Learning Distributions by their Density Levels - A Paradigm for Learning Without a Teacher

Shai Ben-David* Michael Lindenbaum
Computer Science Department, Technion
Haifa 32000, ISRAEL
email: {shai,mic}@cs.technion.ac.il

Abstract

We propose a mathematical model for learning the high-density areas of an unknown distribution from (unlabeled) random points drawn according to this distribution. While this type of a learning task has not been previously addressed in the Computational Learnability literature, we believe that this it a rather basic problem that appears in many practical learning scenarios.

From a statistical theory standpoint, our model may be viewed as a restricted instance of the fundamental issue of inferring information about a probability distribution from the random samples it generates. From a computational learning angle, what we propose is a new framework of un-supervised concept learning. The examples provided to the learner in our model are not labeled (and are not necessarily all positive or all negative). The only information about their membership is indirectly disclosed to the student through the sampling distribution.

We investigate the basic features of the proposed model and provide lower and upper bounds on the sample complexity of such learning tasks. Our main result is that the learnability of a class of distributions in this setting is equivalent to the finiteness of the VC-dimension of the class of the high-density areas of these distributions. One direction of the proof involves a reduction of the density-level-learnability to p-concepts learnability, while the sufficiency condition is proved through the introduction of a generic learning algorithm.

Keywords
Learning Theory, PAC, Vapnik-Chervonenkis dimension, ϵ-approximation, un-supervised learning.

*This research was supported by the David and Ruth Moskowitz Academic Lectureship

1 Introduction

Identification of high-probability-density areas is a basic step in many real-life un-supervised-learning tasks. Consider, for example, the identification of high-risk groups in a population: A physician may wish, on the basis of records of patients effected by some disease, to infer the attribute values of the subgroups of the population which are at high risk of contracting this disease. One may assume that the overall distribution of the population over the attributes space is known to the researcher and serves as a baseline relative to which risk (i.e., the density of the distribution of sick people) is defined. Note that, in the situation we consider here, the physician has access to files of sick people only. Consequently, we may view his data as a sample drawn from the unknown distribution that he wishes to assess – namely, that of people effected by the disease.

A similar analysis is relevant to a wide range of issues in social studies including the identification of accident-prone drivers (from records of drivers involved in accidents + general statistics of the entire population) and certain aspects of marketing analysis.

A different area in which such tasks frequently arise is pattern recognition. In many pattern recognition scenarios, one is faced with a large collection of feature vectors, every one of which characterizes an instance of a class. The classes themselves, as well as the labels associated with each feature vector are not given. Experience shows that feature vector which correspond to the same class tend to cluster together, forming high density area in the feature vector space. A common demand is to identify these clusters and to report their number, position, size and shape, thereby getting insight to the nature or structure of the data, [DH73]. The identification of high-probability-density areas plays a central role in such task of classification via clustering.

Yet another relevant scenario arises in computer vision, when one wishes to identify familiar features in noisy images. A simple but real example may be the detection of straight object boundaries in the image. The points are not labeled, and the only information about their membership in a boundary segment is indirectly disclosed to the student through the sampling distribution.

The methods developed by the pattern recognition and computer vision research communities to handle such problems are usually heuristic and rely on the particular tasks. Partitions based on graph algorithms, (minimum spanning tree,) for example, are used for clustering in a feature space [DH73]. Methods that look for maximal consistency of a concept with the data are used for finding the straight edges [IK88].

In the context of Computational Learning Theory such tasks fall into the realm of *un-supervised learning*. It seems that un-supervised learning has, so far, attracted only limited attention in the Computational Learnability research, mainly under the title of *learning from positive examples*.Natarajan [Nat91] provides a necessary and sufficient condition for distribution-free learnability from positive examples. This condition is very restrictive and rules out most of the interesting examples one may wish to consider in Computer Vision or Pattern Recognition tasks. Kim [Kim91], restricts his attention to limited classes of 'nicely

behaving' distributions over R^n and offers an algorithm for learning geometrical objects with respect to such classes.

From a wider mathematical perspective, the learning tasks mentioned above can all be viewed as instances of the fundamental problem of inferring information about some unknown probability distribution on the basis of independent draws from that distribution. In its most demanding form, one wishes to come up with an approximation of the unknown distribution. This is a well studied problem in statistics and pattern recognition literature. Some variants of this fundamental task have been recently investigated in the context of computational learning theory by Kearns et al [KMRRSS94].

The starting point of this work is the observation that, for many un-supervised learning tasks (including those mentioned above), a much weaker type of information suffices. Rather than attempting to infer an approximation to the unknown distribution, we settle for the task of learning its high-probability-density areas. Given a 'threshold level', r, we ask the student to infer $D_r^+ \stackrel{\text{def}}{=} \{x \ : \ d(x) \geq r\}$ – the set of all points having probability density above r (the probability density is, of course, that of the unknown distribution generating the random examples, and is defined relative to some known basic distribution over the domain). We call a class, $\mathcal{D}$, of distributions *density-level-learnable* if there exists a student, $S_{\mathcal{D}}$, that, upon receiving a finite sample drawn according to some $D \in \mathcal{D}$ and a real number, r, outputs an approximation to D_r^+ (the exact definition of 'approximation', as well as the bounds imposed on the input sample size, are defined later in section 2).

We begin the paper by describing, in section 2, our basic framework of density-level-learnability. That section concludes with a brief comparison of our notion of approximating a probability density function to more common approaches. Section 3 investigates learnability in this model and there we establish a sufficiency condition for density level learnability, based on the finiteness of the VC dimension of an appropriate class.

In section 4 we turn our attention to un-supervised concept learning. Here we wish to model situations in which the sample distribution displays a step-like behavior – it has high density inside some 'target' area and a low density elsewhere. This is, for example, the situation in many pattern recognition settings where, ideally, sample points should have been generated by the target only, but due to noise effects, they occur also in other areas of the picture, with some lower probability density.

We apply the framework presented before to propose a model of concept learning from unlabeled examples. The model, *learning Without A Teacher, (WAT)*, reflects a situation in which a student detects (unlabeled) sample points that are randomly generated all over the scene (inside and outside the target concept). The information about the target concept comes through a dependence of the generating distribution upon this target. We assume that for points outside the target the distribution density is lower than a certain threshold α, while inside the target the density exceeds some value $\beta > \alpha$. The section concludes with a proof of a sufficient VC-dimension condition for learnability in this WAT model.

The last section of this paper is devoted to proving the necessary condition results that complement the sufficiency results for learnability in our models. A general result is obtained

by showing a reduction of WAT learnability to PAC learnability. More concrete sample-complexity lower bound are derived from basic probability considerations and estimates for tail probabilities of Bernoulli processes, and by a a reduction of WAT learnability to p-concept learnability.

