Machine Learning, 18, 255–276 (1995)

Learning from a Population of Hypotheses

MICHAEL KEARNS AND H. SEBASTIAN SEUNG

mkearns@research.att.com, seung@physics.att.com

AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974

Editor: Sally A. Goldman

Abstract. We introduce a new formal model in which a learning algorithm must combine a collection of potentially poor but statistically independent hypothesis functions in order to approximate an unknown target function arbitrarily well. Our motivation includes the question of how to make optimal use of multiple independent runs of a mediocre learning algorithm, as well as settings in which the many hypotheses are obtained by a distributed population of identical learning agents.

Keywords: machine learning, computational learning theory, PAC learning, learning agents

1. Introduction

In this paper, we are concerned with the problem of combining a number of potentially poor but statistically independent hypotheses in order to obtain a significantly better approximation to an unknown target function. Our motivating scenario is a world in which a large number of learning agents each collects a small but independent sample and forms a hypothesis based on its sample. Although the data available to individual agents is limited, the entire population regarded as a single entity has collected a large number of independent examples. These examples are no longer directly available, but have been translated into many individual hypotheses, each with potentially large error. We are thus interested in learning not from random examples, but from the population's many hypotheses. The goal is to combine a number of these limited accuracy hypotheses in order to obtain a new hypothesis with arbitrarily small error.

There are two lines of prior research in computational learning theory and related fields that immediately come to mind in our setting. The first is the recent work on combining "expert" opinions in an optimal on-line fashion (see (Cesa-Bianchi, et. al, 1993) for recent results and an extensive bibliography). Briefly, in the research on experts, we assume that have access to the predictions of a panel of experts, and our goal is to make predictions with a mistake rate approaching that of the best expert. Since typically no assumptions are made regarding the sequence being predicted or the experts (for instance, the sequence may be arbitrarily time-dependent, so an expert's performance on any part of the sequence may be a poor predictor of its future performance), approaching the best expert's mistake rate is the most that can expected in such models (Cesa-Bianchi, et. al, 1993).

In contrast, in this paper we make assumptions about both the desired predictions and the "experts" (which we do not regard as being especially expert). The desired

predictions are represented by a fixed, unknown target function chosen from a restricted, known class, and each "expert" (or hypothesis) is the result of training on a small but *independent* random sample of the target function. By making these assumptions, we allow the possibility of somehow combining the independent hypotheses in a way that considerably outperforms any single hypothesis.

The second loosely related line of research is the work on boosting weak learning algorithms(Schapie, 1990; Freund, 1990; Freund, 1992), in which the goal is to combine a collection of hypotheses from a mediocre learning algorithm in order to obtain an arbitrarily accurate hypothesis. Although our goals are similar, a crucial difference is that in the boosting work, we have control over the executions of the weak learning algorithm and thus by modifying the training distribution we can force each subsequent hypothesis to have a slight prediction advantage where the previous hypotheses have failed. Here we assume no such mechanism, and each hypothesis is trained on the same fixed distribution. Indeed, it is interesting to note that natural schemes for combining hypotheses that are successful in the boosting setting, such as majority vote (Freund, 1990), often fail in our setting.

1.1. Overview of Results

We now give a summary of the paper. In Section 2, we introduce and motivate our model, which we call *population learning*. Briefly, in this model a population learner is provided with an oracle that on each call produces a function that is consistent with an independent random sample of the unknown target function. Thus, each call to the hypothesis oracle causes a new sample of m random examples to be drawn, and for a function consistent with these m examples to be returned to the population learner. The method by which the consistent function is chosen can sometimes be crucial and is a parameter of our model. For several of our results, we concentrate on the case where the returned function is chosen randomly from among all consistent hypotheses (that is, by a *Gibbs* learner). We regard m as a fixed constant over which the population learner has no control, but the population learner may draw as many hypotheses as desired in order to obtain arbitrarily small error.

In Section 3, we analyze a simple population learning problem and introduce the important and natural idea of the distribution induced on hypotheses by the hypothesis oracle. This allows us to develop some general theory for population learning in Section 4. We first introduce our central technical tool, the *separation functions*. These functions essentially quantify how the distance between two possible target functions (measured with respect to the target distribution) translates to the distance between the two corresponding induced distributions on hypotheses (measured by Kullback-Leibler divergence or variation distance). Intuitively, if this translation results in an extreme contraction of distances, then population learning is difficult, and if this translation is relatively mild, then population learning can be accomplished with a modest number of hypotheses.

With the notions of induced hypothesis distributions and separation functions in hand, we next turn to the fundamental problem of providing general upper and lower bounds on the number of hypotheses that must be drawn in order to obtain a desired level of

accuracy. This is analogous to the problem of determining upper and lower bounds on sample complexity in standard models of learning from examples.

For the upper bound, we formulate population learning as a problem of classical parametric distribution estimation of the induced distributions on hypotheses. We then invoke the powerful tools of the uniform convergence literature to analyze the maximum likelihood method for this problem, in order to obtain an upper bound which is polynomial in the inverse of the separation functions and a dimension term. We then provide a lower bound that is also polynomial in the inverse of the separation functions, thereby demonstrating that these functions give a coarse and partial characterization of the required number of hypotheses.

Section 5 gives several applications of the general theory. We analyze some simple population learning problems, including problems where the hypotheses are initial intervals of the real line, boolean conjunctions, and perceptrons. We also consider both cases where the Gibbs algorithm is used to choose consistent hypotheses, and where an arbitrary consistent hypothesis is chosen.

Section 6 mentions several areas for further research.

We wish to emphasize that although some of the methods we propose here are computationally efficient in the limited settings we consider, our primary concern in this paper is with the *statistics* of learning from a population of hypotheses, that is, with the number of independent hypotheses that are necessary and sufficient for learning in our model (whether by a computationally efficient algorithm or not). In general we have left the important problem of computational feasibility to future investigations.

2. The Population Learning Model

Imagine a world populated by a large number of initially identical *learning agents*. Each agent wanders through the world, acquiring a limited number of independent examples of an unknown target function, and then applies an internal algorithm for learning from examples to the data it has collected in order to obtain a hypothesis function. We assume that all agents use the same internal algorithm for learning from examples, so agents differ only in the data they have gathered and its subsequent effects on their hypotheses. In this paper, we wish to investigate the problem of learning not from examples, but from the *hypotheses* computed by the independent agents.

A *population learning problem* will be defined as a triple $(\mathcal{F}, D, m)$ (we will add some further components shortly). Here $\mathcal{F}$ is the class of possible $\{0,1\}$-valued target functions over the input space X, D is a probability distribution over X (or density in the case of continuous X), and $m \geq 1$ is a natural number called the *agent sample size*, which is the number of random examples seen by each agent.

We assume that $\mathcal{F}$, D and m are all known to the algorithm trying to solve the population learning. We also assume that every agent sees the same number m of random examples. In general, throughout the paper we will at any time be discussing a fixed population learning problem, so for notational brevity we will not explicitly indicate dependences on $\mathcal{F}, D$ and m except where necessary. Note also that we are studying a "distribution-specific" model of learning, in the sense that D is fixed and known.

As is typical of concept learning models, we seek algorithms that can find good approximations to an unknown *target function* $f \in \mathcal{F}$ with respect to the distribution D. However, in our model the algorithm (called a *population learning algorithm*) does not have direct access to random examples of f, but only to a large collection of hypotheses that have been independently computed using random examples of f. More precisely, for the population learning problem $(\mathcal{F}, D, m)$ a population learning algorithm is given access to the oracle $POP(f)$ that runs in unit time and behaves as follows on each call:

- Draw m inputs $S = \{x_1, \ldots, x_m\}$ randomly and independently according to D. Let S_f denote the set of inputs in S paired with the labels given by the target function $f \in \mathcal{F}$.
- Choose an element h of the *version space* $VS(S_f)$, which is the set of all functions in $\mathcal{F}$ that are consistent with the labeled sample S_f (further details of this step are discussed below).
- Return h.

Thus, we may think of each call to the oracle $POP(f)$ as returning the hypothesis of a single learning agent from a large population of agents, each member of which saw m independent random examples of f. If we make ℓ calls to this oracle, we obtain a pool $h_1, \ldots, h_\ell$ of hypotheses. Although we expect each h_i to have limited accuracy (because each h_i was obtained using only m examples), the *total* number of independent random examples that was used to train the entire pool is $\ell \cdot m$.

