilon))$. For instance, by the embedding technique[7] of Helmbold et al. [23], the same lower bound holds for various common geometric concept classes such as axis-parallel rectangles in $\mathbb{R}^n$ and half-spaces.

Moreover, we remark that such a lower bound is the best possible when $\eta$ 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 [27, p. 190]. This analysis of disagreement minimization is valid only for finite size classes. Disregarding the dependence on $\delta$, the sample complexity there is $O(\frac{\log|\mathcal{C}|}{\epsilon(1-2\eta)^2})$. Since we are assuming that the learning function *knows* the distribution $\mathcal{P}$, it knows in principle a smallest $\epsilon$-cover of $\mathcal{C}$ w.r.t. $\mathcal{P}$. Dudley contains a proof [15, Theorem 9.3.1] that smallest $\epsilon$-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 $\epsilon$. Using Laird's analysis of minimizing disagreements on a smallest $\epsilon/2$-cover of $\mathcal{C}$, one could obtain the sample size bound $m = O(\frac{\log K + d_{VC}(\mathcal{C})\log\frac{1}{\epsilon}}{\epsilon(1-2\eta)^2})$, which matches our lower bound as $\epsilon \to 0$.

It is natural to expect that the ideas behind Theorem 6 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 [7] and further explored by Bartlett and Helmbold [8] and Barve and Long [9]. Notice, however, that we are still considering a *batch* learning setting.

Following Bartlett [7], we define a distance between probability distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ on $\mathcal{X}$ as follows:

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

A sequence of probability distributions $\{\mathcal{P}_j\}_{j=1...m}$ is called *$\gamma$-admissible* if $\mathrm{dist}(\mathcal{P}_j, \mathcal{P}_{j+1}) \leq \gamma$, for $j = 1, ..., m-1$. 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 "*$(\mathcal{P}, \prod_j \mathcal{P}_j)$-learning in the classification noise model*". Here we assume that the testing distribution $\mathcal{P} = \mathcal{P}_m$, the last distribution in the sequence.

The following lower bound has two parts. The first is a generalization of results in [7, 30, 5]. The second is a generalization of a result proved by Aslam and Decatur [6]. 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). [8]

**Theorem 7**
*1) Let $\mathcal{C}$ be a concept class on $(\mathcal{X}, \mathcal{B})$, $d_{VC}(\mathcal{C}) = d_{VC} \geq 86$, $\epsilon < 1/16$ and $\delta \leq 1/40$. Then* for every $m$ *there exists a $\gamma$-admissible sequence of distributions $\{\mathcal{P}_j\}_{j=1...m}$ on $\mathcal{X}$, with $\gamma = \frac{640\epsilon^2(1-\mathcal{H}(\eta))}{d_{VC}-2}$, such that $(\mathcal{P}_m, \prod_j \mathcal{P}_j)$-learning in the classification noise model is impossible.*

*2) Let $\mathcal{C}$ and $\mathcal{P}_m$ be an $\epsilon$-binary pair on $(\mathcal{X}, \mathcal{B})$, with $\epsilon \leq 1/2$ and $\delta < 1/2$. Then* for every $m$ *there exists a $\gamma$-admissible sequence of distributions $\{\mathcal{P}_j\}_{j=1...m}$ 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.* 1) Let $\{x_1, ..., x_{d_{VC}}\} \subseteq \mathcal{X}$ be shattered by $\mathcal{C}$. Define $\{\mathcal{P}_j\}_{j=1...m}$ to be the following $\gamma$-admissible sequence of distributions on $\mathcal{X}$ with $t = \frac{d_{VC}-2}{40\epsilon(1-\mathcal{H}(\eta))}$:

$$\mathcal{P}_j(x_i) = \begin{cases} 0 & \text{for } i = 1, ..., d_{VC}-2; \\ 1-16\epsilon & \text{for } i = d_{VC}-1; \\ 16\epsilon & \text{for } i = d_{VC} \end{cases}$$

$j = 1, ..., m-t$,

$$\mathcal{P}_j(x_i) = \begin{cases} \frac{\gamma(j-m+t)}{d_{VC}-2} & \text{for } i = 1, ..., d_{VC}-2; \\ 1-16\epsilon & \text{for } i = d_{VC}-1; \\ \gamma(m-j) & \text{for } i = d_{VC} \end{cases}$$

$j = m-t+1, ..., m$, and $\mathcal{P}_j(x) = 0$ elsewhere, $j = 1, ..., m$.

