Predicting {0,1}-Functions on Randomly Drawn Points

Extended Abstract

D. Haussler
N. Littlestone
M.K. Warmuth

Department of Computer and Information Sciences, University of California, Santa Cruz, CA. 95064.

Summary. We consider the problem of predicting $\{0,1\}$-valued functions on $\mathbf{R}^n$ and smaller domains, based on their values on randomly drawn points. Our model is related to Valiant's learnability model, but does not require the hypotheses used for prediction to be represented in any specified form. First we disregard computational complexity and show how to construct prediction strategies that are optimal to within a constant factor for any reasonable class F of target functions. These prediction strategies use the 1-inclusion graph structure from Alon, Haussler and Welzl's work on geometric range queries to minimize the probability of incorrect prediction. We then turn to computationally efficient algorithms. For indicator functions of axis-parallel rectangles and halfspaces in $\mathbf{R}^n$, we demonstrate how our techniques can be applied to construct computationally efficient prediction strategies that are optimal to within a constant factor. Finally, we compare the general performance of prediction strategies derived by our method to those derived from existing methods in Valiant's learnability theory.

1. The Prediction Model.

Let F be a class of $\{0,1\}$-valued functions on a fixed domain X. The domain can be finite, countably infinite, or $\mathbf{R}^n$ for some $n \geq 1$. We consider the problem of predicting the value of an unknown target function $f \in F$ on a randomly drawn point $x_t \in X$, given the value of f on randomly drawn points $x_1, ..., x_{t-1} \in X$. A rule for making such predictions is called a *prediction strategy for F*.

Given feedback from some external agent that indicates whether or not the prediction is correct, a prediction strategy can be iterated indefinitely. In each iteration, a random point is drawn, a prediction is made based on feedback from previous iterations, and new feedback is received. Each such iteration is called a *trial*.

Let us assume that an adversary who knows our prediction strategy is allowed to choose the target function $f \in F$ and the probability distribution P on X. All points are then drawn independently according to P. We are interested in finding a prediction strategy for F that for each $t \geq 1$, minimizes the probability that an incorrect prediction (*mistake*) is made on trial t. Since an adversary chooses the target function and the distribution, our objective is to minimize this probability in the worst case, over all $f \in F$ and distributions P on X. We use $\hat{\mathbf{M}}(t)$ to denote this worst case probability.

This prediction model is based on the non-probabilistic prediction model studied in [L87]. There the points $x_1, x_2, \cdots$ are directly selected by the adversary and the object is to make the fewest total number of mistakes. Our probabilistic variant of this model has the advantage that it is applicable even in situations where this worst case total number of mistakes is infinite, as often occurs when the domain X is infinite (e.g. $\mathbf{R}^n$, or Σ^*, for some finite alphabet Σ).

Our probabilistic prediction model is motivated by, and closely related to, the learnability model of Valiant [Val84] [BEHW86,87a,b] [AL87] [R87] [KLPV87]. In that model the learning algorithm must output a representation of a hypothesis in some hypothesis class H when given random examples of some unknown target function in F. The hypothesis must (with high probability) be a good approximation to the target function in the sense that (with high probability) it correctly predicts the value of the function on further examples drawn from the same distribution. As in our prediction model, these probabilities are calculated with respect to the worst case over all target functions in F and all distributions on the domain.

Our prediction model is essentially a streamlined version of the learnability model, in which the requirement that hypotheses be represented in a specified form

We gratefully acknowledge the support of ONR grant N00014-86-K-0454.

has been dropped, and only predictive performance is measured. We also simplify the model by considering only one probability (i.e. the probability of a mistake) instead of two (i.e. the probability that the algorithm produces a hypothesis with small probability of a mistake). The exact relationship between our prediction model and the learnability model is discussed further in Sections 6 and 7 below, and in detail in [HKLW88] and [PW88].

2. Summary of Results

As in [BEHW86,87b] and [L87], our results are of two types: the first type consists of results that indicate how well a learner can possibly do from the available information, if computational complexity is not an issue; the second type considers what can be done in polynomial time.

In our main result (Theorem 5.1) we show that, ignoring computational complexity, for any reasonable[1] class F of target functions we can construct a prediction strategy for F for which the worst case probability of a mistake on trial t, $\hat{M}(t)$, is within a constant factor of the best possible; we call such a prediction strategy *essentially optimal*. Specifically, when this prediction strategy is applied to target functions from F, then $\hat{M}(t)$ is at most $\frac{2VCdim(F)}{t}$, for any $t \geq 1$. Here $VCdim(F)$ denotes the *Vapnik-Chervonenkis dimension of F* as defined in [HW87] (following [VC71], [Vap82]). From a result in [EHKV87], it can be shown that if F is nontrivial (see Section 5) and $VCdim(F)$ is finite, then for any prediction strategy for F, $\hat{M}(t)$ is $\Omega(\frac{VCdim(F)}{t})$, and if $VCdim(F)$ is infinite, results from [BEHW86,87b] imply that $\hat{M}(t) \geq \frac{1}{2}$, for all $t \geq 1$ (Theorem 5.2). This shows that the upper bound is tight to within a constant factor.

Note that this essentially optimal prediction strategy has a pleasant property when applied to a class F of finite VC dimension. For any target concept in F and any distribution on X, the expected number of mistakes during the second half of any sequence of trials is at most $\sum_{i=t/2}^{t} \frac{2VCdim(F)}{i} \leq 2\ln(2)VCdim(F)$. For example, if $VCdim(F) = 5$ then when you iterate the strategy for 100 trials, the expected number of mistakes during the last 50 trials is at most 7, and when you iterate it for 1,000,000 trials the expected number of mistakes during the last 500,000 trials is still at most 7.

Our main result consists of two parts. First we introduce a technique for analyzing prediction algorithms using a combinatorial bound $\overset{*}{M}(t)$ that we call the *permutation mistake bound;* this bound is an upper bound for $\hat{M}(t)$ (Lemma 5.1). The bound is defined by looking at each possible sequence of examples and considering the fraction of all permutations of the sequence on which the algorithm makes a mistake in predicting the label of the last point. (The precise definition is given later.) This bound suggests a method of constructing prediction algorithms for which $\hat{M}(t)$ will be small: design the algorithm so that $\overset{*}{M}(t)$ is small. Using this idea, together with the 1-inclusion graph from [AHW87], we show how to construct a prediction strategy for an arbitrary F for which $\hat{M}(t) \leq \frac{2VCdim(F)}{t}$ (Lemmas 5.2-5.4).

If one can tell in polynomial time whether or not there is any function in F that is consistent with a sequence of examples, then our general construction gives a prediction strategy that runs in time polynomial in the trial index t, where the exponent is proportional to $VCdim(F)$. This prediction strategy is not practical when $VCdim(F)$ is large. However, we can improve this time bound considerably for many important classes of functions, including indicator functions for halfspaces and axis-parallel rectangles in $\mathbf{R}^n$. We do this by directly constructing efficient prediction strategies with permutation mistake bounds proportional to $\frac{VCdim(F)}{t}$ that run in time polynomial in both t and $VCdim(F)$ (see Examples 5.1 and 5.2).