Despite the fact that a population learning algorithm has access only to the h_i, for sufficiently large ℓ in principle it may be possible to combine $h_1, \ldots, h_\ell$ in some manner to obtain a new hypothesis $\hat{f}$ that is considerably more accurate than any of the h_i. Indeed, as ℓ becomes large one might expect to be able to obtain $\hat{f}$ with arbitrarily small error. It is exactly this type of statement that we wish to formalize and quantify in this paper.

A crucial detail left unspecified by the given description of $POP(f)$ is *which* element of $VS(S_f)$ is returned by the oracle. The insistence that the chosen hypothesis be consistent with the examples is in fact largely inconsequential to the general theory we will develop, but is a reasonable working assumption. The method used to choose from the version space amounts to an assumption on what common algorithm for learning from examples is used by the learning agents. There are many reasonable and interesting assumptions that could be made here. In this paper we will both develop a general theory that applies regardless of what algorithm is used by the agents, and also study the details of a model in which the agents use the so-called *Gibbs* algorithm.

In the general case, we add another item A (called the *agent algorithm*) to the description of a population learning problem $(\mathcal{F}, D, m, A)$. Here A may be any randomized algorithm that takes as input a set S_f of labeled examples of some $f \in \mathcal{F}$ and outputs some $h \in VS(S_f)$. Again, as for the other items in the quadruple defining a population learning problem, we shall usually leave any dependences on A implicit for notational brevity.

Under agent algorithm A, the previously underspecified second step of the oracle $POP(f)$ is completed as follows: the $h \in VS(S_f)$ chosen for output by the oracle is simply $A(S_f)$ (the output of A when given the labeled sample S_f). It is important to note that the agent algorithm A is part of the description of a population learning problem and thus is considered to be "known" by the population learning algorithm. Thus, we allow population learning algorithms to be designed for the particular agent algorithm A in question (as well as the particular $\mathcal{F}, D$ and m).

A special case of interest occurs when the agent algorithm A is the well-studied *Gibbs algorithm*, which is known to be a near-optimal learning algorithm in terms of its expected error as a function of the number of examples m (Haussler, Kearns, & Schapire, 1994). This algorithm simply chooses h uniformly at *random* from the version space $VS(S_f)$. This models a population in which each agent learns by choosing a consistent hypothesis from $\mathcal{F}$ without bias, in the sense that given consistency with the training data, all functions are equally likely to be chosen.

A *population learning algorithm* P for a population learning problem $(\mathcal{F}, D, m, A)$ is an algorithm that for any target function $f \in \mathcal{F}$ is given access to the oracle $POP(f)$ and two inputs $0 < \epsilon, \delta \leq 1$, and eventually halts by outputting a function $\hat{f} \in \mathcal{F}$ that with probability at least $1 - \delta$ satisfies $D[f \Delta \hat{f}] \leq \epsilon$.

Given any fixed population learning problem $(\mathcal{F}, D, m, A)$, in this paper we are primarily interested in the *population size* required for learning. Thus, for a population learning problem $(\mathcal{F}, D, m, A)$ we define the function $\ell(\epsilon, \delta)$ to be the minimum over all population learning algorithms P for $(\mathcal{F}, D, m, A)$ of the maximum number of calls (over all target functions $f \in \mathcal{F}$) made by P to the oracle $POP(f)$ on inputs ϵ and δ. Note that $\ell(\epsilon, \delta)$ depends on all four parameters of the population learning problem.

Several points regarding the model bear mentioning before we embark on our investigation. First, note that we fix the population learning problem $(\mathcal{F}, D, m, A)$, and then seek an algorithm that works for all values of ϵ and δ for this problem. Thus, we think of the agent sample size m as a *constant*, and a population learning algorithm can obtain more information about the target only by drawing a larger number of hypotheses that each have this same constant amount of training.

Second, note that we assume that the oracle $POP(f)$ returns *exact descriptions* of hypotheses, as opposed to only returning "black boxes" (input-output oracles) for hypotheses. Thus, in principle a population learning algorithm may not only evaluate the sampled hypotheses, but may use the defining parameters of the sampled hypotheses in any way it sees fit. For instance, if the function class $\mathcal{F}$ is a class of neural networks of some fixed architecture, the population learning algorithm has access to the values of the weights in the hypotheses returned by $POP(f)$. Although the algorithms we propose will technically use this capability, in general we suspect that there is little additional power gained over black-box use of the hypotheses. For instance, for every specific population learning problem analyzed in Section 5, our algorithms are easily coverted to make only black-box use of hypotheses with no change in the required population size.

Finally, the population learning model could be viewed as an instance of what statisticians call *meta-analysis*, in which multiple sources of perhaps secondary data are combined to give a unified hypothesis.

3. An Illustrative Example: The High-Low Game

In this example, the domain X is the real interval $[0,1]$, and $\mathcal{F}$ is the class of all *initial intervals*. Thus, each target function is a real number $f \in [0,1]$, and the positive examples are the subinterval $[0,f]$, with the interval $(f,1]$ being the negative examples. Let D be the uniform distribution on $[0,1]$. These settings are also known as the "high-low game", since each example x of f simply indicates whether x is smaller or larger than f.

Let us examine the population learning problem $(\mathcal{F}, D, m = 1, A = \mathit{Gibbs})$. In this problem, for target f the oracle $POP(f)$ behaves as follows: a single $x \in [0,1]$ is chosen uniformly at random. If $x \leq f$ (positive example), then a random $h \in [x,1]$ is chosen uniformly and returned. If $x > f$ (negative example), then a random $h \in [0,x)$ is chosen uniformly and returned.

An important observation that applies to any population learning problem is that for any target $f \in \mathcal{F}$, the oracle $POP(f)$ induces a well-defined probability distribution q_f over $\mathcal{F}$. Thus, for any $h \in \mathcal{F}$, we let $q_f[h]$ denote the probability that h is output by the oracle $POP(f)$ (or the density of q_f at h in the continuous $\mathcal{F}$ case). Note that q_f depends crucially on the agent algorithm A. A population learning algorithm has access to random draws from q_f as its sole source of information. The function class $\mathcal{F}$ gives rise to the associated class of induced distributions $\mathcal{Q} = \{q_f : f \in \mathcal{F}\}$.

It is the analysis of the problem of learning the distribution q_f, and the relationship between this problem and approximating the target function f, that will form the backbone of our entire approach. We will shortly obtain general upper bounds on required population size by analyzing the classical maximum likelihood approach to estimating q_f. For the specific case of the high-low game, it turns out to be sufficient for the analysis to compute $\mathbf{E}_{h \in q_f}[h] = \mathbf{E}[h]$, which is the expected value of the hypotheses $h \in [0,1]$ generated by the distribution q_f. (Throughout the paper, we use the subscript $h \in q_f$ on an expectation or probability to denote that h is chosen randomly according to q_f, and $h \in S$ to denote that h is chosen uniformly from the set S.) We may write

$$\begin{aligned}
\mathbf{E}[h] &= \int_0^1 \mathbf{E}[h|x]dx \\
&= \int_0^f \mathbf{E}_{h\in[x,1]}[h]dx + \int_f^1 \mathbf{E}_{h\in[0,x]}[h]dx \\
&= \int_0^f \left(x + \frac{1-x}{2}\right)dx + \int_f^1 \frac{x}{2}dx \\
&= \frac{f}{2} + \frac{1}{4}.
\end{aligned}$$

Here we have broken the expectation into two easily analyzed parts: the first where the single example x is positive (in which case h is drawn randomly from $[x,1]$ and thus has expected value $x + (1-x)/2$), and the second where x is negative (in which case h has expected value $x/2$). This calculation immediately suggests the following population learning algorithm: draw $h_1, \ldots, h_\ell$ from the oracle $POP(f)$ and let $h_{avg} = (1/\ell)\sum_{i=1}^{\ell} h_i$; then solve $h_{avg} = \hat{f}/2 + 1/4$ for the final hypothesis $\hat{f}$. Correctness and

convergence of this procedure can be proven via Chernoff bounds, giving the following theorem.