2 Learning a distribution by its density levels

We start by presenting the learning framework. Our model of learning is based on some fixed measure space, $(\mathcal{X}, \mathcal{B}, \mu)$. I.e., $\mathcal{X}$ is a domain set, $\mathcal{B} \subseteq 2^{\mathcal{X}}$ is an algebra of measurable sets and μ is a probability measure. The measure μ is used as a reference, relative to which the density of the unknown probability distribution (the distribution generating the learning examples) is defined.

One may think of this measure μ as the uniform measure when $\mathcal{X}$ is a finite set or as the Euclidean volume for domains which are bounded subsets of some R^n. In applications such as identifying high-risk groups in a population, μ would typically be the distribution of the entire population over the attribute space.

Definition 1: Let D be a probability distribution over $(\mathcal{X}, \mathcal{B})$ and let d be its density function with respect to μ (we assume that all the distributions, D, discussed are absolutely continuous w.r.t. μ). The r^+-**level** of D is $D_r^+ = \{x \ : \ d(x) \geq r\}$.

The task of the student is, given a positive 'level', r, to infer the set D_r^+ from unlabeled examples generated independently at random according to the distribution D. To make this task achievable, the student would also get, as input, a class $\mathcal{D}$ of distributions to which the distribution D belongs[1].

Thus, while deviating from the PAC model by considering only unlabeled examples, we do adopt the basic PAC framework of learning target sets (the sets D_r^+) from randomly drawn examples on the basis of some a-priory knowledge – a class to which the target belongs. We also borrow from the PAC scene the idea of approximate learning. In our setting the quality of an approximation depends not only an a 'size of error' parameter, ϵ, but also on a parameter ρ measuring the difference between the true probability density at a point and its hypothesized value.

Definition 2: A hypothesis, h, is (ϵ, ρ)-**close** to an r^+-level of a distribution, D, if

$$\mu(\{x : x \in D_r^+ \Delta h \ , \ |r - d(x)| \ > \rho\}) \leq \epsilon$$

[1] A different interesting approach is that of *agnostic learning*. In an agnostic model no assumption is made about the membership of the target in the class provided to the student. Rather than asking for a close approximation of the target, the student is only required to pick a hypothesis which is close to the best approximation of the target by a member of the class.

In other words, an (ϵ, ρ)-close hypothesis may add to D_r^+ an arbitrarily large set of points whose density is below r, as long as this density is not below $(r - \rho)$. Similarly, such a hypothesis may miss any set of points whose density is above r and below $(r + \rho)$. However, making mistakes larger then ρ, in assessing the density of a point, is limited to a set of points of μ-measure at most ϵ.

An alternative way to define this notion of success is through a loss function.

Definition 3: Let D be a distribution as above and $r \geq 0$,

- For $\rho > 0$, let $l_\rho(x) = \begin{cases} 1 & \text{if } |d(x) - r| > \rho \\ 0 & \text{otherwise} \end{cases}$
- Let $l_1(x) = |d(x) - r|$.
- For $h \subseteq \mathcal{X}$, let $L_\rho(h) = \int_{D_r^+ \Delta h} l_\rho(x) d\mu, \quad L_1(h) = \int_{D_r^+ \Delta h} l_1(x) d\mu$

Note that h is (ϵ, ρ)-close to D_r^+ iff $L_\rho(h) \leq \epsilon$. We have chosen to present our results via the L_ρ notion of approximation. It is not hard to verify that their natural variations, defined in terms of the L_1 notion, hold as well.

Having defined our notion of approximation, we now proceed to define the correlating notions of a successful student and of learnability, in the spirit of the corresponding definitions in the PAC framework.

Definition 4: Let $\mathcal{D}$ be a family of probability distributions over a domain space $(\mathcal{X}, \mathcal{B}, \mu)$. Let m denote a natural number and ρ, ϵ, δ are positive real valued parameters.

1. A **student** is a function from $\mathbb{R}^+ \times \bigcup_{m \in \mathbb{N}} \mathcal{X}^m$ to subsets of $\mathcal{X}$. A student, S, is $(m, \epsilon, \delta, \rho)$-**successful** if, for every $r \in \mathbb{R}^+$, and for an m-tuple, $\bar{x}$, of points in $\mathcal{X}$ which are generated independently at random according to some $D \in \mathcal{D}$ then, it provides an hypothesis $S(r, \bar{x})$, which is (ϵ, ρ)-close to D_r^+, with probability exceeding $(1 - \delta)$.

2. A class $\mathcal{D}$ of probability distributions is **density level learnable** if, for every ϵ, δ, ρ, there exists a finite m and an $(m, \epsilon, \delta, \rho)$-successful student for $\mathcal{D}$.

3. In the course of this paper we shall freely use variants of the above definitions – fixing some of the parameters and universally quantifying over the unmentioned parameters. In particular, a class $\mathcal{D}$ is (r, ρ)-**learnable** if for every positive ϵ, δ, there exists a finite m and a function (student) from $\mathcal{X}^m$ to subsets of $\mathcal{X}$, that upon seeing m-tuples of points generated by some $D \in \mathcal{D}$, outputs hypotheses that are (ϵ, ρ) close to D_r^+ with probability exceeding $(1 - \delta)$.

2.1 Density-Levels approximation vs. more common approaches

In this subsection we wish to clarify how the information obtained by a density-level student relates to more common tools for approximating unknown distributions. The bottom line would be that density-level approximation is a fine tool that enables the definition of learning tasks that are strictly weaker (i.e., less informative) than those defined by the other common methods. We have already mentioned in the introduction that this weakness allows for the learnability of wider classes of distributions, and still does not hurt a wide range of applications of un-supervised learning. One should also bare in mind that, as far as lower bound results go, results in a weaker model of learning readily imply similar results for stronger models.

Definition 5: Let $(\mathcal{X}, \mathcal{B}, \mu)$ be a measure space and let $\mathcal{F}_\mathcal{X}$ denote the class of real valued functions on $\mathcal{X}$ the are measurable with respect to $\mathcal{B}$.

1. For $f \in \mathcal{F}_\mathcal{X}$ let $f_r^+ \stackrel{\text{def}}{=} \{x : f(x) \geq r\}$.
2. A set, $A \subseteq \mathcal{X}$, is (ϵ, ρ)-**close to** f_r^+ if, $\mu(\{x : x \in f_r^+ \Delta A \ , \ |r - f(x)| \ > \rho\}) \leq \epsilon$
3. Two of functions $f, g, \in \mathcal{F}_\mathcal{X}$ are (ϵ, ρ)-**close** if, for all $r \in \mathbb{R}$, f_r^+ is (ϵ, ρ)-close to g_r^+.