We then compare the method developed here of constructing prediction strategies by making the permutation mistake bound small with another technique for constructing prediction strategies based on the learning results of [BEHW87b]. The latter technique constructs a hypothesis that is consistent with the previous trials and chosen from a fixed class H of possible hypotheses, and uses this hypothesis to make its prediction. We show that for any H the probability of making a mistake at trial t for such a prediction strategy is $O(\frac{VCdim(H)}{t}\log\frac{t}{VCdim(H)})$ (Theorem 6.1). Note that this bound depends on the VC dimension of the hypothesis space, whereas the optimal bound depends on the VC dimension of the target class. Any consistent algorithm must have $VCdim(H) \geq VCdim(F)$.

This result gives a useful upper bound, since most prediction strategies that are derived directly from learning algorithms are of the above type. However, even if $VCdim(H) = VCdim(F)$, this bound is still worse by a $\log\frac{t}{VCdim(F)}$ factor than that of the essentially optimal strategy given in Section 5. We show that this performance gap is real by exhibiting for every $d \geq 1$ a class of functions F with $VCdim(F) = d$, a target function $f \in F$, a consistent prediction strategy that always chooses a hypothesis from F and for each t a distribution on the domain of f, such that the probability of a mistake on trial t is $\Omega(\frac{VCdim(F)}{t}\log\frac{t}{VCdim(F)})$. (Theorem 6.2).

[1] When $X = \mathbf{R}^n$ we impose certain measurability constraints on F as in [BEHW86,87b].

3. Overview of Methods used in Main Result

In the model that we consider, the input to a prediction strategy is in the form of a sequence of independent random points selected according to an arbitrary, unknown distribution. Lack of knowledge about the distribution can make the performance of the strategy difficult to analyze. However, consider a fixed sequence of points $\bar{x} = (x_1, ..., x_t) \in X^t$. Whatever the underlying distribution, the process of independent random selection induces a uniform distribution on the permutations of this sequence; one is as likely to draw one permutation of $\bar{x}$ as another. As in [VC71] and [Vap82], we use this observation to obtain performance bounds for arbitrary distributions from combinatorial arguments about permutations of sequences.

Let the *permutation mistake bound* $\overset{*}{\mathrm{M}}(t)$ denote the supremum, over all $f \in F$ and all sequences $\bar{x}$ of t points in X, of the fraction of permutations of $\bar{x}$ for which the prediction strategy makes a mistake predicting the value of f on the last point, given its value on the previous points. For many prediction strategies, the permutation mistake bound can be calculated by a simple counting argument. Using the observation above, it is easy to show that $\hat{\mathrm{M}}(t) \le \overset{*}{\mathrm{M}}(t)$, so this function provides a convenient upper bound on our primary performance measure. All our bounds for $\hat{\mathrm{M}}(t)$, including those for the computationally efficient prediction strategies that we give for halfspaces and axis-parallel rectangles, are obtained by bounding $\overset{*}{\mathrm{M}}(t)$.

We now describe the principles that the essentially optimal strategy is based on. Let F be a class of $\{0,1\}$-valued functions on X. Two functions in F can be considered equivalent with respect to the fixed sequence $\bar{x}$ if their values agree on all the points in $\bar{x}$; hence $\bar{x}$ naturally partitions F into a set of at most 2^t equivalence classes. $VCdim(F)$ can be defined as the largest t such that there exists a sequence $\bar{x} = (x_1, ..., x_t)$ that induces 2^t distinct equivalence classes with respect to F.

Suppose that points $x_1, ..., x_{t-1}$ are labeled with the values of an unknown $f \in F$, and your task is to predict the value of f on x_t. If you are lucky then there is only one equivalence class consistent with the labels of $x_1, ..., x_{t-1}$ and thus the value of f on x_t is determined. However if there remains a pair of consistent equivalence classes (one for $f(x_t) = 0$ and the other for $f(x_t) = 1$) then the situation is ambiguous. We try to resolve this ambiguity in a way that will minimize $\overset{*}{\mathrm{M}}(t)$.

To do this we construct for the given F and $\bar{x}$ a graph G called the *1-inclusion graph* [AHW87], whose nodes are the equivalence classes of F induced by $\bar{x}$ and whose edges represent possible pairs of equivalence classes that may remain at trial t for some permutation of $\bar{x}$. By directing the edges of G, we can represent a prediction strategy for the permutations of $\bar{x}$. We show that if there is a way of directing the edges of every 1-inclusion graph derived from F so that the maximum outdegree of any node is k, then this defines a prediction strategy with $\overset{*}{\mathrm{M}}(t) \le k/t$.

For example, consider the class of indicator functions of closed intervals over **R**; each function in the class is 1 on some closed interval and 0 elsewhere. This function class has VC dimension 2. Now consider the equivalence classes induced by this function class on any sample of t distinct points $x_1 < x_2 < \cdots < x_t$. We illustrate the 1-inclusion graph for these equivalence classes in Figure 1. Each equivalence class is represented by the subset of $\{x_1, \ldots, x_t\}$ on which the functions in the equivalence class are 1. Though the degree of the 1-inclusion graph grows in proportion to t, the outdegree of each node will be at most two if all edges in the illustrated graph are directed upward. It turns out that this defines the simple prediction strategy of guessing 0 on x unless x lies between two points known to have value 1.

Our key combinatorial lemma (Lemma 5.2) shows that for any function class F the density (number of edges divided by number of nodes) of any 1-inclusion graph G derived from F is at most $VCdim(F)$. From this it follows that we can direct the edges of G so that the outdegree of each node is at most $2VCdim(F)$. This then implies that $\overset{*}{\mathrm{M}}(t) \le \frac{2VCdim(F)}{t}$, and hence $\hat{\mathrm{M}}(t) \le \frac{2VCdim(F)}{t}$.

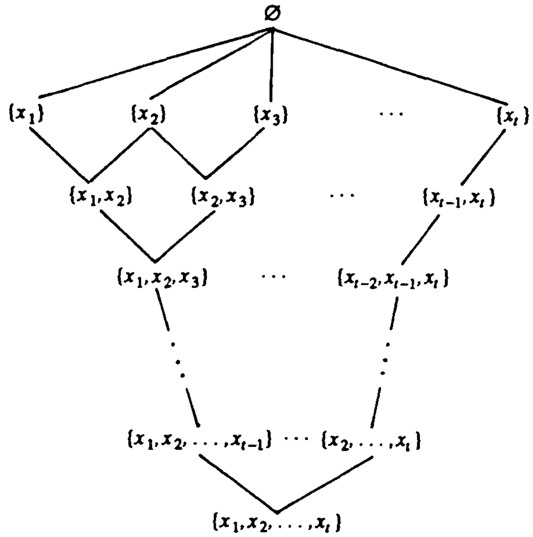

Figure 1.