THEOREM 1 *Let $\mathcal{F}$ be the class of initial intervals over $[0,1]$, and D the uniform distribution on $[0,1]$. Then for the population learning problem $(\mathcal{F}, D, m = 1, A = Gibbs)$, $\ell(\epsilon, \delta) = O(1/\epsilon^2 \log 1/\delta)$.*

This bound compares favorably with the $\Theta(1/\epsilon \log 1/\delta)$ sample size that is required for learning $\mathcal{F}$ from the random examples themselves (rather than the hypotheses) with respect to the same distribution. Thus, even when each agent has seen only a *single* example of the target function, a relatively small sampling of hypotheses can be combined to find a much more accurate hypothesis. Note that our algorithm for this simple problem is also computationally efficient.

3.1. *Remarks on the High-Low Game*

Several other points regarding this simple example bear mentioning. First of all, the choice of the agent algorithm A can sometimes have great effect: let A be the consistent algorithm that for a positive example x chooses the hypothesis $h = x + \gamma$, and for a negative example x chooses the hypothesis $h = x - \gamma$ (for some small $\gamma \geq 0$). Then it easy to see that as γ approaches 0, q_f approaches the uniform distribution on $[0,1]$ *independent of* f. This demonstrates that for the high-low game with $m = 1$, it is not possible to obtain a single finite upper bound on $\ell(\epsilon, \delta)$ that holds simultaneously for all choices of A, and we must analyze the required population size for different agent algorithms on a case-by-case basis.

Second, however, the effects of the particular agent algorithm A can sometimes be overcome by a sufficiently large agent sample size m. Thus, we will later show that in the $m = 2$ case of the high-low game, we *can* upper bound $\ell(\epsilon, \delta)$ by a polynomial in $1/\epsilon$ and $1/\delta$ simultaneously for all agent algorithms. In general, we expect larger agent sample size to make population learning easier (or at least not more difficult). However, there are some subtleties involved with this intuition that we discuss later.

Finally, the high-low game is a simple problem for which several natural and naive approaches to population learning fail. For instance, it is tempting to conjecture that a general approach to population learning is majority voting: sample hypotheses $h_1, \ldots, h_\ell$ and let $\hat{f}$ be the majority vote of these hypotheses. In the high-low game, it is easy to see that this scheme is equivalent to choosing $\hat{f}$ to be the median of $h_1, \ldots, h_\ell$. However, when the target function $f = 0$, it can be shown that the median converges to the value 0.1865... as $\ell \to \infty$, and thus will not achieve arbitrarily small error even given an infinite population size.

4. Development of the General Theory

Throughout this section, we assume a fixed population learning problem $(\mathcal{F}, D, m, A)$. Thus far, we have observed that each target function $f \in \mathcal{F}$ gives rise to an induced

distribution $q_f \in \mathcal{Q}$ over $\mathcal{F}$ which is exactly the distribution sampled by the oracle $POP(f)$; note that each q_f depends on all four parameters of the population learning problem in addition to f. One natural approach to population learning would be to learn an approximation $\hat{q}$ to q_f, and somehow use $\hat{q}$ to find a good approximation to f itself. Our approach to the high-low game can be viewed as a special case of this approach, where all that was needed was an approximation to the mean of q_f.

In order to formalize this approach, we must specify what is meant by learning the distribution q_f (or more precisely, what measure is used to evaluate a hypothesis distribution), and then study quantitatively how the problem of learning the distribution q_f relates to the original problem of learning the target function f.

We will find it convenient to consider two different standard measures for the distance between two probability distributions. The first is the *Kullback-Leibler divergence* (which is not a metric, since it lacks symmetry):

$$KL(q_{f_1}||q_{f_2}) = \sum_{h \in \mathcal{F}} q_{f_1}[h] \log \frac{q_{f_1}[h]}{q_{f_2}[h]}.$$

The second is the *variation distance*:

$$V(q_{f_1}, q_{f_2}) = \sup_{\mathcal{F}' \subseteq \mathcal{F}} |q_{f_1}[\mathcal{F}'] - q_{f_2}[\mathcal{F}']|.$$

Both measures have analogues for densities in the continuous case; in developing our general theory, however, we shall restrict ourselves to the case of distributions for simplicity. We will use the following theorem due to Kullback (1967):

THEOREM 2 *For any distributions* q_{f_1}, q_{f_2}

$$KL(q_{f_1}||q_{f_2}) \geq V^2(q_{f_1}, q_{f_2}).$$

4.1. *The Separation Functions*

Having defined these two closeness measures for probability distributions, we now introduce their associated *separation functions*. This is our most important definition, and is motivated as follows: suppose that in a population learning problem, two potential target functions $f_1, f_2 \in \mathcal{F}$ have disagreement $D[f_1 \Delta f_2] = \epsilon$. If we had access to random examples of the target function, we could distinguish between f_1 being the target and f_2 being the target in $\mathcal{O}(1/\epsilon)$ examples.

In population learning, however, all we have access to is either q_{f_1} or q_{f_2}. If despite the ϵ separation between f_1 and f_2, the separation between q_{f_1} and q_{f_2} is much smaller than ϵ, then we may require a very large population size to achieve error ϵ. On the other hand, if the ϵ separation between f_1 and f_2 implies a "significant" (say ϵ^2) separation between q_{f_1} and q_{f_2}, then a modest population size may suffice. We thus define the separation functions as:

$$\sigma_{KL}(\epsilon, m) = \min_{f_1, f_2 \in \mathcal{F}, D[f_1 \Delta f_2] \geq \epsilon} KL(q_{f_1} || q_{f_2})$$

and

$$\sigma_V(\epsilon, m) = \min_{f_1, f_2 \in \mathcal{F}, D[f_1 \Delta f_2] \geq \epsilon} V(q_{f_1}, q_{f_2}).$$

(In cases where the minimum does not exist, we instead take the infimum.) Here we are violating our convention of leaving dependence on the agent sample size m implicit for reasons we shall discuss shortly in Section 4.2. Both separation functions take ϵ as an argument, and find the *closest* (with respect to either Kullback-Leibler divergence or variation distance) that two ϵ-separated functions in $\mathcal{F}$ (with respect to D) can become in the space $\mathcal{Q}$ of induced distributions. Note that by Theorem 2 we have $\sigma_{KL}(\epsilon, m) \geq (\sigma_V(\epsilon, m))^2$ always.

Shortly we will provide evidence for the significance of the separation functions by showing that they provide a rough characterization of the population size required for any population learning problem. Specifically, we give upper and lower bounds on the population size $\ell(\epsilon, \delta)$ that are polynomial expressions in $1/\sigma_{KL}(\epsilon, m)$ and $1/\sigma_V(\epsilon, m)$ (as well as $1/\delta$ and various complexity measures of the population learning problem). We first engage in a brief discussion of the dependence of the separation functions on the agent sample size m.

4.2. *The Role of Agent Sample Size*

Let us briefly digress from the main development in order to discuss a primary but unfortunately unfulfilled goal of our investigation, and to clear the air of any confusion that this failure may cause. As we have indicated, a "nice" separation function would have behavior such as $\sigma_{KL}(\epsilon, m) \geq \epsilon^2$, so that large distances in the metric induced on $\mathcal{F}$ by D would translate to large distances (either Kullback-Leibler divergence or variation distance) in $\mathcal{Q}$. We will soon see that such nice behavior leads to relatively modest upper bounds on the required population size.

In the population learning model, we essentially regard m as a fixed constant, representing the limited amount of training received by each learning agent in the population. In particular, we do *not* allow m to increase according to the desired error bound ϵ given to the population learning algorithm — m is independent of ϵ, and all the population learning algorithm can do to achieve smaller and smaller ϵ is to take more and more hypotheses of this fixed sample size m. Thus, an important question for us is how small m can be while the separation functions still have nice behavior.

More precisely, note that in general we expect that as m increases, each induced distribution q_f (which of course implicitly depends on m) becomes more peaked around f. For this reason, we expect that as m increases, $KL(q_{f_1} || q_{f_2})$ and $V(q_{f_1}, q_{f_2})$ become larger for any two functions f_1, f_2, and thus $\sigma_{KL}(\epsilon, m)$ and $\sigma_V(\epsilon, m)$ should also increase with m. While this much *seems* clear, the challenging problem is to obtain conditions on m that are independent of ϵ but that guarantee that $\sigma_{KL}(\epsilon, m)$ and $\sigma_V(\epsilon, m)$ are polynomially large in ϵ.

