Relating Data Compression and Learnability

Nick Littlestone, Manfred K. Warmuth*
Department of Computer and Information Sciences
University of California at Santa Cruz

June 10, 1986

Abstract

We explore the learnability of two-valued functions from samples using the paradigm of Data Compression. A first algorithm (compression) choses a small subset of the sample which is called the kernel. A second algorithm predicts future values of the function from the kernel, i.e. the algorithm acts as an hypothesis for the function to be learned. The second algorithm must be able to reconstruct the correct function values when given a point of the original sample. We demonstrate that the existence of a suitable data compression scheme is sufficient to ensure learnability. We express the probability that the hypothesis predicts the function correctly on a random sample point as a function of the sample and kernel sizes. No assumptions are made on the probability distributions according to which the sample points are generated.

This approach provides an alternative to that of [BEHW86], which uses the Vapnik-Chervonenkis dimension to classify learnable geometric concepts. Our bounds are derived directly from the kernel size of the algorithms rather than from the Vapnik-Chervonenkis dimension of the hypothesis class. The proofs are simpler and the introduced compression scheme provides a rigorous model for studying data compression in connection with machine learning.

1 INTRODUCTION

In many learning problems one is learning a concept which is a subset of some *sample domain*. We consider the situation in which the points presented to the learner are selected at random from the sample domain according to some probability distribution. We study bounds on the rate of learning which are independent of the probability distribution, following an approach introduced by Valiant [V84]. In this paper we show that for a certain naturally arising class

*Both authors gratefully acknowledge the support of ONR grant N00014-86-K-0454 and the second author the support of the Faculty Research Committee of the University of California at Santa Cruz.

of learning algorithms, the bounds depend only on a simple characteristic of the algorithm (the size of what we call the kernel).

The learning model is as follows: Let X denote the sample domain. We are to learn concepts, which are subsets of X, from samples. Concepts are represented by their indicator function, i.e. a *concept* ξ is a mapping from X into $\{0,1\}$. During the learning process we will be given a sequence of observations of a particular concept ξ of some *class* C of concepts. Learning corresponds to finding a hypothesis that predicts the concept ξ with small error. The hypothesis is itself a concept, though not necessarily in C.

Observations are elements of $L(X) = X \times \{0,1\}$ and m-samples are sequences of $(X \times \{0,1\})^m$ which we denote by $L(X^m)$. We call m the *size* of such a sample. We are given a sample whose zero-one labels are determined by a particular ξ of class C: For any point $y \in X$, let $L_\xi(y) = (y, \xi(y))$ and for a sequence $\bar{x}$ of m observations, let $L_\xi(\bar{x}) = (L_\xi(x_1), ..., L_\xi(x_m))$.

As an example, the sample domain might be E^2, the Euclidean plane. A class of concepts would be some collection of figures, e.g. the set of all triangular regions. The aim is to learn a particular triangle. We receive observations of that triangle, i.e. points on the plane with labels 0 or 1 according to whether or not they are in the triangle.

Let P be an arbitrary but fixed probability distribution on X (in our example, on the points of the plane). The points of the sample are drawn according to this distribution and labeled with some ξ of C. After drawing m samples the *learning algorithm* forms an hypothesis. As in [V84] and [BEHW86] the hypothesis is evaluated with the same distribution P. The *error* of the hypothesis is the probability (according to P) that the hypothesis disagrees with ξ on the next random point of X, drawn according to P. A learning algorithm must have the following properties ([V84], [BEHW86]):

(L1) The error can be made arbitrarily small with arbitrarily high probability by taking m large enough. The bounds on m are to be independent of the concept ξ we are trying to learn and of the underlying distribution P.

(L2) The bounds on m are to be polynomial in the inverse of the error probabilities. Also the computation of the hypothesis as well as the computation of the value of the hypothesis for a given point must be polynomial in the length of the sample.

A class of concepts for which there exists an algorithm that fulfills (L1) is called *learnable*. If (L2) holds as well then the class is *polynomially learnable*. Condition (L1) is formalized by demanding error greater than ϵ with probability at most δ for small ϵ and δ, uniformly for all concepts in C. Condition (L2) implies that the number of required samples is a function $m(\epsilon, \delta)$ that grows polynomially in $1/\epsilon$ and $1/\delta$.