4. Significance and Relation to Other Work

This work can be viewed as an extension of the general line of research initiated by Valiant in [Val84], aimed at developing a useful theoretical foundation for the analysis of machine learning algorithms used in Artificial Intelligence, Robotics and other areas. The approach here differs from the approach emphasized in the model introduced by Valiant in that we focus on the predictive performance of learning algorithms, whereas much of the work in Valiant's model has focused on learning representations, that is, on finding hypotheses that approximate the target function and are represented in a particular way. The learning problem has often been taken to be one of finding a hypothesis from the target class itself. The difficulty with this approach is that for many seemingly simple target classes it is NP-hard to learn the target class using hypotheses from the same class [KLPV87][2]. Here we work with a prediction model in which it is possible to explore issues of computationally practical learnability without being encumbered by restrictions on the hypothesis space (other than computational restrictions requiring the hypotheses to be evaluatable in polynomial time).

This point of view suggests new approaches to designing learning algorithms, and some of these are considered in this paper. All algorithms that we are aware of that have been constructed using the Valiant model work by finding consistent (or "almost" consistent) hypotheses in a hypothesis class of small size or of small VC dimension. We give here a completely different type of algorithm based on a direct strategy for minimizing the probability of predictive error, and demonstrate how it achieves improved predictive performance. The design techniques and analytic tools we develop here, especially those involving the permutation mistake bound, may also be useful in designing and analyzing other learning algorithms in situations where efficient on-line predictive performance is an issue (e.g. in Robotics). Regarding applications, it should also be noted that our measure of predictive performance is quite naturally extended to the more general measures used in statistical decision theory, in which the prediction algorithm can choose from a variety of responses, and receives a real-valued reward or penalty for each response depending on the current situation.

We also believe that the predictability model forms a more appropriate basis for obtaining hardness results for learning problems. If it can be shown that there are no efficient and effective prediction strategies for a class of functions F, then this implies that F is not polynomially learnable by forming hypotheses from any reasonable hypothesis class. Such hardness results have been obtained (modulo certain cryptographic assumptions) in [PW88] using the model introduced in this paper.

Finally, we have introduced a new combinatorial property of target classes of finite VC dimension in Lemma 5.2. The methods used in obtaining this result are similar to those used in combinatorial geometry to bound the number of cells, faces, edges, etc. in an arrangement of hyperplanes (see e.g. [E87]), hence this further underscores the unexpectedly close relationship between geometry and learning (see also [HW87] [W88] [EGS88a] [EGS88b] [EGHSSSW88] [CEGSW88] for applications of the VC dimension in geometry.) The general, essentially optimal prediction strategy we give is the first learning algorithm that directly exploits deeper combinatorial properties of target classes of finite VC dimension to make its predictions; others have simply made predictions by forming arbitrary consistent hypotheses from classes of small VC dimension. This further exploration of the structure of classes of finite VC dimension and its application to learning is one of the main contributions of this paper.

The technical section of the abstract follows.

Notation. We denote by X a set called the *domain*. We assume X is either finite, countably infinite, or $\mathbf{R}^n$ for some $n \geq 1$, where $\mathbf{R}$ denotes the real numbers. By F we denote a nonempty class of $\{0,1\}$-valued functions on X. We call such a class of functions a *concept class*[3].

For $t \geq 1$, $\bar{x} = (x_1, ..., x_t) \in X^t$ and $f \in F$, $sam(\bar{x}, f) = ((x_1, f(x_1)), ..., (x_t, f(x_t)))$. We call $sam(\bar{x}, f)$ the *sample of f generated by $\bar{x}$*, and each pair $(x_i, f(x_i))$ is called an *example of f*. We call an example (x_i, a) a *positive example* if $a = 1$ and a *negative example* if $a = 0$. We call a the *label* of the example. The *sample space* of F, denoted S_F, is defined by $S_F = \{sam(\bar{x}, f) : f \in F, \bar{x} \in X^t, t \geq 0\}$.

Let P be a probability distribution on X. For any $t \geq 1$, P^t denotes the corresponding product distribution on X^t. For any distribution T and random variable ψ, $E_T(\psi)$ denotes the expectation of ψ with respect to the distribution T.

5. Prediction Strategies

Definition. A *prediction strategy* for F is a mapping $Q : S_F \times X \rightarrow \{0,1\}$. For any $f \in F$, $t \geq 1$, and $(x_1, ..., x_t) \in X^t$, let

[2] For example, in [KLPV87] it is shown that the class of Boolean functions represented by 2-term DNF expressions is not polynomially learnable by any algorithm that must represent its hypothesis as a 2-term DNF (unless NP = RP); however, this class is polynomially learnable if we allow hypotheses represented as 2-CNF expressions.

[3] When $X = \mathbf{R}^n$, we make some measurability assumptions about F, as described in [BEHW86,87b].

$$\mathbf{M}^t_{Q,f}(x_1, ..., x_t) = \begin{cases} 1 & \text{if } Q(sam((x_1, ..., x_{t-1}), f), x_t) \neq f(x_t) \\ 0 & \text{otherwise} \end{cases}$$

Thus $\mathbf{M}^t_{Q,f}(x_1, ..., x_t)$ indicates whether or not the prediction strategy Q makes a mistake predicting the value of the concept f on x_t when given the value of f on $x_1, ..., x_{t-1}$.

Let $\hat{\mathbf{M}}_{Q,f}(t) = \sup E_{P^t}(\mathbf{M}^t_{Q,f})$ where the supremum is taken over all probability distributions P on X, and let $\hat{\mathbf{M}}_{Q,F}(t) = \sup \hat{\mathbf{M}}_{Q,f}(t)$ where the supremum is taken over all $f \in F$. This is the measure of predictive performance used in this paper. It represents the worst case probability of a mistake at the t^{th} trial.

$\hat{\mathbf{M}}_{Q,F}(t)$ involves a supremum over P^t where P is an arbitrary probability distribution on X; hence it is often difficult to measure. We deal with this by using a related bound $\overset{*}{\mathbf{M}}_{Q,F}(t)$ in whose definition drawing sequences according to P^t is replaced by drawing random permutations of fixed sequences of X^t. We will show that $\hat{\mathbf{M}}_{Q,F}(t) \leq \overset{*}{\mathbf{M}}_{Q,F}(t)$. The latter bound is easier to compute; we can estimate it using counting arguments.

Definition. Let Γ_t denote the set of all permutations of $\{1, ..., t\}$. For a prediction strategy Q, and $f \in F$, let

$$\overset{*}{\mathbf{M}}_{Q,f}(t) = \sup \frac{1}{t!} \sum_{\sigma \in \Gamma_t} \mathbf{M}^t_{Q,f}(x_{\sigma(1)}, ..., x_{\sigma(t)}),$$

where the supremum is taken over all $(x_1, ..., x_t) \in X^t$. Let $\overset{*}{\mathbf{M}}_{Q,F}(t) = \sup \overset{*}{\mathbf{M}}_{Q,f}(t)$ where the supremum is taken over all $f \in F$. $\overset{*}{\mathbf{M}}_{Q,F}(t)$ is called the *permutation mistake bound* of Q with respect to class F.