Observation 1:

1. Let D, D' be two probability distributions over $(\mathcal{X}, \mathcal{B}, \mu)$ and let d, d' be their respective density functions. If, for some (ϵ, ρ), d is (ϵ, ρ)-close to d', then, for every r, D_r^+ and $D_r'^+$ have a mutual (ϵ, ρ)-close hypothesis.
2. If, for every r, h_r is an (ϵ, ρ)-close hypothesis for D_r^+, then the function $f(x) \stackrel{\text{def}}{=} \sup\{s : x \in h_s\}$ is (ϵ, ρ)-close to the density function, d, of D.

In other words, learning to (ϵ, ρ) approximate all the density levels, D_r^+, of a distribution, D, is equivalent to coming up with a function which is (ϵ, ρ)- close to the density function of D.

We now turn to the comparison of the (ϵ, ρ) notion of proximity to the common L_1 and L_∞ measures. For functions $f, g, \in \mathcal{F}_\mathcal{X}$, let $L_1^\mu(f, g)$ denote $\int_\mathcal{X} |f(x) - g(x)| d\mu$ and let $L_\infty(f, g)$ denote $\sup_{x \in \mathcal{X}} |f(x) - g(x)|$.

Claim 1:

- If $L_\infty(f, g) \leq \rho$ then, for all ϵ, f is (ϵ, ρ)-close to g.
- If $L_1^\mu(f, g) \leq M$ then, for all ϵ, ρ satisfying $\epsilon \cdot \rho \geq M$, f is (ϵ, ρ)-close to g.

The other direction of these implications is false. For every positive (ϵ, ρ), there exist (nicely behaving) density functions f_1, g_1, f_2, g_2 such that f_1 is (ϵ, ρ)-close to g_1, f_2 is (ϵ, ρ)-close to g_2 and, yet, $L_1^{\mu}(f_1, g_1)$ and $L_{\infty}(f_2, g_2)$ are arbitrarily large.

The main point we would like to note about clain 1 is that both L_{∞} and L_1^{μ} approximations of the density function d result in a function that is (ϵ, ρ)-close to d, this notion of closeness implies, in turn, that *for all* r the density levels of the hypothesis function are (ϵ, ρ)-close to the r-density levels of the learnt distribution. In contrast, our model of (r, ρ)-learnbility enables the separation of certain significant levels without bothering about the complexity of the density function in other levels.

As it turns out, most of the results of this paper can be readily extended to apply to learning density functions in the L_1 norm. We stick with the notion of (r, ρ)-learnbility mainly in order to allow for the flexibility of caring for only certain significant density levels and ignoring the behaviour of the target distributions on all irrelevant levels.

3 Characterizing density level learnability.

The fundamental theorem of PAC learnability, namely the [BEHW89] characterization of learnability, states that the finiteness of the VC-dimension of a concept class is both necessary and sufficient for its PAC-learnability. Furthermore, Blumer et al show that for classes having a finite VC-dimension, any consistent student is successful.

We would like to state an analogous result for learning density levels in the sense described above. Unlike the traditional PAC-learning framework, once one considers un-supervised examples consistency becomes a vacuous notion. (Even in the context of learning from positive examples, where one assumes that all the unlabeled examples belong to the target concept, once a maximal concept exists in a class, it will be consistent with any given sample). We shall replace the notion of consistency by a weaker notion that we call (r, η)-consistency.

Definition 6: Let C be a collection of subsets of a domain set, $\mathcal{X}$, and let A be a finite subset of $\mathcal{X}$. For positive reals r, η, and a hypothesis h (i.e. $h \subseteq \mathcal{X}$), we say that h **is** (r, η)-**consistent** with A relative to C if, for every $c \in \mathcal{C}$:

- If $c \cap h = \emptyset$ then $\frac{|A \cap c|}{|A|} < r \times \mu(c) + \eta$
- If $c \subseteq h$ then $\frac{|A \cap c|}{|A|} > r \times \mu(c) - \eta$

The definition comes to state that, as far as elements of $\mathcal{C}$ are concerned, h is a conceivable hypothesis for D_r^+. The idea behind this definition is that A stands for a sample, providing η-good empirical estimates of the D-probability of every member of $\mathcal{C}$.

The main result of this paper is a characterization of the density-level learnability of classes of distributions in terms of the VC-dimension of their density-level classes. The result may be viewed as variant of the basic [BEHW89] characterization of PAC learnability.

Here, the notion of the (r,η)-consistency plays the role of the consistency condition in the PAC framework. We show that the finiteness of the VC-dimension of the density level class is a necessary and sufficient condition for learnability. Furthermore, we show that, for density level classes having a finite VC-dimension, for every $0 \leq \rho \leq 1$ there exists an η for which any (r,η)-consistent student is (ϵ,ρ) - successful.

The proof of the sufficiency condition is based on the theory of ϵ-approximations. This theory investigates conditions under which randomly drawn samples provide good estimators for the probabilities of a set of events, simultaneously.

Definition 7: For a measure space $(\mathcal{X},\mathcal{B},D)$, a (finite) subset $Y \subset \mathcal{X}$ is an ϵ-approximation of D for the class $\mathcal{C} \subset \mathcal{B}$ if, $\forall c \in \mathcal{C}$, $||Y \cap c|/|Y| - D(c)| < \epsilon$.

($D(c)$ denotes the probability of the event c.) It turns out that if a class, $\mathcal{C}$, has a small VC-dimension, then there are many small ϵ-approximations for it. The following theorem is due to Vapnik and Chervonenkis [VC71].

Theorem 1: [[VC71]] There is a positive constant s such that if $\mathcal{C}$ is any concept class of VC-dimension at most d, and $\epsilon,\delta > 0$, then with probability at least $1-\delta$, a randomly selected subset of size

$$N_{approx}(d,\epsilon,\delta) \stackrel{\text{def}}{=} \frac{s}{\epsilon^2}(d \log \frac{d}{\epsilon} + \log \frac{1}{\delta})$$

is an ϵ-approximation of D for $\mathcal{C}$.

We are now ready to prove the sufficiency part of the learnability characterization, the necessity part is deferred to section 5.