To see the difficulty, let us lower bound $\sigma_{KL}(\epsilon, m)$ in terms of ϵ and m using some standard methods from uniform convergence analysis and see why they are insufficient for our purposes. Suppose we consider two functions $f_1, f_2 \in \mathcal{F}$, let $\epsilon = D[f_1 \Delta f_2]$, and let m be the fixed agent sample size. For any numbers $0 \leq r, s \leq 1$ let us define $KL(r||s) = r\log(r/s) + (1-r)\log((1-r)/(1-s))$; it is easy to show that this is lower bounded by $\max\{r\log 1/s - 1, (1-r)\log 1/(1-s) - 1\}$.

Now it is also true (Kullback, 1967) that for any $\mathcal{F}' \subseteq \mathcal{F}$,

$$KL(q_{f_1}||q_{f_2}) \geq KL(q_{f_1}[\mathcal{F}']||q_{f_2}[\mathcal{F}']). \tag{1}$$

Thus to lower bound $KL(q_{f_1}||q_{f_2})$ let us choose $\mathcal{F}'$ to be the $\epsilon/2$-ball around f_2 in $\mathcal{F}$ with respect to D, that is

$$\mathcal{F}' = \{f \in \mathcal{F} : D[f_2 \Delta f] \leq \epsilon/2\}.$$

Now using uniform convergence methods (Vapnik, 1982; Haussler, 1992) one can show

$$q_{f_2}[\mathcal{F}'] \geq 1 - c\frac{(2m)^d}{d!}e^{-\alpha\epsilon^2 m}$$

and

$$q_{f_1}[\mathcal{F}'] \leq c\frac{(2m)^d}{d!}e^{-\alpha\epsilon^2 m}$$

for constants $c, \alpha > 0$, where d is the Vapnik-Chervonenkis dimension of $\mathcal{F}$. Thus using Equation (1) and the lower bound on $KL(r||s)$ we obtain

$$\begin{aligned} KL(q_{f_1}||q_{f_2}) &\geq \frac{1}{2}\log\frac{1}{c\frac{(2m)^d}{d!}e^{-\alpha\epsilon^2 m}} - 1 \\ &\geq c_1\epsilon^2 m - c_2 d\log m + c_3\log d! + c_4 \end{aligned}$$

for constants $c_1, c_2, c_3, c_4 > 0$. This first term of the lower bound has the desired "nice" behavior: if two functions are at a distance ϵ, then their induced distributions have Kullback-Leibler divergence $\Omega(\epsilon^2)$. Unfortunately, despite this machinery, the lower bound is negative until $m = \Omega(1/\epsilon^2)$, a condition that is unacceptable for the reasons outlined above.

In fact, it is possible to argue that the desired condition on m needed to enforce niceness of the separation functions cannot be expressed solely in terms of a parameter of the function class $\mathcal{F}$ such as the Vapnik-Chervonenkis dimension. It appears that the best we could hope for is a statement of the form: provided $m \geq F(\mathcal{F}, D)$, we have $\sigma_V(\epsilon, m) \geq \epsilon^2$ (or some similarly large function of ϵ), for some function F of the function class $\mathcal{F}$ and distribution D. We have been unable to obtain such a result so far.

In any case, our belief that the separation functions may be sensitive functions of m, combined with our inability to quantify this sensitivity, prompts us to explicitly indicate the dependence on m for these functions.

4.3. *A General Upper Bound on Population Size*

An important observation regarding the separation functions is given in the following lemma, whose proof is immediate from the definitions of $\sigma_{KL}(\epsilon, m)$ and $\sigma_V(\epsilon, m)$.

LEMMA 1 *For any* $f, \hat{f} \in \mathcal{F}$, *if* $KL(q_f || q_{\hat{f}}) < \sigma_{KL}(\epsilon, m)$ *or* $V(q_f, q_{\hat{f}}) < \sigma_V(\epsilon, m)$ *then* $D[f \Delta \hat{f}] < \epsilon$.

Given the machinery we have developed thus far, we can now recast population learning as a problem in parametric distribution estimation. The population learner receives ℓ hypotheses $h_1, \ldots, h_\ell$ drawn independently at random from a distribution. The learner knows that this distribution is a member of the class $\mathcal{Q}$, which is parametrized by $\mathcal{F}$. We study the case where the learner uses the method of maximum likelihood estimation, and thus outputs a hypothesis $\hat{f}$ that is a maximum of $\prod_{i=1}^{\ell} q_{f'}[h_i]$ with respect to f'. This method treats $f' \in \mathcal{F}$ as an *abstract parameter* that does nothing more than parametrize the distributions $q_{f'} \in \mathcal{Q}$. This method may be of more theoretical than practical relevance, since the likelihoods $q_f[h]$ are generally difficult to compute. Nevertheless, the bounds on the population size required by maximum likelihood are a useful first step towards bounds for more practical learning algorithms.

The classical analysis of the error of the maximum likelihood method, involving the Fisher information, requires that the distribution class $\mathcal{Q}$ be a smooth function of continuous, real-valued parameters. As will be illustrated by specific examples in Section 5, $\mathcal{F}$ (and hence $\mathcal{Q}$) often admits no continuous parametrization. Furthermore, even in the case of a continuous parametrization, the likelihood can be nondifferentiable in its parameters, as noted by Amari(Amari, Fujita, & Shinomoto, 1992). Hence classical statistics is not typically applicable to the learning problems of interest here.

Instead we proceed by invoking uniform convergence theorems (Haussler, 1992; Pollard, 1984; Dudley, 1978) to bound fluctuations in empirical log-loss. These theorems are relevant because maximizing the likelihood is equivalent to minimizing the empirical log-loss, which is $-1/\ell \sum_{i=1}^{\ell} \log q_{f'}[h_i]$. Hence maximum likelihood is but a specific case of the general class of empirical loss minimization algorithms. Combined with Lemma 1, which relates log-loss in $\mathcal{Q}$ to loss in the parameter space $\mathcal{F}$, the uniform convergence bounds lead to the following upper bound on population size, whose proof is omitted due to space considerations, but is a fairly straightforward application of the main theorem of Haussler(1992).

THEOREM 3 *Let* $(\mathcal{F}, D, m, A)$ *be any population learning problem. Then*

$$\ell(\epsilon, \delta) = \mathcal{O}\left(\frac{dim(\mathcal{Q}) \cdot M}{(\sigma_{KL}(\epsilon, m))^2} \log \frac{M}{\sigma_{KL}(\epsilon, m)} + \log \frac{1}{\delta}\right)$$

and

$$\ell(\epsilon, \delta) = \mathcal{O}\left(\frac{dim(\mathcal{Q}) \cdot M}{(\sigma_V(\epsilon, m))^4} \log \frac{M}{\sigma_V(\epsilon, m)} + \log \frac{1}{\delta}\right).$$

Here $dim(\mathcal{Q})$ is the combinatorial dimension (Haussler, 1992) of the distribution class $\mathcal{Q}$, and M is a bound on the empirical log-loss of any distribution in $\mathcal{Q}$.

Let us take a moment to absorb this result. First of all, the combinatorial dimension $dim(\mathcal{Q})$ is a generalization of the Vapnik-Chervonenkis (VC) dimension (Vapnik & Chervonenkis, 1971) and can be considered a standard and natural notion of the "complexity" of the population learning problem. In the finite $\mathcal{F}$ case, $dim(\mathcal{Q}) \leq \log|\mathcal{F}|$. We refer the interested reader to Haussler(1992) for details. Secondly, although the appearance of the bound M in the population size upper bound might initially seem worrisome (since we have no a priori reason to assume a finite bound on $-\log q_f[h]$ for all $f, h \in \mathcal{F}$), this is often a technicality: we can typically get around any difficulty using quite general "clamping" techniques that choose a hypothesis from a restricted subclass that excludes degenerate distributions with large loss.

The bounds in Theorem 3 depend on ϵ and m through the separation functions. Although it seems intuitively clear that the separation functions should tend to increase with m and decrease with ϵ, we have not succeeded in characterizing this dependence rigorously, and it appears that uniform convergence theory may be too coarse a tool for this task (see the extensive discussion of this issue in Section 4.2). This technical difficulty is related to the difficulty of performing the quenched average in statistical mechanical analyses of learning (Seung, Sompolinsky, & Tishby, 1992). In the absence of general bounds, we must settle for calculation of the separation functions for some specific learning problems, to be done in Section 5.