W.l.o.g., we restrict our attention to $\mathcal{X} = \{x_1, ..., x_{d_{VC}}\}$ and the class $\mathcal{C}' = \{c \in 2^{\{x_1,...,x_{d_{VC}}\}} : I_c(x_{d_{VC}-1}) = I_c(x_{d_{VC}}) = 0\}$. The sequence $\{\mathcal{P}_j\}_{j=1...m}$ is $\gamma$-admissible since the subset of $\mathcal{X}$ that has the largest variation in probability is $\{x_1, ..., x_{d_{VC}-2}\}$. Since $\gamma t = 16\epsilon$ the testing distribution $\mathcal{P}_m$ is actually the following: $\mathcal{P}_m(x_i) = 16\epsilon/(d_{VC}-2)$, $i = 1, ..., d_{VC}-2$, $\mathcal{P}_m(x_{d_{VC}-1}) = 1-16\epsilon$, $\mathcal{P}_m(x) = 0$ elsewhere. Let $N = |\mathcal{C}'_\epsilon|$ and $\mathcal{D}$ be uniform over $\mathcal{C}'_\epsilon$. Again we apply (2) and Lemma 2 and remark that for this model of noise $I(X^m, \hat{L}^m ; \mathrm{C} \mid R) = I(\hat{L}^m ; \mathrm{C} \mid X^m) = H(\hat{L}^m \mid X^m) - H(\hat{L}^m \mid X^m, \mathrm{C}) = H(\hat{L}^m \mid X^m) - m\mathcal{H}(\eta)$. We get the following necessary condition on $m$

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

[7] The proof of Theorem 6 requires that the entropy of the observed labels be small. Therefore, embedding techniques based on initial segments (like those of Haussler et al. [21]) are more difficult to apply in this context.

[8] The bound related to part 1) was proved in [30] by appealing to the central limit theorem and in [5] with worse constants and only for $\eta$ bounded away from 0. Theorem 7, part 1 can also be derived by the Bayesian argument in [9, Theorem 18], but the constants therein are exceedingly large.

From computations which are close to those exhibited in the proof of Theorem 6 (for a very similar computation we refer to the proof of Corollary 1 in [5]) it follows that $\log(N-1) > \frac{11}{25}(d_{VC}-3)$ for $d_{VC} \geq 7$ and $\epsilon < 1/16$.

Set now $X^m = (X_1, ..., X_m)$ and let $B^m = (B_1, ..., B_m)$ be a bernoullian noise vector where the $B_j$'s are i.i.d $\sim$ bernoul($\eta$), $\hat{L}^m = (\hat{L}_1, ..., \hat{L}_m)$, $\hat{L}_j = I_c(X_j) \oplus B_j$, $j = 1, ..., m$ and $\oplus$ is the exclusive-OR. (Note that $B_j = 1$ means that a labeling error has occurred in $j$-th position.) Define the following three random variables: $U$ = number of $j$'s such that $X_j = x_{d_{VC}-1}$, when $j$ ranges in $\{m-t+1, ..., m\}$; $V$ = number of $j$'s such that $X_j = x_{d_{VC}-1}$ *and* $B_j = 1$, when $j$ ranges in $\{m-t+1, ..., m\}$; $W$ = number of $j$'s such that $B_j = 1$, when $j$ ranges in $\{1, ..., m-t\}$. Obviously $U \sim$ binomial$(t, 1-16\epsilon)$, $V \sim$ binomial$(t, (1-16\epsilon)\eta)$ and $W \sim$ binomial$(m-t, \eta)$. Given $X^m$, the cardinality of the range of the variable $\hat{L}^m$ can in fact be upper bounded in terms of $U, V$ and $W$ by