Theorem 2: Let $\mathcal{D}$ be a family of probability distributions over some domain space $(\mathcal{X},\mathcal{B},\mu)$. Let $\mathcal{C}_{\mathcal{D}}^{lev}$ denote the class of all r^+-levels of members of $\mathcal{D}$, i.e.,

$$\mathcal{C}_{\mathcal{D}}^{lev} \stackrel{\text{def}}{=} \{D_r^+ : D \in \mathcal{D}\ ,\ r \in \mathbb{R}\}.$$

1. If VC-dim$(\mathcal{C}_{\mathcal{D}}^{lev}) < \infty$, then $\mathcal{D}$ is density level learnable.

2. Furthermore, for any ρ and r, let

$$\mathcal{C}_{r,\rho}^{*} \stackrel{\text{def}}{=} \{D_r^+ \setminus D'^+_{r-\rho} \ : D,D' \in \mathcal{D}\} \cup \{D^+_{r+\rho} \setminus D'^+_r \ : D,D' \in \mathcal{D}\},$$

   and let k denote its VC-dimension, then for $0 < \epsilon,\delta < 1$ if a sample of size at least $m = N_{approx}(k,\rho\epsilon/4,\delta)$ is drawn at random according to some $D \in \mathcal{D}$ then, with probability exceeding $1-\delta$, any hypothesis $h \in \{D_r^+ : D \in \mathcal{D}\}$ which is $(r,\rho\epsilon/4)$-consistent relative to $\mathcal{C}_{r,\rho}^{*}$, is (ϵ,ρ)-close to D_r^+.

Proof: First, note that part 2 of the theorem implies part 1. That is because, using standard VC calculation arguments[2] if VC-dim$(\mathcal{C}_{\mathcal{D}}^{lev}) = l$ then VC-dim$(\mathcal{C}_{r,\rho}^{*}) \leq 2l \log l$. We therefore proceed to prove part 2.

[2]The argument that Dudley, [D84], uses for the calculation of the dimension of classes of intersections are applicable here as well.

The proof is in two parts. First, we show that if a sample is an ϵ-approximation for $\mathcal{C}^*_{r,\rho}$, then any sufficiently consistent hypothesis in $\{D^+_r : D \in \mathcal{D}\}$ is (ϵ, ρ)-close to D^+_r. Then we show that with such a sample, there exists some hypothesis that is sufficiently consistent (namely, the true target, D^+_r). Therefore, the $(r, \rho\epsilon/4)$-consistency criterion fails to provide a satisfactory hypothesis only if the sample is not an ϵ-approximation, an event of probability at most δ.

- First, note that a hypothesis, h, is (ϵ, ρ)-close to D^+_r iff $\mu(D^+_{r+\rho} \setminus h) + \mu(h \setminus D^+_{r-\rho}) \leq \epsilon$. Let us show that a hypothesis, h, as assumed, satisfies $\mu(D^+_{r+\rho} \setminus h) \leq \epsilon/2$ (the other half needed to prove the in equation is proved similarly). Let h be a hypothesis in $\{D^+_r : D \in \mathcal{D}\}$. Note that, for the target distribution, D, $D^+_{r+\rho} \setminus h \in \mathcal{C}^*_{r,\rho}$. If A is a random sample drawn according to D then, by theorem 1, $|A| > m$ implies that, with probability exceeding $(1 - \delta)$,

  $$\begin{aligned} \frac{|A \cap (D^+_{r+\rho} \setminus h)|}{|A|} &\geq Prob(D^+_{r+\rho} \setminus h) - \rho\epsilon/4 \\ &\geq (r+\rho)\mu(D^+_{r+\rho} \setminus h) - \rho\epsilon/4 \quad = r\mu(D^+_{r+\rho} \setminus h) + \rho\mu(D^+_{r+\rho} \setminus h) - \rho\epsilon/4. \end{aligned}$$

  If, by way of contradiction, $\mu(D^+_{r+\rho} \setminus h) > \epsilon/2$, then substituting $\epsilon/2$ for $\mu(D^+_{r+\rho} \setminus h)$ in the second term to the right of the equation sign above implies that $\frac{|A \cap (D^+_{r+\rho} \setminus h)|}{|A|} \geq r\mu(D^+_{r+\rho} \setminus h) + \rho\epsilon/4$, contradicting the $(r, \rho\epsilon/4)$-consistency criterion for h.

- We still have to show that such a student has a non-empty set of sufficiently consistent hypotheses to choose from. This follows directly, however, from our choice of $\eta = \rho\epsilon/4$ and from the definition of (ϵ, ρ) consistency, if the target $T = D^+_r$ itself is taken as the hypothesis.

□

Note that the (γ, η)-consistency criterion depends not only upon the the given examples, but also upon 'examples that are not given'. It will reject any concept that contains sufficiently large areas having too few examples in them.

Interestingly, the sample size is roughly proportional to the intuitive tradeoff between the two measures of accuracy, ρ and ϵ. Given a fixed number of available examples, one can decide whether to invest them in reducing the density uncertainty associated with the density level or in the corresponding 'spatial' uncertainty.

Remark: This paper does not consider the calculation of the VC-dimension for r^+-level classes corresponding to commonly used densities. It seems however, that simple density families elicit simple density levels classes. The r^+-level of a Gaussian is an ellipsoid, and the class of ellipsoids, or more generally polynomial sets has a finite VC-dimension [D84]. The commonly used class of linear combinations of (k) Gaussians in $\mathbb{R}^n$ is more complicated as the r^+-levels are not polynomial anymore. Nevertheless, the theory of fewnomials, which generalizes the topological properties of polynomials to other simple functions [K91], together with the technique developed in [BL93], imply that the VC-dimension is finite (see also [KM95]).

4 Un-Supervised Concept Learning

So far we have been following the approach that views the example-generating distribution as the primal target of learning process. A different approach, prevalent in the context of supervised learning, views subsets of the domain as the target to be learnt, and uses the distribution as just a source of information and as means for measuring the success of a given hypothesis. This is, for example, the attitude underlying the definition of the PAC model. In the PAC model, not only that the targets to be learnt are subsets of the domain, but the secondary role of the example-generating distributions is emphasized by the 'distribution - freeness' assumption, i.e. the requirement that successful learning of any given target should occur regardless of the choice of the distribution. It seems that existing models of *un*-supervised learning do not share this view. As far as we can tell, the only framework of un-supervised learning which, in some sense, may be viewed as aiming at learning subsets of the domain is the task of clustering.

In this section we shall introduce an un-supervised model that shares the PAC model approach of viewing domain subsets as the primal targets and of requiring that the student be able to figure out the right target subsets as long as the example-generating distribution belongs to a certain class of legal distributions.

Our learning task is to figure out a target subset of the domain from (unlabeled) examples generated by any distribution that has a lower-bound on its density inside the target and some strictly lower upper-bound on the density of the distribution in the complement of the target.

We shall apply our density levels learnability results to prove that the finiteness of the VC dimension of a class suffices for its un-supervised learnability in this new model. Then, in the next section, we shall derive lower bounds on the sample size needed for such learning. Theses lower bounds imply that the finiteness of the VC dimension of a class is also a necessary condition for its learnability in this model.