A more positive statement about Theorem 3 is that the dependence of $\ell(\epsilon, \delta)$ on ϵ is captured in the polynomial dependence on $1/\sigma_{KL}(\epsilon, m)$ and $1/\sigma_V(\epsilon, m)$. This demonstrates the importance of the separation functions: good *lower bounds* on the separation functions lead to good *upper bounds* on the required population size. If we can prove, for instance, that $\sigma_{KL}(\epsilon, m)$ is bounded below by ϵ^2, then we have shown that $\ell(\epsilon, \delta)$ has an $\mathcal{O}(1/\epsilon^4)$ dependence on ϵ. If, on the other hand, $\sigma_{KL}(\epsilon, m)$ grows like ϵ^n where n is a complexity measure such as the Vapnik-Chervonenkis dimension, we face the possibility of exponentially large population size. Indeed, in the following subsection we show that this possibility can in fact be realized, and complete our rough characterization of $\ell(\epsilon, \delta)$ by providing a lower bound expressed in terms of the separation functions.

4.4. A General Lower Bound on Population Size

THEOREM 4 *Let $(\mathcal{F}, D, m, A)$ be any population learning problem. Then*

$$\ell(\epsilon, \delta) = \Omega\left(\frac{1}{\sqrt{\sigma_{KL}(2\epsilon, m)}}\right) \quad \text{and} \quad \ell(\epsilon, \delta) = \Omega\left(\frac{1}{\sigma_V(2\epsilon, m)}\right).$$

Proof: The proof is most easily done for the variation distance; the Kullback-Leibler lower bound then follows from Theorem 2. Thus let $\epsilon' = 2\epsilon$, and let $f_1, f_2 \in \mathcal{F}$ be such that $D[f_1 \Delta f_2] \geq \epsilon'$ and $V(q_{f_1}, q_{f_2}) = \sigma_V(\epsilon', m)$. Such functions must exist by the definition of $\sigma_V(\epsilon', m)$. Let P be a population learning algorithm requiring at most

ℓ calls to the oracle $POP(f)$ to obtain error smaller than ϵ (for some small constant δ) for any $f \in \mathcal{F}$.

To prove the lower bound, we will choose the target function randomly between f_1 or f_2, and we may assume without loss of generality that under these conditions, P outputs either f_1 or f_2. Let us define two complementary sets of ℓ-tuples of functions in $\mathcal{F}$: $\mathcal{T}_{f_1} = \{T \subseteq \mathcal{F}^\ell : P(T) = f_1\}$ and $\mathcal{T}_{f_2} = \{T \subseteq \mathcal{F}^\ell : P(T) = f_2\}$. Here $P(T) \in \{f_1, f_2\}$ is the output of algorithm P when the sequence $T = (h_1, \ldots, h_\ell)$ is returned by the oracle. We assume that P is deterministic; the same proof holds with only minor modification if P is randomized. Thus, $\mathcal{T}_{f_1}$ is the set of all sequences of ℓ functions causing P to output f_1, and similarly for $\mathcal{T}_{f_2}$.

We now analyze the probability (over the random choice of f_1 or f_2 as the target function f, and the subsequent random choices of $POP(f)$ from q_f) that algorithm P outputs the wrong function; notice that if this event occurs, the error of P's hypothesis is at least ϵ'. We may write

$$\begin{aligned}\mathbf{Pr}_{f\in\{f_1,f_2\},T\in q_f^\ell}[P(T) \neq f] &= \frac{1}{2}q_{f_1}^\ell[\mathcal{T}_{f_2}] + \frac{1}{2}q_{f_2}^\ell[\mathcal{T}_{f_1}] \\ &= \frac{1}{2} + \frac{1}{2}(q_{f_2}^\ell[\mathcal{T}_{f_1}] - q_{f_1}^\ell[\mathcal{T}_{f_1}]).\end{aligned}$$

Here we have used the equality $q_{f_1}^\ell[\mathcal{T}_{f_2}] = 1 - q_{f_1}^\ell[\mathcal{T}_{f_1}]$. Now

$$\begin{aligned}\left|q_{f_2}^\ell[\mathcal{T}_{f_1}] - q_{f_1}^\ell[\mathcal{T}_{f_1}]\right| &\leq \sum_{(h_1,\ldots,h_\ell)\in\mathcal{T}_{f_1}} \left|q_{f_2}^\ell[(h_1,\ldots,h_\ell)] - q_{f_1}^\ell[(h_1,\ldots,h_\ell)]\right| \\ &\leq \sum_{(h_1,\ldots,h_\ell)\in\mathcal{F}^\ell} \left|q_{f_2}^\ell[(h_1,\ldots,h_\ell)] - q_{f_1}^\ell[(h_1,\ldots,h_\ell)]\right| \\ &\leq \sum_{(h_1,\ldots,h_\ell)\in\mathcal{F}^\ell} \Big|q_{f_2}^{\ell-1}[(h_1,\ldots,h_{\ell-1})]q_{f_2}[h_\ell] \\ &\qquad - q_{f_1}^{\ell-1}[(h_1,\ldots,h_{\ell-1})]q_{f_1}[h_\ell]\Big|\end{aligned}$$

Now it is not hard to show that for any $A, A', B, B' \leq 1$

$$|AA' - BB'| \leq |A' - B'| + |A - B|.$$

Applying this to the above equation gives

$$\begin{aligned}\left|q_{f_2}^\ell[\mathcal{T}_{f_1}] - q_{f_1}^\ell[\mathcal{T}_{f_1}]\right| &\leq \sum_{h_\ell\in\mathcal{F}} |q_{f_2}[h_\ell] - q_{f_1}[h_\ell]| \\ &\quad + \sum_{(h_1,\ldots,h_{\ell-1})\in\mathcal{F}^\ell} \Big|q_{f_2}^{\ell-1}[(h_1,\ldots,h_{\ell-1})] \\ &\qquad - q_{f_1}^{\ell-1}[(h_1,\ldots,h_{\ell-1})]\Big|\end{aligned}$$

Now the first term in this final expression is bounded above by $\sigma_V(\epsilon', m)$, and so by induction the sum of the two terms is bounded above by $\ell \cdot \sigma_V(\epsilon', m)$. Thus

$$\mathbf{Pr}_{f\in\{f_1,f_2\},T\in q_f^\ell}[P(T)\neq f]\geq 1/2-(1/2)\ell\cdot\sigma_V(\epsilon',m).$$

The expected error of P is thus $\frac{1}{2}(1-\ell\cdot\sigma_V(\epsilon',m))\epsilon'$. Thus to obtain expected error smaller than $\epsilon=\epsilon'/2$ requires $\ell=\Omega(1/\sigma_V(2\epsilon,m))$, as desired. □

Note that we suspect the existence of stronger lower bounds, since Theorem 4 lower bounds only the dependence on the separation functions. It seems plausible that a lower bound also incorporating $dim(\mathcal{Q})$ is the right answer, but the given bound is sufficient for an initial characterization of population size.

Let us review where we are. At this point we have shown that the population size is roughly characterized by $\sigma_V(\epsilon,m)$ or $\sigma_{KL}(\epsilon,m)$ and the dimension term $dim(\mathcal{Q})$. A natural question to pose is how different are the given bounds from the usual bounds on the number of random examples required for learning from examples? The answer to this lies in how dramatically the separation functions may contract distances. For instance, if we could somehow prove that for any population learning problem we have $\sigma_V(\epsilon,m)\geq\epsilon^2$ then we would have shown (at least for finite classes, where $dim(\mathcal{Q})$ is bounded by $\log|\mathcal{F}|$) that the population size required for learning is always polynomially bounded by the number of random examples required for learning.

Unfortunately, and not surprisingly, the answer is not so simple in general, as the separation functions can greatly contract distances. For instance, one can show that for $\mathcal{F}$ the class of all parity functions over n boolean variables, D the uniform distribution over $\{0,1\}^n$, and for small values of m, even when the agent algorithm A is the Gibbs algorithm we have $\sigma_V(1/2,m)\leq 1/2^n$ (in this problem, $\epsilon=1/2$ is the only relevant value since every pair of target functions disagree on $1/2$ the inputs). Theorem 4 immediately implies an exponential lower bound on the population size for this problem, whereas it is well-known that $\mathcal{O}(n)$ random examples suffice for learning from examples. Thus, given a population learning problem, in general we must expect to make a specific argument for polynomial population size.