$$\binom{m-t}{W} \cdot 2^{t-U} \cdot 2 \cdot \binom{U}{V} \tag{13}$$

Consider indeed the first $m-t$ components of $X^m$. Since $I_c(x_{d_{VC}-1}) = I_c(x_{d_{VC}}) = 0$, the first $m-t$ components of the vector $\hat{L}^m$ can be set to 1 only by the noise. Hence the number of possible values taken by the first $m-t$ components of $\hat{L}^m$ is exactly $\binom{m-t}{W}$. On the other hand, among the last $t$ components of $\hat{L}^m$, all components $\hat{L}_j$ such that $X_j = x_{d_{VC}-1}$ and $B_j = 0$ have value $v$, while all those such that $X_j = x_{d_{VC}-1}$ and $B_j = 1$ have value $1-v$, where $v$ can be either 0 or 1.

Now, (see, e.g., [13, p. 284]) for $0 \leq b \leq a$ we have $\binom{a}{b} \leq 2^{a\mathcal{H}(b/a)}$ (for $a = 0$ consider both terms to be zero). In view of (13), *no matter how we choose $\mathcal{D}$ over $\mathcal{C}'$* this yields

$$H(\hat{L}^m \mid X^m) \leq (m-t)E_{\mathcal{M}}[\mathcal{H}(W/(m-t)] + t - E_{\mathcal{M}}[U] + 1 + E_{\mathcal{M}}[U\mathcal{H}(V/U)].$$

The function $f(w) = \mathcal{H}(w/(m-t))$ is clearly concave over $[0, m-t]$, while, by direct computation of the Hessian matrix, it can be shown that even the function $g(u,v) = u\mathcal{H}(v/u)$ is concave over the convex set $\{(u,v) : u \in [0,m], v \in [0,u]\}$ (motivated by continuity we set $g(u,u) = g(u,0) = 0$ for $u \in [0,m]$). Therefore by Jensen's inequality $E_{\mathcal{M}}[\mathcal{H}(W/(m-t)] \leq \mathcal{H}(E_{\mathcal{M}}[W]/(m-t)) = \mathcal{H}(\eta)$ and $E_{\mathcal{M}}[U\mathcal{H}(V/U)] \leq E_{\mathcal{M}}[U]\mathcal{H}(E_{\mathcal{M}}[V]/E_{\mathcal{M}}[U]) = t(1-16\epsilon)\mathcal{H}(\eta)$.

Thus, for every $\mathcal{D}$ over $\mathcal{C}'$, $H(\hat{L}^m \mid X^m) - m\mathcal{H}(\eta)$ is bounded from above by

$$(m-t)\mathcal{H}(\eta) + t - t(1-16\epsilon) + 1 + t(1-16\epsilon)\mathcal{H}(\eta) - m\mathcal{H}(\eta)$$

$$= 16t\epsilon(1-\mathcal{H}(\eta)) + 1$$

$$= \tfrac{2}{5}(d_{VC}-2) + 1$$

Putting together as in (12) implies

$$\frac{11}{25}(1-\delta)(d_{VC}-3) - 1 \leq \frac{2}{5}(d_{VC}-2) + 1$$

which, for $\delta \leq 1/40$ and $d_{VC} \geq 86$ is not fulfilled. This contradiction and the fact that it holds for every $m$ prove part 1.

2) This proof is similar to that of Theorem 4. Thus we only give the main steps in a sketchy form.

For $x_1, x_2 \in \mathcal{X}$, let $\mathcal{C}$ be the class $\mathcal{C} = \{c_1, c_2\}$, $c_1 = \{x_1, x_2\}$, $c_2 = \{x_1\}$. Let $\{\mathcal{P}_j\}_{j=1...m}$ be the following $\gamma$-admissible sequence of distributions on $\mathcal{X}$: $\mathcal{P}_j(x_1) = 1, j = 1, ..., m-t, \mathcal{P}_j(x_1) = 1-\gamma(j-m+t), j = m-t+1, ..., m$, $\mathcal{P}_j(x_2) = 1 - \mathcal{P}_j(x_1), j = 1, ..., m$, with $\gamma t = \epsilon$.