Definition 8: Let $\mathcal{C}$ be a collection of measurable sets in a domain space, $(\mathcal{X}, \mathcal{B}, \mu)$ (i.e., $\mathcal{C} \subseteq \mathcal{B}$), let α, β be real numbers such that $0 \leq \alpha < \beta$.

- A probability distributions $\mathcal{D}$ is (α, β) - **sound for** $\mathcal{C}$ if there exists some $c \in \mathcal{C}$ such that $D^+_\alpha = D^+_\beta = c$.

- An (α, β)-**student** is a function from $\bigcup_{m \in \mathbb{N}} \mathcal{X}^m$ to subsets of $\mathcal{X}$. A student, S, is (m, ϵ, δ)-**successful** for a class C if, for every probability distribution D which is (α, β)-sound for C, given an m-tuple, $\bar{x}$, of points in $\mathcal{X}$ which are generated independently at random according to D, it provides an hypothesis $S(\bar{x})$, which is ϵ-close to D^+_α, with probability exceeding $(1 - \delta)$. (By 'h is ϵ-close to c' we mean that the μ-measure of their symmetric difference is at most ϵ).

- $\mathcal{C}$ **is** (α, β)-**learnable** if there exists an (α, β)-student S, such that for every $\epsilon, \delta \in (0, 1)$ for some finite m, S is (m, ϵ, δ)-successful for C.

- **WAT-learnable** if $\mathcal{C}$ is (α, β)-learnable, for every $0 \leq \alpha < \beta$.

The WAT learning framework is motivated by some practical problems in pattern recognition and computer vision. The scenario we have in mind is that of a target object that, ideally, should have been the only source of data points, but some disturbance, or noise, generates misleading data in other parts of the scene as well. Another practical example is the detection of edge points in a noisy image, when viewed as a binary random sample. Typically in such images the density of the detected points is high in the vicinity of the true edges but it sharply drops in places farther off. The identification of high-risk groups in a population (as discussed in the introduction above) in cases where some factors sharply increase the investigated risk is one more setting of a similar nature.

It should be noted that WAT learning is a special case of density-level learning. Namely, a class C is (α, β)-learnable (in the WAT sense), if and only if, every class of distributions that are (α, β)-sound for C is (r, ρ) - learnable in the density-level sense, for $r = \frac{\alpha+\beta}{2}$, $\rho = \frac{\alpha-\beta}{2}$. In particular, WAT-learnability of a class C is implied by the density-level learnability of the class of all distributions which are (α, β)-sound for C, for some $0 \leq \alpha < \beta$.

From the PAC point of view, WAT learnability may be regarded as distribution-free learnability from positive examples that are generated by a classification-noise source. More concretely, let D be a probability distribution which is (α, β)-sound for C and let d denote its density relative to the underlying distribution, μ. A sample generated by the distribution D may be viewed as the collection of the positive examples generated by first drawing a point $x \in \mathcal{X}$ according to μ and labeling it according to its membership in some target concept $c \in C$, then inverting its label randomly with probability $d(x)$ if the label was 0 and with probability $1 - d(x)$ if the label was 1 and finally outputting only the positively labeled sample points.

There are some differences between this 'noisy-PAC' scenario and the common PAC classification-noise model (see, for example, [AL88]). First we assume that the student receives, as input, only positively labeled examples (rather than revealing to the student all the drawn examples and providing him their labels). The second difference is that we allow the noise to depend upon the drawn example (rather then having a fixed probability of inverting the label of any sampled point). Finally, one should also note that, in our case the underlying distribution, μ, is fixed, and may therefore be viewed as 'known to the student'. This known distribution defines the measure relative to which the quality of the student's hypothesis is defined (but the distribution that generates the examples has an unknown noise component on top of this μ). In the distribution-free PAC scene, the distribution relative to which the distance between a hypothesis and the target is defined is the distribution that generates the examples and it is unknown to the student.

The relation between the complexity of learnability in this un-supervised model and learnability in the common PAC scenario is not simple: In the un-supervised scene, the examples are not labeled, implying that the student gains 'less information' from each example he sees. Indeed, lemma 2 below states that if a class is 'learnable without a teacher' then it

is also PAC learnable. On the other hand, when the target concept is relatively small, a PAC student may see, even after viewing considerably large samples, only few positive examples. In such a case he may conclude that the target has low probability, but is left with very little information regarding its identity. In the WAT framework, as all members of the sample the student gets are positively labeled, he is guaranteed to see many positive points when he views large samples. Consequently, when dealing with targets having a small probability weight, the restriction we make, to distributions concentrating over the target, may give the un-supervised student an edge over the PAC learner.

The sufficiency of the finiteness of the VC-dimension of a class for its WAT learnability is a straightforward consequence of theorem 2 above.

Lemma 1: Every concept class that has a finite VC-dimension is learnable without a teacher.

Simply applying theorem 2 gives the sufficient number of examples for (α, β)-learnability.

Corollary 3: Consider the concept class $\mathcal{C}$. If $d = VCdim(\mathcal{C}) < \infty$, then for every $0 < \alpha < \beta,\ 0 < \delta, \epsilon < 1$ and sample size at least $N_{approx}(2d \log d, \frac{1}{8}(\beta - \alpha)\epsilon, \delta)$, $((\alpha + \beta)/2, \frac{1}{8}(\beta - \alpha)\epsilon)$-consistent hypothesis relative to the difference class $\mathcal{C}^* = \{c \setminus c' \ : c, c' \in \mathcal{C}\}$ is ϵ-close to the target with probability at least $1 - \delta$.

Proof: Apply Theorem 2, for $r = (\alpha + \beta)/2$ and $\rho = (\beta - \alpha)/2$. Observe that for (α, β)-distributions and for these values of r and ρ, an (ϵ, ρ)-close hypothesis (to D_r^+) is also ϵ-close to D_r^+ in the PAC sense. □

In the next section we show that the finiteness of the VC-dimension of a class is also a necessary condition for its WAT learnability.

4.1 A comparison to a common signal detection technique

Those who are familiar with common engineering techniques for signal detection may ask whether this well established theory can be applied also here, to detect high density regions. In Appendix A we examine the most common detection tool, namely the matched filter, in the context of data provided through (α, β)-distributions. We show that while the matched filter performs well for a restricted class of concepts, it may fail for some general concept classes.

5 Sample-Complexity Lower Bounds

So far we have established only an upper bound on the number of examples sufficient for density level learning and un-supervised learning. We have shown that the finiteness of the VC-dimension guarantees learnability in these models.