In [BEHW86] necessary and sufficient conditions for learnability are given in terms of the Vapnik-Chervonenkis dimension [VC 71] of the concept class. Bounds on the rate of learning are given which are functions of the Vapnik-Chervonenkis dimension. The results are non-constructive in the sense that

they lead to no specific algorithm for learning. Using the results of that paper, the steps to constructing a learning algorithm and verifying that it learns with a polynomial number of sample points are:

(S1) Construct an algorithm which, given any finite sample, generates a hypothesis that is consistent with the sample. In the main theorem it is required that the hypothesis be a member of the class of concepts being learned. Later they allow other hypothesis classes.

(S2) Find the Vapnik-Chervonenkis dimension of the class from which the hypotheses are chosen. A finite dimension demonstrates learnability and yields bounds on the speed of learning.

Our approach is motivated by various examples found in that paper. In a number of those examples, an algorithm is given in which the hypothesis is specified in terms a fixed-size subsample of the given labeled sample. For example, the concept of an orthogonal rectangle in the plane is determined by four observations. In a sense the sample of size m is compressed to a sample of fixed size. From the compressed sample the labels of the original m-sample can be reconstructed. Similarly other figures in E^r like polygons, half spaces, etc, are determined by a small number of points. Another example of this is the algorithm of [BL86] which uses two points to determine a half-plane in E^2.

In this paper, we show that if an algorithm has this characteristic of data compression (more explicitly specified below) then that alone is sufficient to guarantee learnability; it is not necessary to refer to the Vapnik-Chervonenkis dimension. In other words, we can leave out step (S2). Also, we do not require that the hypotheses themselves lie in any particular class of concepts; they can be arbitrary Borel sets of X. Bounds on the rate of learning are given in terms of the size of the compressed subsample (we call this the *kernel* size). In examples of learning geometric concepts which we have examined, these bounds are better than the bounds derived from the Vapnik-Chervonenkis dimension. The precise general relation between the bounds yielded by the two approaches is not known. The difficulty of finding a general relationship between the bounds reflects a substantial difference between the two approaches which should make them valuable supplements of each other. We expect the data compression algorithms described here to exist for a wide variety of concept classes, providing an easily applied alternative to the approach of [BEHW86].

Our basic results relating data compression to learnability are based on the conditions (L1) and (L2). The proof technique used is amenable to relaxation of these conditions. After presenting the basic compression scheme we suggest some extensions.

Notation Sequences or tuples are denoted with barred variables, i.e. the elements of X^m are denoted with $\bar{x}$. The i-th point of $\bar{x}$ is x_i, for $1 \leq i \leq m$. A *subsequence* of $\bar{x}$ is a sequence $x_{t_1}, x_{t_2}, ..., x_{t_k}$, s.t. $1 \leq t_1 < t_2 < ... < t_k \leq m$. I_Q denotes the indicator function of set Q.

2 THE BASIC COMPRESSION SCHEME

In this section we study learnability in relation to the basic compression scheme presented in the introduction.

We consider data compression schemes of the following form: Given a concept class C, a data compression scheme with kernel-size k consists of a pair of mappings

$$\kappa : \bigcup_{m=k}^{\infty} L(X^m) \to L(X^k) \text{ and } \rho : L(X^k) \times X \to \{0,1\}$$

such that

(C1) For any $\xi \in C$ and any $\bar{x} \in X^m$, for any $m \geq k$, $\kappa(L_\xi(\bar{x}))$ is a subsequence of length k of $L_\xi(\bar{x})$.

(C2) For any $\xi \in C$, any m, any $\bar{x} \in X^m$, and any point x_i of $\bar{x}$, $\rho(\kappa(L_\xi(\bar{x})), x_i) = \xi(x_i)$.