Lemma 5.1. *Let β be a real number and χ be a random variable defined on X^t for some $t \geq 1$. If for all sequences $(x_1, ..., x_t)$ of X^t,*

$$\frac{1}{t!} \sum_{\sigma \in \Gamma_t} \chi(x_{\sigma(1)}, ..., x_{\sigma(t)}) \leq \beta,$$

then this implies that $E_{P^t}(\chi) \leq \beta$, for any distribution P on X.

Proof. For any permutation $\sigma \in \Gamma_t$

$$\int_{X^t} \chi(x_1, ..., x_t) dP^t(x_1, ..., x_t)$$

$$= \int_{X^t} \chi(x_{\sigma(1)}, ..., x_{\sigma(t)}) dP^t(x_1, ..., x_t).$$

Thus

$$E_{P^t}(\chi) = \int_{X^t} \chi(x_1, ..., x_t) dP^t(x_1, ..., x_t)$$

$$= \frac{1}{t!} \sum_{\sigma \in \Gamma_t} \int_{X^t} \chi(x_{\sigma(1)}, ..., x_{\sigma(t)}) dP^t(x_1, ..., x_t)$$

$$= \int_{X^t} \frac{1}{t!} \sum_{\sigma \in \Gamma_t} \chi(x_{\sigma(1)}, ..., x_{\sigma(t)}) dP^t(x_1, ..., x_t)$$

$$\leq \int_{X^t} \beta dP^t(x_1, ..., x_t) = \beta. \qquad \square$$

Corollary 5.1. *For all concept classes F, all prediction strategies Q for F and all $t \geq 1$, $\hat{\mathbf{M}}_{Q,F}(t) \leq \overset{*}{\mathbf{M}}_{Q,F}(t)$.*

Proof. The above lemma implies that $\hat{\mathbf{M}}_{Q,f}(t) \leq \overset{*}{\mathbf{M}}_{Q,f}(t)$ for every $f \in F$; the result follows from this. □

We give an example of a simple strategy whose performance can be analyzed using the permutation mistake bound.

Example 5.1. Let the domain $X = \mathbf{R}^n$ and let the concept class F_n consist of indicator functions of axis-parallel rectangular regions in $\mathbf{R}^n$. The positive examples of each concept in F_n form a single closed and bounded rectangular region. The boundaries are hyperplanes of dimension $n-1$ which are parallel to the coordinate axes.

We consider the following prediction strategy for this concept class. As long as no positive examples have been seen, the prediction strategy predicts 0 for each new point. When positive examples have been seen, the prediction strategy keeps track of the smallest axis-parallel rectangular region which contains all of the positive examples seen so far. It predicts 1 if the new point is contained in this (closed) rectangular region and 0 otherwise.

The permutation mistake bound of this strategy is $2n/t$. We will demonstrate this for $n = 2$; the argument directly generalizes to higher dimensions. Consider any sequence s of points of $\mathbf{R}^2$. For each permutation of the sequence, we imagine that the prediction strategy has seen the correct classification of all but the last element of the permutation. It must make a prediction for the last element. We wish to bound the number of permutations for which the above strategy will make a mistake. The prediction strategy will never make a mistake for a negative example. Assume that the sequence contains at least one positive example. Let R denote the smallest axis-parallel rectangle containing the positive examples of the entire sequence s. Each edge of R will contain at least one point of s. Pick one such point from each edge of R to form a set E of size at most 4. Now note that if no point of E is the last point of a given permutation, then the rectangle constructed by the prediction strategy will match R, and the strategy will make the correct prediction for the last point. If the sequence s is of length t, then the prediction strategy can make a mistake for at most a fraction $4/t$ of the permutations: those with a point of E at the final element. (Note that if a point of E occurs more than once in s, then a mistake will not be made when that point is final). This upper bound is easily achieved, thus the permutation mistake bound is $4/t$.

We give an example of another strategy with small permutation mistake bound, using linear programming. In this case the details of the algorithm are expressly constructed to keep the bound as small as possible.

Example 5.2. Let $X = \mathbf{R}^n$ and let F_n be the set of indicator functions of closed halfspaces of $\mathbf{R}^n$. Given any finite sample s of a target function f in this class there exists a hyperplane strictly separating the positive examples in s from the negative examples. Such a separating hyperplane can be found in polynomial time using linear programming (with bit complexity model, using Karmarkar's algorithm [K84]). This gives a simple polynomial prediction strategy: given a sample $s \in S_{F_n}$ and $x \in \mathbf{R}^n$, solve a linear programming problem $Z(s)$ and predict 1 if x lies in the halfspace determined by the solution of $Z(s)$ that contains the positive points of s, 0 otherwise. The exact strategy depends on how $Z(s)$ is defined.

For the strategy LP that we construct, we can show a permutation mistake bound of $\frac{n+4}{t}$, implying that $\hat{\mathrm{M}}_{LP,F_n}(t) \leq \frac{n+4}{t}$. We can show that not all strategies using linear programming to find a separating hyperplane have a useful permutation mistake bound. There exist such strategies for which the permutation mistake bound is 1, which is a trivial bound on $\hat{\mathrm{M}}$. It turns out that the problem is due primarily to non-uniqueness of the optimal solution. We deal with this problem by randomizing our choice of the objective function. (In the full version of the paper we extend our prediction model to include such randomized strategies.)

As in the previous example, counting tight constraints (due to sample points) at an extremal consistent hypothesis is the key to getting a small permutation mistake bound. The extremal hypothesis used here is an optimal solution to a linear programming problem. We use Caratheodory's theorem for convex hulls to show that we do not need to count redundant constraints in calculating the permutation mistake bound. Details are omitted in this abstract.

Now we turn to the general essentially optimal prediction strategy based on the 1-inclusion graph. We first develop the background for the strategy.

Definition. Let X be a nonempty set and F be a nonempty class of $\{0,1\}$-valued functions on X. For any $S \subseteq X$, $\Pi_F(S)$ denotes the set of all $A \subseteq S$ such that there exists a function $f \in F$ that is 1 on A and 0 on $S - A$. The *Vapnik-Chervonenkis dimension* of F, denoted $VCdim(F)$, is

$$\sup \{ |S| : S \subseteq X \text{ and } \Pi_F(S) = 2^S \}.$$

Definition. Given any sequence $\bar{x} = (x_1, ..., x_t) \in X^t$ and any concept class F over X, we construct a graph that we call the *1-inclusion graph for F with respect to $\bar{x}$*, denoted $G_F(\bar{x})$ ([AHW87][4]). The nodes of $G_F(\bar{x})$ are just the elements of $\Pi_F(\{x_1, ..., x_t\})$. Let v and w be any two such nodes. They are connected with an edge if and only if the sets v and w differ by exactly one element x of $\bar{x}$ and x appears only once in $\bar{x}$. Each edge is labeled with the corresponding x.