We now turn our attention to the issue of *lower bounds* on the sample complexity of these learning tasks. Here, we show that the finiteness of the VC-dimension is also necessary for learnability and provide explicit lower bounds on number of needed examples. We consider only the un-supervised learning model. Recalling that this model is a special case of density level learning, both the necessary condition and the sample-complexity lower bound readily hold for the density-level-learning model as well.

We start by constructing a crude reduction from the un-supervised learning model to the PAC learning model. As was already mentioned in the previous section, the exact relations between the sample complexities of learning in these two models are not clear. Nevertheless the reduction allows us to apply the sample complexity lower bound, derived for the PAC model ([BEHW89]), to infer a lower bound on the sample complexity of un-supervised learning. While this bound confirms that the VC-dimension indeed characterizes un-supervised learnability, it does not reflect the details of the dependence of the sample size on the parameters of our models. To clarify these issues, we go on to derive, from first principles of probability theory, a second, refined, lower bound.

Lemma 2: Let $\mathcal{C}$ be a concept class over $(\mathcal{X}, \mathcal{B}, \mu)$. If, for some $0 \leq \alpha < \beta$, $\mathcal{C}$ is (α, β)-*learnable* with $m(\epsilon, \delta)$-many examples, then it is PAC (ϵ, δ)-learnable with respect to the fixed distribution μ, with $O(\frac{m(\epsilon,\delta)}{\epsilon})$-many examples.

Proof:

A PAC student receives labeled examples generated randomly according to μ. By ignoring the negative examples he is left with examples generated according to a distribution that has density 0 outside the target concept, T, and density $\frac{1}{\mu(T)}$ inside the target concept. Noting that any (α, β)-distribution for a target T must satisfy $\beta \leq \frac{1}{\mu(T)}$, we may conclude that the distribution of positive examples generated by μ is (α, β) sound for C, for every target concept $T \in C$.

Ignoring some of the examples does not add information, we may therefore invoke the (α, β) learnability of C and conclude its PAC learnability from sufficiently many positive examples.

To conclude the proof, we should now just compute how many μ-generated examples are needed to guarantee, with high enough probability, that at least m of them are positive. A straightforward calculation shows that $O(m/\mu(T))$ suffice.

Therefore, as a PAC student may ignore concepts of μ-weight below ϵ, $O(m/\epsilon)$-many μ-generated examples suffice for PAC (ϵ, δ)-learnability. □

Combining this lemma with the [BEHW89] necessary condition for PAC learnability, complements lemma 1 to provide an un-supervised variant of the basic [BEHW89] characterization of PAC learnability.

Theorem 4: A concept class is learnable without a teacher iff it has a finite VC-dimension.

As a necessary condition for WAT learning is also necessary for density-level learning, we can now complement theorem 2 to confirm that the difficulty of learning density levels is indeed determined by the VC-dimension of the arising level set classes. Formally,

Theorem 5: Let $\mathcal{D}$ be a family of probability distributions over some domain space $(\mathcal{X}, \mathcal{B}, \mu)$. Let $\mathcal{C}_{\mathcal{D}}^{lev}$ denote (as in theorem 2) the class of all r^+-levels of members of $\mathcal{D}$. Then, $\mathcal{D}$ is density level learnable iff VC-dim$(\mathcal{C}_{\mathcal{D}}^{lev}) < \infty$.

The lower bound provided by the reduction to the PAC framework does not relate, however, the required sample complexity to the ϵ, α, β parameters (or to the ϵ, ρ parameters). It is only natural to expect that, the smaller ϵ and $(\beta - \alpha)$ get, the harder the learning task should become.

In the following lines, we relate the necessary number of examples required for learning to the common statistical problem of calculating the tail probabilities of binomial distribution. Let $\{x_i\}_{i=0}^n$ $(x_i \in \{0,1\})$ be a sequence of i.i.d. binary random variables. For any $0 \leq q < p \leq 1$, let $LB(p, q, \delta)$ denote the minimal sequence length, n, required to guarantee that, with confidence level $1 - \delta$, when the x_i's are generated by the Bernouli-p distribution (i.e., the distribution defined by $prob(x_i = 1) = p$), the statistics $\hat{p} = \frac{1}{n} \sum_{i=1}^n x_i$, is higher than than q. Simon [S93] proves that, if $m < LB(p, q, \delta)$ then no function of the variables $(x_1, \ldots, x_m)$ can distiguish, with probability greater than $(1 - \delta)$, between the case where the x_i's are generated by the Bernouli-p distribution and the case where they are generated by the Bernouli-q distribution.

In the appendix we prove the following lemma, using standard techniques from information theory.

Lemma 3:

$$LB(p, q, \delta) = \Omega\left(\frac{p(1-p)}{(p-q)^2} \log \frac{1}{\delta}\right).$$

To obtain our next lower bound, we shall apply the following non-asymptotic version of this lower bound on tail probabilities. The lemma was communicated to us by Dichterman [D95].

Lemma 4: [Dichterman]

$$LB(p, q, \delta = 0/05) \geq \left(\frac{p(1-p)}{p-q}\right)^2$$

.

Now, we may prove the following lower bound on the sample complexity required for (α, β)-learnability.

Lemma 5: For any $0 \leq \alpha < 1/2$, $\beta = (1 - \alpha)$, and $\epsilon > 0$, if $\mathcal{C}$ has at least two elements that are distinct and have a non-empty intersection then (α, β)- learning of $\mathcal{C}$ with accuracy ϵ and confidence > 0.98 requires at least $\frac{1}{2\epsilon}\left(\frac{\alpha\beta}{\beta-\alpha}\right)^2$ many examples.

Proof: Pick $s, t \in \mathcal{C}$ and points $a, b, c \in \mathcal{X}$ such that $a \in s \cap t$, $b \in s \setminus t$ and $c \in t \setminus s$. Define the underlying distribution μ by setting $\mu(a) = (1 - 2\epsilon)$, $\mu(b) = \mu(c) = \epsilon$, and $\mu(\mathcal{X} \setminus \{a, b\}) = 0$.

We shall now define a pair of distributions, $\{D_s, D_t\}$, that are both (α, β)- sound for $\mathcal{C}$ and require the lemma's sample size to distinguish between them. D_s is defined by the following density function (with respect to μ)

$$\begin{array}{ll} If\ x \notin \{b, c\}\ then & d_s(x) = 1 \\ If\ x = b\ then & d_s(x) = \beta \\ If\ x = c\ then & d_s(x) = \alpha. \end{array}$$

and D_t is defined analogously by switching b and c. By the above result of Simon and lemma 4, no student can distinguish between D_s and D_t with probability greater than 0.95 without examining at least $\left(\frac{\alpha\beta}{\beta-\alpha}\right)^2$ many examples in the set $\{b, c\}$. The lemma now follows by noting that, as for both D_s and D_T the probability of hitting this set is exactly 2ϵ, at least 0.4 of the randomly generated samples of a given size, m, will hit it no more than $2\epsilon m$ many times. □

Note that this bound conveys the intuition that the learning task should be harder as the sample densities, α outside versus β inside the target, get closer together.