$\mathcal{C}$ and $\mathcal{P}_m$ are an $\epsilon$-binary pair so that we can assume $\mathcal{C}_\epsilon = \{c_1, c_2\}$. Let $\mathcal{D} = (d, 1-d)$ be a distribution over this $\mathcal{C}_\epsilon$. By the noise model it can be easily verified that

$$H(\hat{L}_j \mid X_j) = H(\hat{L}_j \mid X_j, \mathrm{C}) = \mathcal{H}(\eta), \quad j = 1, ..., m-t,$$

while $H(\hat{L}_j \mid X_j)$ equals

$$\mathcal{H}(\eta) + \gamma(j-m+t)\big[\mathcal{H}\big((1-\eta)d + \eta(1-d)\big) - \mathcal{H}(\eta)\big],$$

and $H(\hat{L}_j \mid X_j, \mathrm{C}) = \mathcal{H}(\eta)$, for all $j = m-t+1, ..., m$. Therefore

$$\begin{aligned} I(X^m, \hat{L}^m; \mathrm{C} \mid R) &= I(\hat{L}^m; \mathrm{C} \mid X^m) \\ &= H(\hat{L}^m \mid X^m) - H(\hat{L}^m \mid X^m, \mathrm{C}) \\ &= \beta\big[\mathcal{H}\big((1-\eta)d + \eta(1-d)\big) - \mathcal{H}(\eta)\big], \end{aligned}$$

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 Lemma 3, yielding

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

and then (3) on $\beta = \beta(d)$ by setting

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

We still have $\sup_{d \in [0,1]} \beta(d) \geq \lim_{d \to 0+} \beta(d)$. We employ De l'Hôspital's rule to get

$$\lim_{d \to 0+} \beta(d) = \frac{(1-2\delta)\ln\frac{1-\delta}{\delta}}{\epsilon(1-2\eta)\ln\frac{1-\eta}{\eta}} \tag{14}$$

Since $\frac{1-\eta}{\eta} = 1 + \frac{1-2\eta}{\eta}$ we have $\ln\frac{1-\eta}{\eta} = \ln(1 + \frac{1-2\eta}{\eta}) \leq \frac{1-2\eta}{\eta}$ the inequality following from $\ln(1+x) \leq x$. Plugging the last inequality into the denominator of (14) gets

$$\beta \geq \frac{(1-2\delta)\,\eta\,\ln\frac{1-\delta}{\delta}}{(1-2\eta)^2} \tag{15}$$

From $\beta = \frac{\epsilon}{2}\left(\frac{\epsilon}{\gamma} + 1\right)$, and (15) we obtain the claimed bound on $\gamma$ holding for every $m$. □

Hence the noise rate $\eta$ affects in a significant way the drifting constant $\gamma$. The specialization to $(\mathcal{P}, \mathcal{P}^m)$-learning in the classification noise model (i.e., with $\gamma = 0$) is obtained by setting $t = m, \mathcal{P}_j = \mathcal{P}_m = \mathcal{P}, j = 1, ..., m$. In part 1 if $A$ is a $(\mathcal{P}, \mathcal{P}^m)$-learning function for $\mathcal{C}$ under these conditions then its sample complexity must satisfy (12). In the proof of part 1 we showed that its LHS is $\Omega(d_{VC})$ and that 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 part 2, under these conditions, we can replace $\beta$ by $m\epsilon$ and from (15) 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*, i.e., $A$ can ask 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 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 [28], Sakakibara [29], Angluin et al. [3] and the references in those papers. Perhaps the references closest to our work are Eisenberg and Rivest [17] and Turàn [33]. 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, the main theorem 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 $(X_1, \hat{L}_1), \ldots, (X_{i-1}, \hat{L}_{i-1})$. Therefore the query models have the important property that

$$I(X_i, c \mid (X_1, \hat{L}_1), \ldots (X_{i-1}, \hat{L}_{i-1}), R) = 0. \qquad (16)$$

This is equivalent to a "data processing inequality" [13, 24].