In the Section 5, we make such arguments for several population learning problems by lower bounding a separation function. In doing so, we illustrate a case where it is possible to analyze the effects of increasing the agent sample size m, and a case where we can prove small population sizes regardless of the agent algorithm A.

4.5. More General Learning Models

It is worth noting that all of the theory we have developed in this section for the population learning model can actually be applied to a much more general setting of *learning from secondary data*. The only properties of the population learning model that we have used in this section are:

- The existence of a *primary* metric space Z. In the population learning model, the primary space was $Z=\mathcal{F}$ and the metric was simply that induced by the distribution D.

- The existence for each $z \in Z$ of an induced distribution q_z over some *secondary* abstract data space Y. In the population learning model, for $f \in \mathcal{F}$, q_f happens to be over $Y = \mathcal{F}$, and is defined by $POP(f)$.

Thus in general, we could study the problem of learning a point close to a target point $z \in Z$ when given access only to q_z. The separation functions can be defined, and both our upper and lower bounds will apply to this more general setting.

5. Applications of the General Theory

We now give polynomial upper bounds on the population size required for several population learning problems of interest. The general approach is to lower bound a separation function and then apply Theorem 3. It should be noted that since Theorem 3 is obtained by Haussler(1992) in an extremely general setting, we suspect the existence of considerably better upper bounds than those we provide here; for now, however, we restrict our efforts towards proving polynomial bounds, leaving improvement of the polynomial degree for future research.

5.1. *The High-Low Game with Any Agent Algorithm*

Recall that in Section 3, we argued that in the high-low game with agent sample size $m = 1$, it was impossible to obtain an upper bound on population size that held simultaneously for all consistent agent algorithms. In the following theorem, we show that with $m = 2$, we can obtain such a uniform bound. We include a proof sketch that is illustrative of the type of reasoning used to prove such bounds.

THEOREM 5 *Let $\mathcal{F}$ be the class of initial intervals over $[0,1]$, and D the uniform distribution on $[0,1]$. Then for any consistent agent algorithm A, the population learning problem $(\mathcal{F}, D, m = 2, A)$ satisfies $\ell(\epsilon, \delta) = \mathcal{O}(1/\epsilon^8 \log 1/\epsilon + \log 1/\delta)$.*

Proof: We demonstrate that the separation function for the variation distance obeys $\sigma_V(\epsilon, 2) = \Omega(\epsilon^2)$; the stated upper bound on $\ell(\epsilon, \delta)$ can then be obtained as outlined in Section 4.3 and Theorem 3.

Let $f \in [0,1]$ be a potential target function. Recall that in the population learning problem $(\mathcal{F}, D, m = 2, A)$, $POP(f)$ draws two points uniformly from $[0,1]$, labels them according to f, and applies the consistent agent algorithm A to the resulting sample to obtain the returned hypothesis $h \in [0,1]$. Without loss of generality, we will use x_L to denote the smaller of the two chosen sample points, and x_R to denote the larger.

To prove that $\sigma_V(\epsilon, 2) = \Omega(\epsilon^2)$ it suffices to show that for any ϵ and any target functions $f_1, f_2 \in [0,1]$ such that $D[f_1 \Delta f_2] \geq \epsilon$, $V(q_{f_1}, q_{f_2}) = \Omega(\epsilon^2)$. For $S, S_L, S_R \subseteq [0,1]$, let us use $q_f[S | x_L \in S_L, x_R \in S_R]$ to denote the probability that q_f generates a hypothesis h falling in S given that in the two-point sample, x_L fell in S_L and x_R fell in S_R.

For f_1, f_2 satisfying $D[f_1 \Delta f_2] = \epsilon$ (let us assume without loss of generality that $f_1 \leq f_2 = f_1 + \epsilon$), we first have that for any $S \subseteq [0,1]$,

$$q_{f_1}[S|x_L, x_R \notin f_1 \Delta f_2] = q_{f_2}[S|x_L, x_R \notin f_1 \Delta f_2].$$

This is because the behavior of $POP(f)$ depends only on the labeled sample, and not directly on the target function, so as long as both f_1 and f_2 give the same labeling to the sample the conditional distribution of hypotheses is identical regardless of which function is the target.

Now let z be the midpoint between f_1 and f_2, so $z = (f_1 + f_2)/2 = f_1 + \epsilon/2$. It is easy to see that

$$q_{f_1}[[0, z]|x_L \in [f_1, z], x_R \in [z, f_2]] = 1$$

and

$$q_{f_2}[[z, 1]|x_L \in [f_1, z], x_R \in [z, f_2]] = 1.$$

Furthermore, the probability that $x_L \in [f_1, z]$ and $x_R \in [z, f_2]$ is $\epsilon^2/4$. Thus, if we restrict our attention only to the conditional cases of x_L and x_R discussed so far, we have found two regions on which q_{f_1} and q_{f_2} differ by $\Theta(\epsilon^2)$: that is, q_{f_1} is $\epsilon^2/4$ more likely than q_{f_2} to generate a hypothesis in $[0, z]$ and q_{f_2} is $\epsilon^2/4$ more likely than q_{f_1} to generate a hypothesis in $[z, 1]$. It is fairly straightforward to show that the remaining cases of x_L and x_R do not alter this difference, thus giving $q_{f_1}[[0, z]] = q_{f_2}[[0, z]] + \epsilon^2/4$ and $q_{f_2}[[z, 1]] = q_{f_1}[[z, 1]] + \epsilon^2/4$. Either of these suffice to show $V(q_{f_1}, q_{f_2}) \geq \epsilon^2/4$, as desired. □

Better upper bounds for this problem may be possible by direct analysis of the Kullback-Leibler separation function. The proof of Theorem 5 also provides a case where it reasonably straightforward to analyze the beneficial effects of increased agent sample size m. In the proof, we lower bounded $\sigma_V(\epsilon, 2)$ by the probability we drew a sample x_L, x_R such that $x_L \in [f_1, z]$ and $x_R \in [z, f_2]$. The arguments given hold for any m, but now the probability that we draw a set S of m points from D such that there exists $x_L, x_R \in S$ satisfying $x_L \in [f_1, z]$ and $x_R \in [z, f_2]$ can be lower bounded by $1 - 2(1 - \epsilon/2)^m \approx 1 - e^{\alpha\epsilon m}$ for some constant α. Thus in the high-low game, for any consistent A and any m we have $\sigma_V(\epsilon, m) \geq 1 - e^{\alpha\epsilon m}$, giving considerably improved population size upper bounds for large m via Theorem 3. This is a rare case where we can precisely quantify the effects of increasing m, as opposed to the general situation discussed in Section 4.2.

5.2. *Conjunctions with Gibbs and Any Distribution*

The high-low game is a one-dimensional learning problem, so we have not examined the potential effects of high dimension on the separation functions (other than for the class of parity functions with small agent sample size in Section 4.4, where we saw that the contraction of distance was exponentially small in the dimension). We now examine some population learning problems in high-dimensional spaces and find that often the effects are rather modest, and still permit polynomial population size. We begin with the well-studied class of boolean conjunctions, for which we can actually obtain a bound

that holds simultaneously for *any* fixed distribution. Here we restrict our attention to the $m = 1$ case.

THEOREM 6 *Let $\mathcal{F}_n$ be the class of all monotone conjunctions over n boolean variables, and let D be any distribution over $\{0,1\}^n$. Then for the population learning problem $(\mathcal{F}, D, m = 1, A = Gibbs)$ we have $\ell(\epsilon, \delta) = \mathcal{O}(n^5/\epsilon^4 \log n/\epsilon + \log 1/\delta)$.*

Proof: We proceed as usual by demonstrating an appropriate lower bound on $\sigma_V(\epsilon, 1)$. Thus, let f_1 and f_2 be any monotone conjunctions, and let $D[f_1 \Delta f_2] = \epsilon$. Let T_1 be the set of variables appearing in f_1 but not in f_2 and let T_2 be the set of variables appearing in f_2 but not in f_1. Let ϵ_1 be the probability with respect to D that an x is drawn satisfying $f_1(x) = 1, f_2(x) = 0$ and let ϵ_2 be the probability that $f_1(x) = 0, f_2(x) = 1$; note that $\epsilon_1 + \epsilon_2 = \epsilon$.