We shall conclude this section with yet another lower bound result. This last result has the advantage of showing that the task of WAT learning a class $\mathcal{C}$ gets harder as the VC dinmension of $\mathcal{C}$ grows. Its weak side is that it is an asymptotic result and that it applies only when the difference $(\beta - \alpha)$ is below some constant, γ_0.

Lemma 6: There exists a positive constant $0 < \gamma_0 < 1$ such that for every concept class $\mathcal{C}$ and any number d below the VC dimension of $\mathcal{C}$, whenever ϵ, δ and $(\beta - \alpha)$ are all below γ_0 and $(\beta + \alpha)/2 = \frac{1}{2}$, then any algorithm for (α, β)- learning of $\mathcal{C}$ with accuracy ϵ and confidence $(1 - \delta)$ requires $\Omega\left(\frac{d}{\epsilon(\beta-\alpha)^2}\right)$ many examples.

We prove this lemma by reducing the task of WAT learning a class of sets to the task of learning an associated class of probabilistic concepts, and then applying a lower bound of Simon [S93] on the sample size of p-concepts learning. For the relevant definitions of p-concept learning and the related notions of γ-shattering and the dimension $d_C(\gamma)$ we refer the reader to the papers of Kearns and Schapire [KS90] and of Simon [S93].

Proof: Given a class $\mathcal{C}$ of sets over a domain $(\mathcal{X}, \mathcal{B}, \mu)$ and a parameter $0 < \gamma < 0.5$, we define a class of distributions, $D_{\mathcal{C}} = \{D_t : t \in \mathcal{C}\}$. Each distribution D_t is defined by setting its density function (w.r.t. μ) to be $d_t(x) = 0.5 + \gamma$ for every $x \in t$ and $d_t(x) = 0.5 - \gamma$ for every $x \notin t$.

The following claims follow immediatly from the relevant definitions:

Claim 2: The γ-shattering dimension of the class of density functions, $d_{D'_{\mathcal{C}}}(\gamma)$ (where $D'_{\mathcal{C}} = \{d_t : t \in \mathcal{C}\}$), equals the VC dimension of the class $\mathcal{C}$.

Claim 3: If $h \subseteq \mathcal{X}$ is ϵ-close to some $t \in \mathcal{C}$ then the function f_h, defined by $f_h(x) = 0.5 + \gamma$ for every $x \in h$ and $f_t(x) = 0.5 - \gamma$ for every $x \notin h$, is an (ϵ, γ)-good model of probability for the density function d_t.

Claim 4: If $\mathcal{C}$ is $(0.5 - \gamma, 0.5 + \gamma)$-learnable (in the WAT sense) from m many examples with accuracy ϵ and confidence $(1 - \delta)$, then the class of density functions $D'_{\mathcal{C}}$ is learnable in the p-concept sense with an (ϵ, γ)-good model of probability from m^+ many examples, where m^+ is the number of examples needed to guarantee that at least m of them are labeled 1 by the target P-concept.

The proof of our lemma now concludes by applying the following Theorem 3.1 of Simon [S93]:

Theorem 6: [Simon] Let $\mathcal{C}$ be a p-concept class. Assume that $\delta \leq \epsilon \leq \frac{3}{160}$ and $\gamma \leq \gamma_0$ for a sufficiently small constant $0 < \gamma_0 < 1$ (not depending on $\mathcal{C}$). Then any algorithm A which learns $\mathcal{C}$ with an (ϵ, γ)-good model of probability needs $\Omega\left(\frac{d_{\mathcal{C}}(\gamma)-1}{\epsilon\gamma^2}\right)$ examples.

□

Note that all the above lemma imply similar lower bounds on the more general task of (ϵ, ρ) density level learning.

6 Acknowledgments

We thank Hans Ulrich Simon, Eli Dichterman, Oded Goldreich, Ofer Zeitouni, Neri Merhav, Freddy Bruckstein and Michael Ben-Or for sharing their expertise with us.

References

[AL88] Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

[BL93] S. Ben-David and M. Lindenbaum, 1993, "Localization vs. Identification of Semi-Algebraic Sets", *Proc. COLT 93*, pp. 327-336.

[BEHW89] Blumer, A., A. Ehrenfeucht, D. Haussler and M.K. Warmuth, 1989, "Learnability and The Vapnik-Chervonenkis Dimension", *JACM*, **36**(4), pp. 929-965.

[CEG93] Canetti, R., G. Even and O. Goldreich, 1993, "Lower Bounds for sampling Algorithms for Estimating the Average", Computer Science Technical Report No. 789, Technion - Israel Institute of Technology.

[D95] Dichterman, E., Private communication.

[DH73] Duda, R.O., and P.E. Hart, 1973, "Pattern Classification and Scene Analysis", Wiley.

[D84] Dudley, R.M., 1984, "a course on empirical processes", *Lecture Notes in Mathematics*, 1097, pp. 2-142.

[IK88] Illingworth, J., and J. Kittler, 1988 "A Survey of Hough Transform", *Computer Vision, Graphics, and Image Processing*, **44**, pp. 87-116.

[KM95] Karpinski, M., and A. Macintyre, "Polynomial Bounds for VC-dimension of Sigmoidal Neural Networks", To appear in the Proceedings of STOC '95.

[K91] Khovanskii, A.G. 1991, "Fewnomials", *Translations of Mathematical Monographs*, 88.

[KMRRSS94] Kearns, M., Y. Mansour, D. Ron, R. Rubinfeld, R.E. Schapire, and L. Sellie, 94, "On the Learnability of Discrete Distributions", *Proc. of 26th ACM STOC 94*, pp. 273-282.

[KS90] Kearns, M., and R.E. Schapire, "Efficient Distribution-Free Learning of Probabilistic Concepts" *Proc. of 31st FOCS 90*, pp. 382-392. Full version will appear in JCSS.

[Kim91] Kim, W.M., 1991, "Learning by Smoothing: a Morphological approach", *Proc. of COLT 91*, pp. 43-57.

[Nat91] Natarajan, B.K., 1991, "Probably Approximate Learning of Sets and Functions", *SIAM J. Comput.* **20**(1), pp. 328-351.

[Pap84] Papoulis, A., 1984, "Probability, Random Variables, and Stochastic Processes", 1984, McGraw-Hill.