We call $\kappa(L_\xi(\bar{x}))$ the *kernel* of the sample. The second condition specifies that ρ reconstructs the labels of the sample points correctly. We say that $\rho(\kappa(L_\xi(\bar{x})), .)$ is *consistent* with ξ on $\bar{x}$. Usually both mappings are given by algorithms. We assume that the reconstruction function ρ is Borel measurable. (This holds, for example, for functions on $\mathbb{R}^n$ built recursively from ordinary comparison and arithmetic operations.) Throughout the paper, we also assume that the concepts in C are Borel measurable. Note that to make our notation simple we assume that kernels always have the same size and the sample-size m is always at least k. We define the kernel size of a concept class to be the minimum kernel size of all compression schemes.

A data compression scheme of this form can be used as the basis of a learning algorithm. Given a labeled sample, $L_\xi(\bar{x})$, the algorithm makes the hypothesis that the concept is the set $\{y : \rho(\kappa(L_\xi(\bar{x})), y) = 1\}$. Determining the kernel with κ corresponds to computing the hypothesis, i.e. the kernel encodes the hypothesis. The computation of the value of the hypothesis is achieved with ρ using the kernel as an input. To fullfil the condition L2 for polynomial learnability the algorithms ρ and κ must be polynomial in the length their input and the sample size $m(\epsilon, \delta)$ must be polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$. We show that whenever there is a compression scheme with fixed kernal size then $m(\epsilon, \delta)$ is always polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$.

The basic scheme is appealing because of its simplicity and generality. The sample is compressed to the kernel but ρ must be able to reconstruct the values of the sample. Note that we don't require any bounds on the length of the encoding of the kernel. The points of X might for instance be reals of arbitrary high precision. Compression to a bounded number of bits is discussed in [BEHW87] and is much simpler.

Theorem 2.1 For any compression scheme with kernel size k the error is larger than ϵ with probability (w.r.t. P^m) less than $\binom{m}{k}(1-\epsilon)^{m-k}$ when given a sample of size $m \geq k$.

Proof: Suppose we are learning some concept ξ. Given an ϵ and an m, we want to find a bound on the probability of choosing an m-sample which leads to a hypothesis with error greater than ϵ. In other words, we want to bound the error probability $P^m(E)$ where

$$E = \{\bar{x} \in X^m : P(\{y : \rho(\kappa(L_\xi(\bar{x})), y) \neq \xi(y)\}) > \epsilon\}.$$

Equivalently,

$$E = \{\bar{x} \in X^m : P(\{y : \rho(\kappa(L_\xi(\bar{x})), y) = \xi(y)\}) < 1 - \epsilon\}.$$

Let T be the collection of all k-element subsequences of the sequence $(1, 2, ..., m)$. For any $\bar{t} = (t_1, ..., t_k) \in T$, let

$$A_{\bar{t}} = \{\bar{x} \in X^m : \kappa(L_\xi(\bar{x})) = L_\xi(x_{t_1}, ..., x_{t_k})\}$$

$$E_{\bar{t}} = \{\bar{x} \in A_{\bar{t}} : P(\{y : \rho(\kappa(L_\xi(\bar{x})), y) = \xi(y)\}) < 1 - \epsilon\}$$

$$U_{\bar{t}} = \{\bar{x} \in X^m : P(\{y : \rho(L_\xi(x_{t_1}, ..., x_{t_k}), y) = \xi(y)\}) < 1 - \epsilon\}.$$

$$B_{\bar{t}} = \{\bar{x} \in X^m : mark\rho(L_\xi(x_{t_1}, ..., x_{t_k}), x_i) = \xi(x_i),$$
$$\text{for all } x_i \text{ with } i \notin \bar{t}\}.$$

We have $E_{\bar{t}} = E \cap A_{\bar{t}}$ and since $X^m = \bigcup_{\bar{t} \in T} A_{\bar{t}}$, $E = \bigcup_{\bar{t} \in T} E_{\bar{t}}$. From the definition of $A_{\bar{t}}$ we get

$$E_{\bar{t}} = \{\bar{x} \in A_{\bar{t}} : P(\{y : \rho(L_\xi(x_{t_1}, ..., x_{t_k}), y) = \xi(y)\}) < 1 - \epsilon\}.$$