Nodes of a 1-inclusion graph can have arbitrarily high degree, even for concept classes of Vapnik-Chervonenkis dimension 1. (Figure 1 gives an example for a class of VC dimension 2.) However, it turns out that the edges of any 1-inclusion graph for a concept class of Vapnik-Chervonenkis dimension d can be directed so that the outdegree of every node is at most $2d$. This result, which we develop in the next sequence of lemmas, is the basis for the 1-inclusion graph prediction strategy.

Definition. The density of a graph G, denoted $dens(G)$, is the number of its edges divided by the number of its nodes.

Lemma 5.2. *Let G be any 1-inclusion graph for F. Then $dens(G) \leq VCdim(F)$.*

Proof. Assume $\bar{x} = (x_1, ..., x_t) \in X^t$ and $G = G_F(\bar{x})$, where $VCdim(F) = d$. The proof proceeds by induction on t. When t is zero, G contains one node, and the result follows trivially for all d. For the induction step, we assume that $t > 0$ and that the result holds for $t-1$ for all d. There are two cases.

Case 1. $x_t = x_i$ for some $1 \leq i < t$.

Let $H = G_F(x_1, ..., x_{t-1})$. Since x_t already appears in $\{x_1, ..., x_{t-1}\}$, the node set of H is the same as that of G, and the edge set of G is the same as the edge set of H, except that all edges labeled with x_t are missing. Hence $dens(G) \leq dens(H) \leq d$ by assumption.

Case 2. $x_t \neq x_i$ for all $1 \leq i < t$.

We construct a mapping η of the nodes of G onto the nodes of another 1-inclusion graph H. Let $\bar{y} = (x_1, ..., x_{t-1})$. Then $H = G_F(\bar{y})$. Note that nodes of G are subsets of $\{x_1, ..., x_t\}$ and nodes of H are subsets of $\{x_1, ..., x_{t-1}\}$. The mapping η from nodes of G to nodes of H is defined by $\eta(v) = v - \{x_t\}$. Thus two nodes of G map to the same node of H if they differ only in whether or not they contain x_t. Note that either one or two nodes of G map to each node of H. Now let W be the set of all nodes of H whose inverse image under η contains two nodes of G. If W is empty, then the number of nodes of G equals the number of nodes of H and every edge of H corresponds to at most one edge of G. Thus in this case $dens(G) \leq dens(H) \leq d$ by the induction hypothesis, and we are done. (This case always occurs if $d = 0$.)

[4] Our definition of the 1-inclusion graph reduces to the one given in [AHW87] when all the points in $\bar{x}$ are distinct.

Otherwise, let F' be the concept class over X whose concepts are the indicator functions of the subsets of X which are the elements of W. One can show that the Vapnik-Chervonenkis dimension of F' is at most $d-1$ (if not, then there exists an $S \subseteq \{x_1, ..., x_{t-1}\}$ with $|S| = d$ such that $\Pi_F(S \cup \{x_t\}) = 2^{S \cup \{x_t\}}$, contradicting the fact that $VCdim(F) = d$.) We consider a third 1-inclusion graph, $J = G_{F'}(\bar{y})$. The nodes of J are just the elements of W. Let N_G, N_H, and N_J denote the number of nodes of G, H, and J, respectively. Let E_G, E_H, and E_J denote the number of edges of G, H, and J, respectively. The number of nodes of G is just the number of nodes of H plus the number of nodes of H to which two nodes of G map under η; i.e. $N_G = N_H + N_J$. To count the edges of G, first note that there will be one edge in G for each element of W, since any two nodes which map to the same node of H will be joined by an edge. Call these edges edges of the first type. The number of edges in G of the first type will equal N_J. For each other edge of G, there must be distinct nodes u and v of H such that the edge joins a node in $\eta^{-1}(\{u\})$ to a node of $\eta^{-1}(\{v\})$. Such edges can only occur when u and v are connected by an edge in H. There can be two such edges for a given u and v only if u and v are both in W, and then u and v will also be connected by an edge in J. Thus the number of edges in G of the second type will be bounded by the number of edges of H plus the number of edges of J. Thus we have $E_G \le E_H + E_J + N_J$. Thus

$$dens(G) \le \frac{E_H + E_J + N_J}{N_H + N_J} \le \max\left(\frac{E_H}{N_H}, \frac{E_J + N_J}{N_J}\right)$$

$$= \max(dens(H), dens(J) + 1).$$

From the induction hypothesis, we have $dens(H) \le d$ and $dens(J) \le d-1$, giving the desired result, $dens(G) \le d$. □

Lemma 5.3. *Let G be any graph with density at most d. Then G contains a node with degree at most $2d$.*

Proof. Assume to the contrary that every node has degree greater than $2d$. Then clearly the density of G exceeds d. □

Lemma 5.4. *Let $G = G_F(x_1, ..., x_t)$ be any 1-inclusion graph for a concept class F. Then the edges of G can be directed to form a directed graph with outdegree at most $2VCdim(F)$ such that the direction of the edges does not depend on the order in which the points of $\{x_1, \cdots, x_t\}$ appear in $(x_1, \cdots, x_t)$.*

Proof. Choose a permanently fixed ordering for the set of all finite subsets of X. Iteratively choose from among the minimum-degree vertices of G the smallest one according to this ordering, direct all edges away from the chosen vertex, and remove the vertex and the newly directed edges from G.

It is obvious that the resulting orientation of G is the same for all permutations of $(x_1, \cdots, x_t)$. To show the outdegree bound, note that any graph created by removing nodes from G is the 1-inclusion graph w.r.t. $(x_1, \cdots, x_t)$ for some concept class contained in F; hence by Lemma 5.2, each such graph has density at most $VCdim(F)$. Thus by Lemma 5.3, the node removed at each iteration of the above construction has degree at most $2VCdim(F)$. The outdegree of a vertex is equal to its degree when it was removed from the graph and thus the lemma follows. □

We can now describe the 1-inclusion graph prediction strategy. Assume that the target class is F. Given a labeled sample of points $x_1, \ldots, x_{t-1}$ and a point x_t for which a prediction is desired, we determine a prediction as follows. First construct the 1-inclusion graph $G = G_F(x_1, \ldots, x_t)$ and direct the edges so that the outdegree of each node is at most $2VCdim(F)$ by the method given above in Lemma 5.4. Each node of G is an element of $\Pi_F(\{x_1, \ldots, x_t\})$. The labeling of $x_1, \ldots, x_t$ in the sample can be consistent with at most two of the nodes, one for each possible labeling of x_t. There must be at least one consistent node since the sample is consistent with some concept in F. If there is exactly one consistent node v, then the strategy makes whichever prediction is consistent with that node. If there are two consistent nodes, then they differ only for x_t. Thus there must be an edge joining the nodes. The strategy chooses the node v towards which that edge is directed and makes a prediction consistent with v.