**Theorem 8**
*1) 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 $\epsilon$-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*

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

*2) Let $\mathcal{C}$ and $\mathcal{P}$ be an $\epsilon$-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*

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

*Proof.* 1) Let $\mathcal{D}$ be the uniform distribution over $\mathcal{C}_\epsilon$. From (2) and Lemma 2 we obtain

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

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

$$\begin{aligned} & I(X^m, \hat{L}^m;\, c \mid R) \\ &= \sum_{i=1}^{m} I(X_i, \hat{L}_i;\, c \mid (X_1, \hat{L}_1), \ldots, (X_{i-1}, \hat{L}_{i-1}), R) \\ &= \sum_{i=1}^{m} \Big( I(X_i;\, c \mid (X_1, \hat{L}_1), \ldots, (X_{i-1}, \hat{L}_{i-1}), R) \\ &\qquad + I(\hat{L}_i;\, c \mid X_i, (X_1, \hat{L}_1), \ldots, (X_{i-1}, \hat{L}_{i-1}), R) \Big) \end{aligned}$$

The first term is zero by (16), we remove it and expand the last mutual information expression.

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

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

2) We can apply Lemma 3 after setting $\mathcal{D} = (d, 1-d)$ and recalling that $\mathcal{C}_\epsilon = \mathcal{C}$. Along with the proof of part 1 above this gives

$$\begin{aligned} & \mathcal{H}\big((1-\delta)d + \delta(1-d)\big) - \mathcal{H}(\delta) \\ &\leq \sum_{i=1}^{m} H(\hat{L}_i \mid X_i, (X_1, \hat{L}_1), \ldots, (X_{i-1}, \hat{L}_{i-1}), R) \\ &\quad - m\mathcal{H}(\eta). \end{aligned} \qquad (17)$$

We now upper bound each $H(\hat{L}_i \mid X_i, (X_1, \hat{L}_1), \ldots, (X_{i-1}, \hat{L}_{i-1}), R)$ by $H(\hat{L}_i \mid R)$ and notice that the weakest bound results when each $H(\hat{L}_i \mid R)$ 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),

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

where $p_i = \Pr_{\mathcal{M}}(X_i = x_1 \mid c_2, r), i = 1, ..., m$ represent the probabilistic[9] query strategy of the algorithm in this context. For any $i$ and for any $r$,

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

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 and the entropy is maximized when $p_i = 0$. This yields $\sum_{i=1}^{m} H(\hat{L}_i \mid R) \leq m\mathcal{H}\big((1-\eta)d + \eta(1-d)\big)$. Plugging this bound into (17) and solving for $m$ gives

$$m(d) = \frac{\mathcal{H}\big((1-\delta)d + \delta(1-d)\big) - \mathcal{H}(\delta)}{\mathcal{H}\big((1-\eta)d + \eta(1-d)\big) - \mathcal{H}(\eta)},$$

so

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

This concludes the proof. □ [Theorem 8]

Note that one can easily find a concept class $\mathcal{C}$ and a distribution $\mathcal{P}$ where $N(\mathcal{C}, \epsilon, \mathcal{P}) = \Theta(|\mathcal{C}_\epsilon|)$ for which a learning algorithm can use membership queries to perform a binary search in a smallest $\epsilon$-cover of $\mathcal{C}$. One example is the class of initial segments of [0,1] with the uniform distribution, as mentioned in Eisenberg and Rivest [17]. This shows that, at least in the $\eta = 0$ case, part 1 of Theorem 8 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 the first part of the theorem holds for any query model $\mathcal{M}$. By Lemma 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 $\mathcal{M}$.

Part 1 generalizes a lower bound by Turàn [33, Theorem 1] in two directions.

- It measures the descriptive complexity of $\mathcal{C}$ w.r.t. $\mathcal{P}$ using the size of a largest $\epsilon$-well-separated subclass of $\mathcal{C}$. Thus if $\mathcal{P}$ is the distribution over $d_{VC}$ shattered points mentioned in [16], (so $N = \Omega(d_{VC})$, see [5]), 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 part 1 to the concept class and the distribution mentioned in Theorem 6 yields the tighter bound $m \geq \Omega(\frac{d_{VC}}{(1-2\eta)^2}\log\frac{1}{\epsilon})$.
- Turàn's result is specifically for the noise-free case, and our bound includes a dependence on the noise rate.