First let us describe the behavior of the Gibbs algorithm in this context. Given a positively labeled x, a random consistent hypothesis is obtained by randomly choosing a subset of the variables set to 1 in x, and forming the conjunction of this subset. Given a negatively labeled x, a random consistent hypothesis is obtained by choosing a random subset of all the variables, then rejecting the trial unless the chosen subset contains at least one variable set to 0 in x.

To demonstrate a difference between q_{f_1} and q_{f_2} we may restrict our attention to points where f_1 and f_2 disagree. Thus, suppose that f_1 is the target and we draw x such that $f_1(x) = 1, f_2(x) = 0$ (which happens with probability ϵ_1). Then the expected number of variables in T_1 chosen by the Gibbs algorithm is $|T_1|/2$ (since all these variables must be set to 1 in x), and the expected number of variables in T_2 chosen is at most $(|T_2| - 1)/2$ (since at least one variable in T_2 is set to 0). On the other hand, if f_1 is the target and we draw x such that $f_1(x) = 0, f_2(x) = 1$ (which happens with probability ϵ_2), then the expected number of variables in T_1 chosen is at least $|T_1|/2$ (the fact that T_1 contains at least one variable set to 0 can only introduce a bias towards larger subsets), and the expected number of variables in T_2 chosen is $|T_2|/2$ (since all variables in T_2 must be set to 1 in x). If for any monotone conjunction h, we let $\chi_1(h)$ denote number of variables in h appearing in T_1, these facts are easily combined to give

$$\begin{aligned}
&\mathbf{E}_{h \in q_{f_1}}[\chi_1(h)] - \mathbf{E}_{h \in q_{f_2}}[\chi_1(h)] \\
&\geq \epsilon_1 \left(\frac{|T_1|}{2} \right) + \epsilon_2 \left(\frac{|T_1|}{2} \right) - \epsilon_1 \left(\frac{|T_1|}{2} \right) - \epsilon_2 \left(\frac{|T_1| - 1}{2} \right) \\
&= \frac{\epsilon_2}{2}.
\end{aligned}$$

By symmetric arguments, if $\chi_2(h)$ denotes the number of variables in h appearing in T_2, we have

$$\mathbf{E}_{h \in q_{f_2}}[\chi_2(h)] - \mathbf{E}_{h \in q_{f_1}}[\chi_2(h)] \geq \frac{\epsilon_1}{2}.$$

Since either $\epsilon_1 \geq \epsilon/2$ or $\epsilon_2 \geq \epsilon/2$, we may assume without loss of generality that $\mathbf{E}_{h \in q_{f_1}}[\chi_1(h)] - \mathbf{E}_{h \in q_{f_2}}[\chi_1(h)] \geq \epsilon/4$. Now

$$\mathbf{E}_{h \in q_{f_1}}[\chi_1(h)] - \mathbf{E}_{h \in q_{f_2}}[\chi_1(h)] = \sum_h (q_{f_1}[h] - q_{f_2}[h]) \chi_1(h)$$

$$\begin{aligned} &\leq \sum_h |q_{f_1}[h] - q_{f_2}[h]|\chi_1(h) \\ &\leq n\sum_h |q_{f_1}[h] - q_{f_2}[h]| \\ &\leq 2nV(q_{f_1}, q_{f_2}) \end{aligned}$$

where we have used the fact that $\chi_1(h) \leq n$ always. Thus we have $V(q_{f_1}, q_{f_2}) \geq \epsilon/8n$ or $\sigma_V(\epsilon, 1) \geq \epsilon/8n$. Application of Theorem 3 then yields the stated bound on $\ell(\epsilon, \delta)$. □

5.3. *Learning from a Population of Perceptrons*

The population learning formalism can also be applied to the learning of homogeneous linear threshold functions (perceptrons) with respect to a spherically symmetric input distribution. This learning problem is nontrivial, yet analytically tractable, so that the Kullback-Leibler divergence can calculated to within very tight bounds for the case of agent sample size $m = 1$.

THEOREM 7 *Let $\mathcal{F}_n$ be the class of homogeneous linear threshold functions on R^{n+1}, and let D be any spherically symmetric distribution over R^{n+1}. Then for the population learning problem $(\mathcal{F}_n, D, m = 1, A = Gibbs)$ we have*

$$\ell(\epsilon, \delta) = \mathcal{O}(n^2/\epsilon^4(\log 1/\epsilon)(\log n/\epsilon) + \log 1/\delta).$$

Proof: Each perceptron in the concept class is parametrized as $\mathrm{sgn}(\vec{w} \cdot \vec{x})$, where $\vec{w} \in R^{n+1}$ is constrained to lie on the unit n-sphere S^n (the magnitude of $\vec{w}$ does not matter). As shorthand notation, we will refer to a perceptron by its *weight vector* $\vec{w}$. We assume a spherically symmetric input distribution D on the input space $X = R^{n+1}$. The angle θ_{12} between two unit vectors $\vec{w}_1$ and $\vec{w}_2$ is defined by $\vec{w}_1 \cdot \vec{w}_2 = \cos\theta_{12}$. It is easily shown that the probability of disagreement between two perceptrons is proportional to θ_{12}:

$$D[\vec{w}_1 \Delta \vec{w}_2] = \frac{\theta_{12}}{\pi}. \tag{2}$$

This result depends on $\vec{w}_1$ and $\vec{w}_2$ only through the angle θ_{12} because of the spherical symmetry of the input distribution D.

For the case $m = 1$, the ratio of the probability density $dq_{\vec{w}_1}$ to the uniform density dq_0 is proportional to $1 - D[\vec{w}\Delta\vec{w}_1]$, that is,

$$\frac{dq_{\vec{w}_1}}{dq_0} = 2(1 - D[\vec{w}\Delta\vec{w}_1]) \tag{3}$$

because all version spaces determined by a single example have the same volume. The normalization constant 2 is set by noting that the expectation of $D[\vec{w}\Delta\vec{w}_1]$ for $\vec{w}$ drawn

according to dq_0 is $1/2$. An analogous result holds for $dq_{\vec{w}_2}$. In this continuous setting, the Kullback-Leibler divergence is defined by

$$KL(dq_{\vec{w}_1}||dq_{\vec{w}_2}) = \int dq_{\vec{w}_1} \log \frac{dq_{\vec{w}_1}}{dq_{\vec{w}_2}}. \tag{4}$$

In the appendix, this is evaluated using spherical coordinates. The resulting integral can be tightly bounded for large n using Laplace's method. The result described by Equation (7) depends on $\vec{w}_1$ and $\vec{w}_2$ through their angle, and implies

$$KL(\epsilon) = \frac{\epsilon^2}{\pi\sqrt{n}} + \mathcal{O}(\epsilon^4 n^{-1/2}) + \mathcal{O}(n^{-3/2})$$

for small $\epsilon = D[\vec{w}_1 \Delta \vec{w}_2]$. In particular, this implies that the separation function $\sigma_{KL}(\epsilon, 1) = \Omega(\epsilon^2/\sqrt{n})$. The only obstacle to application of Theorem 3 is the lack of a simple bound on $dim(\mathcal{Q})$ due to the infinite cardinality of $\mathcal{F}$. However, by constructing a maximal ϵ-separated set in $\mathcal{F}$, we can obtain a finite concept class $\mathcal{F}'$ of cardinality $\mathcal{O}((1/\epsilon)^n)$, the learning of which is equivalent to the learning of $\mathcal{F}$. This construction leads to a bound on $\ell(\epsilon, \delta)$ that is equivalent to that provided by Theorem 3 with the substitution of $dim(\mathcal{F}') = \mathcal{O}(n \log 1/\epsilon)$ for $dim(\mathcal{Q})$. □

6. Future Research

There are many open problems in the population learning model. Here is a small sampling:

- **Effects of Agent Sample Size.** It would be nice to prove general quantitative theorems regarding the effect of increasing the agent sample size m. This is perhaps the most important open problem, and some of the difficulties involved in its solution were discussed in Section 4.2. For instance, for the high-low game in Section 5, we showed that $\sigma_V(\epsilon, m)$ grows like $1 - e^{-\alpha\epsilon m}$; can we give general conditions under which such exponential behavior occurs?