Theorem 5.1. *For any concept class F there exists a prediction strategy Q with permutation mistake bound $\hat{\mathrm{M}}_{Q,F}(t)$ (which upper bounds $\hat{\mathrm{M}}_{Q,F}(t)$) of at most $\frac{2VCdim(F)}{t}$.*

Proof. Let $\bar{x} = (x_1, ..., x_t) \in X^t$. Using the above 1-inclusion graph prediction strategy, a mistake is made predicting $f(x_t)$ given $f(x_1), ... f(x_{t-1})$ only when the node corresponding to the equivalence class containing f has an outgoing edge with label x_t. By definition of the 1-inclusion graph, this can happen only when x_t occurs just once in $\bar{x}$. Since the 1-inclusion graph is oriented the same for all permutations of $\bar{x}$, and in this orientation the outdegree of any node is bounded by $2VCdim(F)$, this implies that for any target function $f \in F$, the fraction of permutations of $\bar{x}$ for which a mistake is made predicting the last point is at most $\frac{2VCdim(F)}{t}$. □

We now show that this result is tight to within a constant factor.

Definition. A (nonempty) concept class $F \subseteq 2^X$ is *trivial* if F contains exactly one concept, or if F consists of two functions $f, g : X \to \{0, 1\}$ such that $f = 1-g$.

Theorem 5.2. Assume F is nontrivial and Q is any prediction strategy.

(i) If $VCdim(F)$ is infinite then $\hat{\mathbf{M}}_{Q,F}(t) \geq \frac{1}{2}$ for all $t \geq 1$.

(ii) If $VCdim(F)$ is finite then for all $t > VCdim(F)$, $\hat{\mathbf{M}}_{Q,F}(t) \geq \frac{\max(1, VCdim(F)-1)}{2et}$, where e is the base of the natural logarithm.

Proof. We use the technique from ([EHKV87]). Here we give only the proof of part (ii); part (i) is similar. We first consider the case $VCdim(F) \geq 2$. Let $k = VCdim(F) - 1$ and fix $t \geq k$. Assume $X_0 = \{y_0, ..., y_k\} \subseteq X$ is shattered by F. Let P be the distribution on X defined by

$$P(y_i) = \frac{1}{t}, 1 \leq i \leq k,$$

$$P(y_0) = 1 - \frac{k}{t}, \text{and}$$

$$P(x) = 0, x \in X - X_0.$$

Let $p = P^t\{(x_1, ..., x_t) \in X_0^t : x_t \neq x_i, 1 \leq i < t\}$. Note that $p \geq P^t\{(x_1, ..., x_t) \in X_0^t : x_t \neq y_0 \text{ and } x_t \neq x_i, 1 \leq i < t\} = \frac{k}{t}(1 - \frac{1}{t})^{t-1} \geq \frac{k}{et}$.

Let $F_0 = 2^{X_0} = \Pi_F(X_0)$ and T be the uniform distribution on F_0. Let $\mathbf{M}_Q^t(\bar{x}, f) = \mathbf{M}_{Q,f}^t(\bar{x})$ for all $f \in F_0$ and $\bar{x} \in X_0^t$. Given any sample $s = ((x_1, a_1), ..., (x_{t-1}, a_{t-1})) \in S_{F_0}$ and $x_t \in X_0$ such that $x_t \neq x_i$, $1 \leq i < t$, half of the concepts in F_0 that are consistent with s include x_t and the other half don't include x_t. Thus for any fixed $\bar{x} = (x_1, ..., x_t) \in X_0^t$ such that $x_t \neq x_i$, $1 \leq i < t$, $E_T(\mathbf{M}_Q^t(\bar{x}, f)) = \frac{1}{2}$ for any prediction strategy Q. This implies that $E_{P^t \times T}(\mathbf{M}_Q^t(\bar{x}, f)) \geq \frac{p}{2} \geq \frac{k}{2et}$. Hence there exists $f_0 \in F_0$ such that $E_{P^t}(\mathbf{M}_Q^t(\bar{x}, f_0)) \geq \frac{k}{2et}$, and hence for any $f \in F$ such that $f \cap X_0 = f_0$, $E_{P^t}(\mathbf{M}_{Q,f}^t) \geq \frac{k}{2et}$.

For $VCdim(F) \geq 2$, the above argument shows that for all $t \geq VCdim(F) - 1$ there exists $f \in F$ such that $E_{P^t}(\mathbf{M}_{Q,f}^t) \geq \frac{VCdim(F) - 1}{2et}$. If $VCdim(F) < 2$, then we can use the fact that F is nontrivial to find concepts f_1 and f_2 in F and points a and b in X such that a is in f_1 but not in f_2 and b is either both in f_1 and in f_2 or neither in f_1 nor in f_2. We then set $P(a) = \frac{1}{t}$ and $P(b) = 1 - \frac{1}{t}$. The remainder of the proof is similar to the above argument and shows that $E_{P^t}(\mathbf{M}_{Q,f}^t) \geq \frac{1}{2et}$ for either $f = f_1$ or $f = f_2$. □

We conclude this section by briefly discussing the implementation of the 1-inclusion graph strategy.

Definition [BEHW86]. A concept class F is *polynomially recognizable* if there exists a polynomial algorithm that, given a sample, decides whether there is a concept in F consistent with the sample.

Theorem 5.3. *If F is has finite VC dimension and is polynomially recognizable then the 1-inclusion graph can be created, directed and applied to predict in time polynomial in the size of the sample.*

Proof sketch. A theorem of Sauer [S72] shows that for any $(x_1, ..., x_t) \in X^t$, the cardinality of $\Pi_F\{x_1, ..., x_t\}$ is $O(t^{VCdim(F)})$. Hence the size of the 1-inclusion graph $G_F(\bar{x})$ is $O(VCdim(F)\, t^{VCdim(F)})$ by Lemma 5.2. If $VCdim(F)$ is finite and one can tell in polynomial time whether or not there is any function in F that is consistent with a given sequence of examples, then Lemma 12 of [BEHW86] shows how the sets in $\Pi_F(\bar{x})$ can be listed in polynomial time. Given this list, it is easy to construct $G_F(\bar{x})$, direct its edges as required, and make a prediction, in polynomial time. □

6. Prediction Strategies Using Hypothesis Spaces of Small VC Dimension

The results of the previous section indicate that one approach to constructing good prediction strategies is to construct strategies with small permutation mistake bounds. Here we consider another approach, based on the learning results of [BEHW86, 87b], in which a prediction is made on the basis of a consistent hypothesis chosen from a hypothesis space of small Vapnik-Chervonenkis dimension.

Definition. For any prediction strategy Q and sample $s \in S_F$, the *hypothesis of Q generated by s*, is the function $h : X \rightarrow \{0,1\}$ defined by $h(x) = Q(s, x)$. A hypothesis h is said to be *consistent* if it correctly labels the points of the sample s generating it, i.e. if for all pairs $(x, a) \in s$, $h(x) = a$.

Theorem 6.1. Let H be some collection of functions from X to $\{0,1\}$ with $|H| > 1$. Suppose that for all samples in S_F the hypothesis of prediction strategy Q is an element of H. Also assume that Q always chooses a hypothesis consistent with the sample. Then $\hat{\mathbf{M}}_{Q,F}(t)$ is $O(\frac{VCdim(H)}{t} \log \frac{t}{VCdim(H)})$ for $t > VCdim(H)$.