Thus $E_{\bar{t}} = U_{\bar{t}} \cap A_{\bar{t}}$. The Condition (C2) guarantees that $A_{\bar{t}} \subset B_{\bar{t}}$. Roughly, these sets serve us as follows: We split X^m into the $A_{\bar{t}}$ (which only overlap where m-samples have repeated points). We then look at the intersection of E with each of these $A_{\bar{t}}$. Extending these intersections to the sets $U_{\bar{t}} \cap B_{\bar{t}}$ eliminates explicit dependence of the sets on κ and gives us sets whose probabilities can be easily bounded. We have

$$P^m(E_{\bar{t}}) \leq P^m(B_{\bar{t}} \cap U_{\bar{t}}).$$

It will now be convenient to rearrange the coordinates. Let $\pi_{\bar{t}}$ be any permutation of $1, 2, ..., m$ which sends i to ti, for $i = 1, ..., k$. Let $\phi_{\bar{t}} : X^m \to X^m$ send $(x_1, x_2, ..., x_m)$ to $(x_{\pi_{\bar{t}}(1)}, x_{\pi_{\bar{t}}(2)}, ..., x_{\pi_{\bar{t}}(m)})$. We have

$$P^m(B_{\bar{t}} \cap U_{\bar{t}}) = P^m(\phi_{\bar{t}}(B_{\bar{t}}) \cap \phi_{\bar{t}}(U_{\bar{t}})) = \int_{\phi_{\bar{t}}(U_{\bar{t}})} I_{\phi_{\bar{t}}(B_t)} dP^m.$$

Note that

$$\phi_{\bar{t}}(U_{\bar{t}}) = \{\bar{x} \in X^m : P(\{y : \rho(L_\xi(x_1, ..., x_k), y) = \xi(y)\}) < 1 - \epsilon\}.$$

Thus there exists some set $V_{\bar{t}} \subset L(X^k)$ such that $\phi_{\bar{t}}(U_{\bar{t}}) = V_{\bar{t}} \times X^{m-k}$. By Fubini's Theorem

$$\int_{\phi_{\bar{t}}(U_{\bar{t}})} I_{\phi_{\bar{t}}(B_t)} dP^m = \int_{V_{\bar{t}}} dP^k \int_{X^{m-k}} I_{\phi_{\bar{t}}(B_t)} dP^{m-k}.$$

We have

$$\phi_{\bar{t}}(B_{\bar{t}}) = \{\bar{x} \in X^m : \rho(L_\xi(x_1, ..., x_k), x_i) = \xi(x_i), for i = k+1, ..., m\}.$$

Let

$$W_{x_1,...,x_k} = \{y \in X : \rho(L_\xi(x_1, ..., x_k), y) = \xi(y)\}$$

Now

$$(x_1, ..., x_k) \times X^{m-k} \cap \phi_{\bar{t}}(B_{\bar{t}}) = (x_1, ..., x_k) \times W^m_{x_1,...,x_k} - k.$$

Thus the inner integral equals $P^{m-k}(W^{m-k}_{x_1,...,x_k})$. Since $(x_1, ..., x_k) \in V_{\bar{t}}$ we have $P(W_{x_1,...,x_k}) < 1 - \epsilon$. Thus the inner integral is bounded by $(1-\epsilon)^{m-k}$. This then bounds the entire integral, and we get

$$P^m(E_{\bar{t}}) < (1-\epsilon)^{m-k}.$$

Since the size of T is $\binom{m}{k}$, we have

$$P^m(E) < \binom{m}{k}(1-\epsilon)^{m-k}.$$

□

Remark: The proof depends on the measurability of the sets $W_{x_1,...,x_k}$, $U_{\bar{t}}$, and $B_{\bar{t}}$. The measurability of $W_{x_1,...,x_k}$ and $B_{\bar{t}}$ follows from the measurability of ρ using the fact that compositions of Borel measurable functions are Borel measurable. To see the measurability of $U_{\bar{t}}$, let

$$W = \{(\bar{x}, y) : \rho(L_\xi(x_{t_1}, ..., x_{t_k}), y) = \xi(y)\}$$

The set W is measurable, so the function

$$w(\bar{x}) = P(\{y : (\bar{x}, y) \in W\})$$

is measurable. (This follows by a simple case of Fubini's theorem ([R74]).) Thus $U_{\bar{t}} = \{\bar{x} : w(\bar{x}) < 1 - \epsilon\}$ is measurable.