- **Natural Algorithms.** The maximum likelihood or empirical loss minimization procedure we proposed, while providing very general upper bounds on population size, does not seem like the most natural method of combining hypotheses. On the other hand, we know that certain intuitive methods such as majority vote fail. It would be interesting to obtain good upper bounds on other natural approaches, such as weighted voting schemes.

- **Bounds on $dim(\mathcal{Q})$.** We suspect that except for degenerate classes, the combinatorial dimension $dim(\mathcal{Q})$ can be bounded by a slowly growing function of the Vapnik-Chervonenkis dimension of $\mathcal{F}$. It would be interesting to give conditions for this.

Acknowledgements

We give warm thanks to Rob Schapire for his help on the proof of Theorem 4, and for many interesting discussions on the research presented here.

Technical Appendix

This appendix gives the details of the calculation of the Kullback-Leibler divergence between two densities $dq_{\vec{w}_1}$ and $dq_{\vec{w}_2}$ induced on $\mathcal{F}$ by two perceptrons $\vec{w}_1$ and $\vec{w}_2$. We parametrize the concept class $\mathcal{F} = S^n$ using spherical coordinates:

$$\begin{aligned}
w_1 &= \cos\varphi_n \cos\varphi_{n-1} \cdots \cos\varphi_2 \cos\varphi_1 \\
w_2 &= \cos\varphi_n \cos\varphi_{n-1} \cdots \cos\varphi_2 \sin\varphi_1 \\
w_3 &= \cos\varphi_n \cos\varphi_{n-1} \cdots \sin\varphi_2 \\
&\vdots \\
w_n &= \cos\varphi_n \sin\varphi_{n-1} \\
w_{n+1} &= \sin\varphi_n.
\end{aligned}$$

Here $\varphi_1 \in [-\pi, \pi]$ and $\varphi_2, \ldots, \varphi_n \in [-\pi/2, \pi/2]$. The spherically symmetric measure on S^n is given by

$$dq_0 = \frac{1}{A_n} \cos^{n-1}\varphi_n \cos^{n-2}\varphi_{n-1} \cdots \cos\varphi_2 d\varphi_n d\varphi_{n-1} \cdots d\varphi_1$$

where the normalization constant A_n is the area of S^n, or

$$\begin{aligned}
A_n &= \int_{-\pi/2}^{\pi/2} d\varphi_n \cdots \int_{-\pi}^{\pi} d\varphi_1 \cos^{n-2}\varphi_{n-1} \cdots \cos\varphi_2 \\
&= \frac{2\pi^{(n+1)/2}}{\Gamma[(n+1)/2]}.
\end{aligned}$$

To write the densities $dq_{\vec{w}_1}$ and $dq_{\vec{w}_2}$ in spherical coordinates, we first align our coordinate system so that $\vec{w}_1 = (0, \ldots, 1)$ and $\vec{w}_2 = (0, \ldots, \sin\theta_{12}, \cos\theta_{12})$, which is consistent with $\vec{w}_1 \cdot \vec{w}_2 = \cos\theta_{12}$. This choice of coordinates, which involves no loss of generality, leads to

$$\begin{aligned}
\vec{w} \cdot \vec{w}_1 &= \sin\varphi_n \\
\vec{w} \cdot \vec{w}_2 &= \sin\varphi_n \cos\theta_{12} + \cos\varphi_n \sin\varphi_{n-1} \sin\theta_{12}.
\end{aligned}$$

For the sequel we define

$$R = \sin\varphi_n \cos\theta_{12} + \cos\varphi_n \sin\varphi_{n-1} \sin\theta_{12}$$

for notational brevity.

Substitution of this result in Equations (2) and (3) yields the induced densities

$$dq_{\vec{w}_1} = \left(1 + \frac{2\varphi_n}{\pi}\right) dq_0$$

$$dq_{\vec{w}_2} = \left[1 + \frac{2}{\pi}\sin^{-1} R\right] dq_0.$$

In spherical coordinates, the Kullback-Leibler divergence of Equation (4) takes the form

$$\begin{aligned} &KL(dq_{\vec{w}_1}||dq_{\vec{w}_2}) \\ &= \int dq_0 \left(1 + \frac{2\varphi_n}{\pi}\right) \log \frac{1 + \frac{2\varphi_n}{\pi}}{1 + \frac{2}{\pi}\sin^{-1} R} \\ &= \frac{A_{n-2}}{A_n} \int_{-\pi/2}^{\pi/2} d\varphi_n \cos^{n-1}\varphi_n \int_{-\pi/2}^{\pi/2} d\varphi_{n-1} \cos^{n-2}\varphi_{n-1} \\ &\quad \left(1 + \frac{2\varphi_n}{\pi}\right) \log \frac{1 + \frac{2\varphi_n}{\pi}}{1 + \frac{2}{\pi}\sin^{-1} R}. \end{aligned} \tag{5}$$

The last equality was obtained by performing the integral over $d\varphi_{n-2} \cdots d\varphi_1$, yielding A_{n-2}, the area of S^{n-2}.

For large n, we can derive an asymptotic expansion for this integral using Laplace's method (Erdelyi, 1956):

LEMMA 2 *(Laplace's Method) Let*

$$J(\lambda) = \int_I g(x) e^{-\lambda f(x)} dx \tag{6}$$

where λ is a large positive parameter and the integration domain I in R^N contains some neighborhood of the origin. If the minimum of f in I is at the origin, f and g possess fourth-order Taylor expansions about the origin, and the Hessian of f at the origin is positive definite, then

$$J(\lambda) = \int_{R^N} g_2(x) e^{-\lambda f_2(x)} dx + \mathcal{O}(\lambda^{-5/2})$$

where f_2 and g_2 are the second-order Taylor expansions of f and g.

The proof of this lemma can be found in many textbooks, but the intuition behind it is simple. As λ becomes large, only the neighborhood around the minimum of f contributes to the integral. Hence the integral can be approximated by Taylor expanding f and g.

The integral of Equation (5) can be put in the form of Equation (6) by setting $f = -\log\cos\varphi_{n-1} - \log\cos\varphi_n$, $\lambda = n - 1$, and g equal to the rest of the integrand. We Taylor expand f and g to second order, and perform the resulting Gaussian integrals, yielding

$$KL(dq_{\vec{w}_1}||dq_{\vec{w}_2}) = \frac{2}{\pi^3} \frac{1 - \cos\theta_{12}}{\sqrt{n}} + \mathcal{O}(n^{-3/2}) \tag{7}$$

where we have used $A_{n-2}/A_n = (n-1)/2\pi$.

References

Amari, S., Fujita, N., & Shinomoto, S. (1992). Four types of learning curves. *Neural Computation*, 4:605–618.

Cesa-Bianchi, N., Freund, Y., Helmbold, D.P., Haussler, D., Schapire, R.E., & Warmuth M.K. (1993). How to use expert advice. In *Proceedings of the Twenty-Fifth Annual ACM Symposium on the Theory of Computing*, pages 382–391.

Dudley, R.M. (1978). Central limit theorems for empirical measures. *The Annals of Probability*, 6(6):899–929.

Erdelyi, A. (1956). *Asymptotic expansions*. Dover.

Freund, Y. (1990). Boosting a weak learning algorithm by majority. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 202–216.

Freund, Y. (1992). An improved boosting algorithm and its implications on learning complexity. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 391–398.

Haussler, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150.

Haussler, D, Kearns, M., & Schapire, R.E. (1994). Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. *Machine Learning*, 14:83–113.

Kullback, S. (1967). A lower bound for discrimination information in terms of variation. *IEEE Transactions on Information Theory*, 13:126–127.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag.

Schapire, R.E. (1990). The strength of weak learnability. *Machine Learning*, 5(2):197–227.

Seung, H.S., Sompolinsky, H. & Tishby, N. Statistical mechanics of learning from examples. *Physical Review A*, 45(8):6056–6091.

Vapnik, V.N. (1982). *Estimation of Dependences Based on Empirical Data*. Springer-Verlag.

Vapnik, V.N. & Chervonenkis, A. Ya. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, XVI(2):264–280, 1971.

Received October 20, 1993
Accepted December 17, 1993
Final Manuscript June 30, 1994