We prove a lemma first.

Definition. Let Q be a prediction strategy for F. For each $t \geq 1$, $f \in F$, distribution P on X and $\bar{x} \in X^t$, $\mathbf{ER}_{Q,f,P}^t(\bar{x}) = P\{x \in X : Q(sam(\bar{x}, f), x) \neq f(x)\}$. □

Thus $\mathbf{ER}_{Q,f,P}^t(\bar{x})$ is the probability that the hypothesis of Q generated by $sam(\bar{x}, f)$ disagrees with the target function f on a randomly drawn point. This is the notion of the "error" of a hypothesis used in Valiant's learning model [HKLW88] [Val84].

Lemma 6.1. *$E_{P^t}(\mathbf{ER}_{Q,f,P}^t) = E_{P^{t+1}}(\mathbf{M}_{Q,f}^{t+1})$ for any prediction strategy Q, target function $f \in F$, distribution P on X and $t \geq 0$.*

Proof.

$$E_{P^{t+1}}(\mathbf{M}_{Q,f}^{t+1}) = \int_{X^{t+1}} \mathbf{M}_{Q,f}^{t+1}(x_1, ..., x_{t+1}) \mathrm{d}P^{t+1}(x_1, ..., x_{t+1})$$

$$= \int_{X^t} \left[\int_X \mathbf{M}_{Q,f}^{t+1}(x_1, ..., x_{t+1}) \mathrm{d}P(x_{t+1}) \right] \mathrm{d}P^t(x_1, ..., x_t)$$

by Fubini's theorem. However, for fixed $(x_1, ..., x_t)$,

$$\int_X \mathbf{M}_{Q,f}^{t+1}(x_1, ..., x_{t+1}) \mathrm{d}P(x_{t+1}) = \mathbf{ER}_{Q,f,P}^t(x_1, ..., x_t)$$

by definition. Hence

$$E_{P^{t+1}}(\mathbf{M}_{Q,f}^{t+1}) = \int_{X^t} \mathbf{ER}_{Q,f,P}^t(x_1, ..., x_t) \mathrm{d}P^t(x_1, ..., x_t)$$

$$= E_{P^t}(\mathbf{ER}_{Q,f,P}^t). \quad \square$$

Proof of Theorem 6.1. Let $d = VCdim(H)$. Since $|H| > 1$, $d \geq 1$. Since Q is consistent and uses hypotheses in H, it follows from Theorem A2.4 and Proposition A2.5 of [BEHW87b] (see [VC71] [Vap82]) that for all $f \in F$, distributions P on X, $\varepsilon > 0$ and $t \geq d$,

$$P^t\{\bar{x} : \mathbf{ER}_{Q,f,P}^t(\bar{x}) \geq \varepsilon\} \leq 2(2et/d)^d 2^{-\varepsilon t/2}.$$

Since $\mathbf{ER}_{Q,f,P}^t \leq 1$, this implies that $E_{P^t}(\mathbf{ER}_{Q,f,P}^t) \leq \varepsilon + 2(2et/d)^d 2^{-\varepsilon t/2}$ for all $\varepsilon > 0$. Letting $\varepsilon = \frac{2}{t}(\log(t/d) + d\log(2et/d))$, it follows that $E_{P^t}(\mathbf{ER}_{Q,f,P}^t) \leq \frac{2}{t}(\log(t/d) + d\log(2et/d) + d) \leq \frac{2(d+1)}{t}\log(4et/d)$. The result then follows from Lemma 6.1. $\square$

Note that to be always able to find consistent hypotheses in H, $VCdim(H)$ must be at least $VCdim(F)$. Thus the bound of Theorem 6.1 is greater, by at least a factor proportional to $\log\frac{t}{VCdim(F)}$, than the bound for an algorithm with an optimal permutation mistake bound. The following theorem states that this bound is tight to within a constant factor.

Theorem 6.2. *There exists a family of countable domains $\{X_d\}_{d=1}^{\infty}$ and corresponding concept classes $\{F_d\}_{d=1}^{\infty}$ with $VCdim(F_d) = d$ for which there is a prediction strategy LEGAL such that:*

(1) when LEGAL is presented with a sample from S_{F_d} it chooses a consistent hypothesis from F_d

and

$$\hat{\mathbf{M}}_{LEGAL,F_d}(t) \geq \frac{1}{24}\frac{d}{t}\ln\frac{t}{d}.$$

Proof outline. We describe the function class and the prediction strategy *LEGAL* for which the theorem holds, omitting the proof of the lower bound. Let the domain X_d consist of d copies of the natural numbers, which are called buckets B_i $(1 \leq i \leq d)$. The function class F_d consists of all functions that are 1 on at most one point in each bucket. It is easy to see that the VC dimension of F_d is d.

Given sample $s = ((x_1, a_1), \cdots, (x_{t-1}, a_{t-1}))$ and an unlabeled point x_t in bucket B_i, *LEGAL* predicts as follows: if any point in bucket B_i occurs in s with label 1, then *LEGAL* predicts 1 iff x_t is this point, else *LEGAL* predicts 1 iff x_t is the smallest number in $B_i - \{x_1, \cdots, x_{t-1}\}$. Note that for any sample s in S_{F_d}, the hypothesis of *LEGAL* is 1 on exactly one point from each bucket and is thus in F_d. Clearly the hypothesis is also consistent with the sample s.

The theorem trivially holds for $1 \leq t \leq d$ since the lower bound is non-positive. Thus assume $t > d$. Recall that $\hat{\mathbf{M}}_{LEGAL,F_d}(t) =$ sup $E_{P^t}(\mathbf{M}_{LEGAL,f}^t)$ over all $f \in F_d$ and distributions P on X_d. To obtain a lower bound for $\hat{\mathbf{M}}_{LEGAL,F_d}(t)$, it suffices to exhibit a function $f \in F_d$ and a distribution $P_{t,d}$ on X_d and prove a lower bound on $E_{P_{t,d}^t}(\mathbf{M}_{LEGAL,f}^t)$. We can bound this expectation from below by $\frac{1}{24}\frac{d}{t}\ln\frac{t}{d}$ when f is the constant function 0 and $P_{t,d}$ is the distribution that is uniform over the union of the sets $\{1, ..., n\}$ from all d buckets and 0 elsewhere, where $n = \left| 15/e^{\frac{t/d}{\ln t/d}} \right|$. Details are omitted. $\square$

7. Further Results and Open Problems

It is easy to see that any learning algorithm in the sense of [Val84] or [BEHW86, 87b] can easily be converted into a prediction strategy: we simply use the hypothesis formed by the learning algorithm on the given sample to make our prediction. It is shown in [HKLW88] that the opposite is also true, i.e. any prediction strategy can be converted into a learning algorithm. In this case the hypothesis class used by the learning algorithm is given by the possible states of the prediction strategy, and is thus distinct, in general, from the class of target functions.