Although the proof of part 1 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.

Part 1 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 part 2 is due solely to the difficulty of learning with noise, as it takes only a single noise-free query to learn an $\epsilon$-binary pair.

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

## 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 [32].

There are several directions in which this work can be extended. Theorem 6 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 6 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).

**Acknowledgments**
We thank the COLT program committee members for their helpful comments.

## References

[1] D. Angluin, "Learning regular sets from queries and counterexamples", *Information and Computation*, vol. 75, n. 2, pp. 87-106, 1987.

[2] D. Angluin, "Queries and concept learning", *Machine Learning*, vol. 2, n. 4, pp 319-342, 1988.

[3] D. Angluin, M. Krikis, R. Sloan, G. Turán, "Malicious Omissions and Errors in Answers to Membership Queries", *Machine Learning*, vol. 28, no. 2/3, pp. 211-255, 1997.

[4] D. Angluin, P.D. Laird, "Learning from Noisy Examples", *Machine Learning*, vol. 2, no. 2, pp. 343-370, 1988.

[5] B. Apolloni, C. Gentile, "Sample Size Lower Bounds in PAC Learning by Algorithmic Complexity Theory", *Theor. Comp. Sc., to appear.*

[6] J.A. Aslam, S.E. Decatur, " On the sample complexity of noise-tolerant learning", *Inf. Proc. Lett.*, vol. 57, pp. 189-195, 1996.

[7] P. Bartlett, "Learning with a slowly changing distribution", in *Proc. of the 5th Workshop on Comput. Learn. Th.*, 1992, pp. 243-252.

[8] P. Bartlett, D. Helmbold, manuscript, 1996.

[9] R.D. Barve, P.M. Long, "On the complexity of learning from drifting distributions", *Information and Computation*, vol. 138, no. 2, pp. 170-193, 1997.

[10] G. Benedek, A. Itai, "Learnability by Fixed Distributions", *Theor. Comp. Sc.*, vol. 86, no. 2, pp. 377-389, 1991.

[11] A. Blumer, A. Ehrenfeucht, D. Haussler, M. Warmuth, "Learnability and the Vapnik-Chervonenkis Dimension", *J. of ACM*, vol. 36, pp. 929-965, 1989.

[12] N. Cesa-Bianchi, E. Dichterman, P. Fischer, E. Shamir, H.U. Simon, "Sample-efficient Strategies for Learning in the Presence of Noise", *eCOLT Tech. Rep. 97-003, WWW: http://ecolt.informatik.uni-dortmund.de/.* Preliminary versions in *28th STOC, 1996 and 3rd EuroCOLT, 1997*

[13] T.M. Cover, J.A. Thomas, *Elements of information theory.* NY: John Wiley & Sons, Inc., 1991.

[14] R.L. Dobrushin, "General formulation of Shannon's main theorem in information theory", *Uspekhi Mat. Nauk*, vol. 14, pp. 3-104, 1959; translated in *Amer. Math. Soc. Translations*, Ser. 2, vol. 33, pp. 323-438, 1963.

[15] R.M. Dudley, *A Course on Empirical Processes.* Lecture Notes in Mathematics, vol. 1097, Springer-Verlag, Berlin/New York, 1984.

[16] A. Ehrenfeucht, D. Haussler, M. Kearns, L. Valiant, "A General Lower Bound on the Number of Examples Needed for Learning", *Information and Computation*, vol. 82, no. 3, pp. 247-261, 1989.

[17] B. Eisenberg, R.L. Rivest, "On the Sample Complexity of Pac-Learning using Random and Chosen Examples", in *Proc. of the 3th Workshop on Comput. Learn. Th.*, 1990, pp. 154-162.

[18] C. Gentile, "A note on sample size lower bounds for PAC-learning", manuscript, 1997.