[S93] Simon, H.U., "General Bounds on the Number of Examples Needed for Learning Probabilistic Concepts", to appear in JCSS. An extended abstarct appears in *Proc. COLT 93*, pp. 402-411.

[VC71] Vapnik, V.N. and A.Y. Chervonenkis, 1971, "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities", Theory of Probability and its applications, 16(2), pp. 264-280.

7 Appendix A: Matched filtering as a learning criterion

We have shown here that (r, η)-consistency relative to an appropriate set, is a reliable rule for drawing hypotheses from unclassified examples. It would be interesting to compare the

proposed method to more intuitive and traditional methods. The most intuitive heuristic would be to choose the concept which contains the maximal number of examples. This approach, however, will be biased towards large concepts and will completely fail when nested concepts ($c_1 \subset c_2$) are considered. An attempt to eliminate this drawback by normalizing the number of examples included in each concept by its measure, and choosing a concept that maximizes this empirical density $\hat{\mu}_(c) = \frac{|x \cap c|}{\mu(c)}$, may also fail, as concepts that are subsets of the target may have the same density or even higher. One way to eliminate this problem is to restrict the concept class to contain only concepts that have the same measure.

Such approaches are indeed efficient practical approaches in a somewhat different, yet similar, context: the detection of signals in noise, where they is called "correlation detector" or "matched filter" [Pap84]. There, one gets an input which is either a known signal corrupted with additive, stationary, zero-mean noise, or just the noise itself, and wishes to decide whether the signal is present. It turns out that in a certain sense, (when one restricts himself to linear operators), the decision procedure which will give the lowest error rate is to take the inner product of the noisy signal and a test signal, and to compare it to some threshold which depends on the noise. It is not difficult to prove (by Schwartz inequality), that for all signals with the same energy, the expected value of this inner product is maximal when the test signal is identical to the original clean signal.

In the context of randomly drawn examples, one may regard the $(\alpha = 0, \beta = 1/\mu(T))$-distribution as the clean signal, and any (α, β)-distribution, for $0 < \alpha < \beta < 1/\mu(T)$, as a noisy signal.

We can now show that maximizing the empirical density is indeed a good strategy for learning equi-measure concept classes and (α, β)-distributions are considered. It is not a good strategy however, for the more general concepts classes considered here.

Theorem 7: Let $\mathcal{C}$ be a finite VC-dimension concept class in a measure space $(\mathcal{X}, \mathcal{B}, \mu)$.

1. If, $\forall c \in \mathcal{C},\ \mu(c) = \mu_0$, then, any student which takes at least $N_{approx}(d, \frac{1}{2}\epsilon(\beta - \alpha), \frac{1}{2}\delta)$ samples, and maximizes the empirical density $\hat{\mu}(c)$ is (ϵ, δ)-successful for every (α, β)-distribution.

2. If $c_1, c_2 \in \mathcal{C}$ and $c_1 \subset c_2$, then any student which maximizes the empirical density may fail for certain (α, β)-distributions.

Proof:

The probability of a sample to fall inside any concept c, for which $\mu(T \Delta c) \geq \epsilon$ and $\mu(c) = \mu(T) = \mu_0$, is at most $(\mu_0 - \epsilon)\beta + \epsilon\alpha$. Taking the number of samples specified in the theorem, ensures with high confidence $1-\delta$, that the sample set is an $\frac{1}{2}\epsilon(\beta-\alpha)$-approximation and that all ϵ-far hypotheses have lower empirical density that the target itself.

On the other hand, consider the following (α, β)-distribution, valid when $c_1 \subset c_2$.

$$
\begin{array}{llll}
If\ x \in c_1 & then & D(x) \geq \beta_1 \mu(x) & \\
If\ x \in c_2 \setminus c_1 & then & D(x) = \beta_2 \mu(x) & (\beta_2 < \beta_1) \\
If\ x \notin c_2 & then & D(x) \leq \alpha \mu(x) &
\end{array}
$$

Clearly, even if $T = c_2$, a student which maximize the density may choose c_1 as the hypothesis. $\square$

Thus, using the "matched filter" approach, the student requires a lower sample complexity but can learn only more restricted concept classes. This drawback is usually not significant in the signal detection context, where a fixed-length data vector is often considered.

8 Appendix B: a lower bound on the tail probability

For completeness, we derive here, using standard statistical techniques, a lower bound on the number of examples required to achieve a reliable probability estimate. A similar results, derived using different tools, may be found in [CEG93].

Lemma 7: let $\{x_i\}_{i=1}^n$ be a sequence of i.i.d. binary random variables. ($Prob\{x_i = 1\} = p$). The size, n, of the sequence, required to guarantee, with confidence of $1-\delta$ or higher, that the statistics $\hat{p} = \frac{1}{n}\sum_{i=1}^n x_i$ is higher than $q < p$ is asymptotically, $LB(p,q,\delta) = \frac{p(1-p)}{(p-q)^2}\log\frac{1}{\delta}$.

Proof: Using the known results on the number of typical sequences, $\binom{n}{k} = 2^{nH(k/n)}$, the probability $Prob\{\hat{p} < q\}$ can be estimated as follows:

$$\begin{aligned}
Prob\{\hat{p} < q\} &= \sum_{k<nq} \binom{n}{k} p^k (1-p)^{n-k} \\
&\doteq \sum_{k<nq} 2^{nH(k/n)}\ 2^{k\log p + (n-k)\log(1-p)} \\
&= \sum_{k<nq} 2^{n[\frac{k}{n}(\log p - \log\frac{k}{n}) + (1-\frac{k}{n})\log(1-p) - log\frac{k}{n}]} \\
&= \sum_{k<nq} 2^{-nD(\frac{k}{n}\|p)} \\
&\geq 2^{-nD(q\|p)}, \qquad (1)
\end{aligned}$$

where $D(q\|p) = q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p}$ is the divergence between the two Bernoulli distributions, one associated with probability q and another with probability p. Requiring $Prob\{\hat{p} < q\} < \delta$, implies that

$$n \geq \frac{1}{D(q\|p)}\log\frac{1}{\delta} \qquad (2)$$

Using the convexity of the logarithmic function, we get

$$\begin{aligned}
D(q\|p) &= q\log\frac{q}{p} + (1-q)\log\frac{1-q}{1-p} \\
&< \log\left[\frac{q^2}{p} + \frac{(1-q)^2}{1-p}\right] \\
&= \log\left[1 + \frac{(q-p)^2}{p(1-p)}\right] \\
&< \frac{(q-p)^2}{p(1-p)}, \qquad (3)
\end{aligned}$$

which implies the lemma. □