By applying this conversion to the 1-inclusion graph prediction strategy discussed in Section 5, we can obtain a learning algorithm that, for any target function in F and any distribution on the domain, produces, with probability at least $1 - \delta$, a hypothesis with error at most ε using $O(\frac{VCdim(F)}{\varepsilon}\log\frac{1}{\delta})$ independent random training examples. (As above, the "error" of a hypothesis is the probability that it disagrees with the target function on a randomly drawn point.) This sample size bound is, for some choices of ε and δ, better than the $O(\frac{VCdim(F)}{\varepsilon}\log\frac{1}{\varepsilon} + \frac{1}{\varepsilon}\log\frac{1}{\delta})$ bound given in [BEHW87b] for learning algorithms that produce consistent hypotheses in F, which was the previous best bound. It is still an open problem to find a general learning algorithm that meets the $\Omega(\frac{VCdim(F)}{\varepsilon} + \frac{1}{\varepsilon}\log\frac{1}{\delta})$ lower bound established in [EHKV87].

However, the most significant open problems that remain concern the existence of computationally efficient prediction strategies. As mentioned above, the time complexity of the 1-inclusion graph strategy grows exponentially as $VCdim(F)$ increases. Thus it is not suitable for many types of target classes, (e.g. the Boolean classes of monomials, k-DNF, decision trees, etc. [KLPV87] [EH87]), since here most applications require a domain with a large number of variables, and this implies a large VC dimension. However, in some cases there exists prediction strategies like those we have given for axis-parallel rectangles and half-spaces that have time complexities polynomial in the VC dimension. Most of these are obtained by converting learning algorithms into prediction strategies.

Recent results suggest that there may be no efficient prediction algorithm for some important target classes. Let us say that a general class of Boolean formulae is *polynomially bounded* if for each number n of variables, it only includes formulae that have length at most $p(n)$ bits, for some fixed polynomial p. In [KV88] it is shown that under certain cryptographic assumptions, the class of functions represented by polynomially bounded Boolean formulae has no efficient prediction strategy, for some suitable polynomial. It is still not known whether this is true for polynomially bounded DNF formulae, or for polynomially bounded decision trees.

Acknowledgements. We would like to thank Lenny Pitt, Michael Kearns and Bob Sloan for helpful comments on this material.

References

[AHW87] Alon, N., D. Haussler and E. Welzl, "Partitioning and geometric embedding of range spaces of finite Vapnik-Chervonenkis dimension," *Proc. 3rd Symp. on Computational Geometry,* Waterloo, June 87, pp. 331-340.

[AL87] Angluin, D., and P. D. Laird, "Identifying k-CNF formulas from noisy examples," *Machine Learning,* 2(4), 1987, pp. 343-370.

[BEHW86] Blumer, A., A. Ehrenfeucht, D. Haussler and M.K. Warmuth, "Classifying learnable geometric concepts with the Vapnik-Chervonenkis dimension," *Proc. 18th ACM Symp. on Theory of Computation,* Berkeley, CA., 1986, pp. 273 282.

[BEHW87a] Blumer, A., A. Ehrenfeucht, D. Haussler and M.K. Warmuth, "Occam's Razor," *Inf. Proc. Let.,* 24, 1987, pp. 377-380.

[BEHW87b] Blumer, A., A. Ehrenfeucht, D. Haussler and M.K. Warmuth, "Learnability and the Vapnik-Chervonenkis Dimension," Tech. Rep. UCSC-CRL-87-20, University of California, Santa Cruz, CA., November 1987.

[CEGSW88] Clarkson, K., H. Edelsbrunner, L. Guibas, M. Sharir, E. Welzl, "Complexity bounds for arrangements: the primal approach," in preparation.

[E87] Edelsbrunner, H., *Algorithms in Combinatorial Geometry,* Springer-Verlag, 1987.

[EGS88a] Edelsbrunner, H., L. Guibas and M. Sharir, "The complexity of many faces in arrangements of lines and of segments," *Proc. 4th Ann. Symp. on Computational Geometry,* to appear.

[EGS88b] Edelsbrunner, H., L. Guibas and M. Sharir, "The complexity of many cells in arrangements of planes and related problems," in preparation.

[EGHSSSW88] Edelsbrunner, H., L. Guibas, J. Hirshberger, R. Seidel, M. Sharir, M. Smoezink and E. Welzl, "Implicitly representing arrangements of lines or segments," *Proc. 4th Ann. Symp. on Computational Geometry,* to appear.

[EH87] Ehrenfeucht, A. and D. Haussler, "Learning decision trees from random examples," *Proceedings of the First Workshop on Computational Learning Theory,* MIT, August 1988, Morgan Kaufmann, publisher.

[EHKV87] Ehrenfeucht, A., D. Haussler, M. Kearns, and L. Valiant, "A general lower bound on the number of examples needed for learning," *Information and Computation,* to appear.

[HKLW88] Haussler, D., M. Kearns, N. Littlestone, and M. K. Warmuth, "Equivalence of models for polynomial learnability," *Proceedings of the First Workshop on Computational Learning Theory,* MIT, August 1988, Morgan Kaufmann, publisher.

[HW87] Haussler, D. and E. Welzl, "Epsilon-nets and simplex range queries," *Discrete Comput. Geom.* 2, 1987, pp. 127-151.

[K84] Karmarkar, N., "A New Polynomial-Time Algorithm for Linear Programming," *Combinatorica,* 4, 1984, pp. 373-395.

[KLPV87] Kearns, M., M. Li, L. Pitt, and L. Valiant, "On the learnability of Boolean Formulae," *Proc. 19th ACM Symp. on Theory of Computation,* New York, New York, 1987, pp. 285-295.

[KV88] Kearns, M. and L. Valiant, "Learning Boolean formulae or finite automata is as hard as factoring," Tech. Rep. TR-14-88, Aiken Computer Lab, Harvard Univ., Cambridge, MA, August, 1988.

[L87] Littlestone, N., "Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm," *Machine Learning,* 2(4), pp. 285-318, 1987.

[PW88] Pitt, L. and M.K. Warmuth, "Reductions among prediction problems: on the difficulty of predicting automata," extended abstract, *Proc. 3rd IEEE Conf. on Structure in Complexity Th.,* Washington, DC, June 88, pp. 60-69.

[R87] Rivest, R.L., "Learning decision-lists," *Machine Learning,* 2(3), pp. 229-246, 1987.

[S72] Sauer, N. "On the density of families of sets," *J. Combinatorial Theory (A),* 13, pp. 145-147, 1972.

[Vap82] Vapnik, V.N., *Estimation of Dependences Based on Empirical Data,* Springer Verlag, New York, 1982.

[Val84] Valiant, L.G., "A theory of the learnable," *Comm. ACM,* 27(11), 1984, pp. 1134-1142.

[VC71] Vapnik, V.N. and A.Ya.Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Th. Prob. and its Appl.,* 16(2), 1971, pp. 264-280.

[W88] Welzl, E., "Partition trees for triangle counting and other range search problems," *Proc. 4th Ann. Symp. on Computational Geometry,* to appear.