[19] D. Haussler, A. Barron, "How well do Bayes methods work for on-line prediction of {-1,+1} values?", in *Proc. of the 3rd NEC Symposium on Computation and Cognition*, 1992, pp. 74-100.

[20] D. Haussler, M. Kearns, R. Schapire, "Bounds on the Sample Complexity of Bayesian Learning Using Information Theory and the VC Dimension", *Machine Learning*, vol. 14, pp. 84-114, 1994.

[21] D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting $\{0,1\}$ functions on randomly drawn points. *Information and Computation*, 115(2):284–293, 1994.

[22] D. Haussler, M. Opper, "Mutual Information, Metric Entropy, and Cumulative Relative Entropy Risk", *Annals of Statistics*, 1997. To appear.

[23] D. Helmbold, N. Littlestone, P. Long, "Apple Tasting", manuscript 1997. An extended abstract appeared in *Proc. of the 33rd Symposium on the Foundations of Comp. Sci.*, 1992, pp.493-502.

[24] S. Ihara, *Information theory for continuous systems.* River Edge, NJ: World Scientific, 1993.

[25] M. Kearns, M. Li, "Learning in the presence of malicious errors", *SIAM J. Comput*, vol. 22, pp. 807-837, 1993.

[26] A.N. Kolmogorov, V.M. Tihomirov, "$\epsilon$-entropy and $\epsilon$-capacity of sets in functional spaces", *Amer. Math. Soc. Translations (Ser. 2)*, vol. 17, pp. 277-364, 1961.

[27] P. Laird, *Learning from Good and Bad Data.* Kluwer International Series in Engineering and Computer Science, Kluwer Academic Publishers, Boston, MA, 1988.

[28] W. Maass, G. Turán, "On the complexity of learning from counterexamples and membership queries", in *Proc. of the 31th Symposium on the Foundations of Comp. Sci.*, 1990, pp. 203-210.

[29] K. Sakakibara, "On learning from queries and counterexamples in the presence of noise", *Inf. Proc. Lett.*, vol. 37, no. 5, pp. 279-284, 1991.

[30] H.U. Simon, "General Bounds on the Number of Examples Needed for Learning Probabilistic Concepts", *Journal of Comp. System Sci.*, vol. 52, no. 2, pp. 239-254, 1996.

[31] R. Sloan, "Four types of noise in data for PAC learning", *Inf. Proc. Lett.*, vol. 54, pp. 157-162, 1995.

[32] G. Shackelford, D. Volper, "Learning k-DNF with noise in the attributes", in *Proc. of the 1988 Workshop on Comput. Learn. Th.*, 1988, pp. 97-103.

[33] G. Turàn, "Lower bounds for PAC learning with queries", in *Proc. of the 6th Workshop on Comput. Learn. Th.*, 1993, pp. 384-391.

[34] L. Valiant, "A theory of the learnable", *Communication of ACM*, vol. 27, no. 11, pp. 1134-1142, 1984.

[35] V.N. Vapnik, *Estimation of dependencies based on empirical data.* NY: Springer Verlag, 1982.

[36] V.N. Vapnik, *The Nature of Statistical Learning Theory.* NY: Springer Verlag, 1995.

[37] B. Yu, "Lower Bounds on Expected Redundancy for Nonparametric Classes", *IEEE Trans. on Inf. Th.*, vol. 42, no. 1, pp. 272-275, 1996.

# A Appendix

*Proof of Fact 1 used in Theorem 4*

We exploit a property of the entropy function $\mathcal{H}[p_1, ..., p_n]$ that for $n = 3$ reduces to the following [24, 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

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

For $\mathcal{H}\big[(1-\eta)(1-\epsilon), d(1-\eta)\epsilon+(1-d)\eta, d\eta+(1-d)(1-\eta)\epsilon\big]$ 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 $\mathcal{H}\big[(1-\eta)(1-\epsilon), (1-\eta)\epsilon, \eta\big]$ 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. □

*Proof of Lemma 5, 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 $\eta$, $\mathcal{H}(\eta+\alpha)$ is obviously concave in $\alpha$ and therefore by a first-order Taylor expansion around $\alpha = 0$ we get

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

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