In the following theorem we give explicit bounds for the sample size that guarantee learnability. A similar bound $m \geq \max\left(\frac{4}{\epsilon}\log\frac{2}{\delta}, \frac{8d}{\epsilon}\log\frac{8d}{\epsilon}\right)$ was given in [BEHW86], where d denotes the Vapnik-Chervonenkis dimension of the class to be learned. For example, in the case of learning n-dimensional orthogonal rectangles the dimension is $2n$. The kernel size of the straight forward compression scheme is also $2n$. Thus the bounds stated in the following theorem are better roughly by a factor of two. The dimension and the kernel size are not always equal. In the case of arbitrary halfplanes the dimension is three but there exists an algorithm with kernel size two ([BL86]).

Theorem 2.2: Any compression scheme with kernel-size $k \geq 1$ produces with probability at least $1 - \delta$ a hypothesis with error at most ϵ when given a sample of size

$$m \geq \max(\frac{2}{\epsilon}\ln(\frac{1}{\delta}), \frac{4k}{\epsilon}\ln(\frac{4k}{\epsilon}) + 2k).$$

This holds for arbitrary ϵ and δ.

Proof: Follows from the bound of the previous theorem. Applying the previous theorem it suffices to show that if m fulfills the bound then $\binom{m}{k}(1-\epsilon)^{m-k} \leq \delta$. This can be rewritten as

$$m \geq \frac{\ln \frac{1}{\delta} + \ln \binom{m}{k}}{-\ln(1-\epsilon)} + k$$

which holds if

$$m \geq \frac{1}{\epsilon}(\ln \frac{1}{\delta} + k \ln(m)) + k = \frac{1}{\epsilon}(\ln \frac{1}{\delta}) + k(\frac{1}{\epsilon} \ln(m) + 1)$$

There are two summands in the last expression. The inequality certainly holds if each summand is at most $\frac{m}{2}$. For the first summand this easily leads to the first bound in the maximum expression of the theorem. Similarly the second summand will lead to the second bound. If

$$\frac{m}{2} \geq k(\frac{1}{\epsilon} \ln(m) + 1)$$

holds when m is equal to the second bound then it also holds for larger m. Replacing m by the second bound in the above inequality leads to

$$\frac{2k}{\epsilon} \ln(\frac{4k}{\epsilon}) + k \geq \frac{k}{\epsilon}(\ln(\frac{4k}{\epsilon}) + \ln(\ln(\frac{4k}{\epsilon}) + \frac{\epsilon}{2})) + k,$$

which simplifies to $\frac{4k}{\epsilon} \geq \ln(\frac{4k}{\epsilon}) + \frac{1}{2}$ and can easily be verified. □

3 ADDITIONAL INFORMATION

In this section we extend the basic scheme by allowing additional information besides the kernel. The m-sample is compressed to an element of some finite set Q besides a kernel of size k. The set Q represents the additional information which κ is providing to ρ. Now ρ receives an element of Q and the kernel as an input. More exactly κ and ρ are redefined as follows:

$$\kappa : \bigcup_{m=k}^{\infty} L(X^m) \to Q \times L(X)^k, \rho : Q \times L(X^k) \times X \to \{0,1\}$$

For example if one wants to learn unions of n orthogonal rectangles, then we clearly need $4n$ observations. But which observation belongs to what rectangle? The $4n$ observations must be a subsequence of the original sample. We need the additional information to specify a particular permutation of the $4n$ points. After permuting, the first four points might determine the first rectangle, the next four points the second rectangle and so forth. Given the additional information, ρ knows the locations of the rectangles and can predict accordingly.

We now generalize the theorems of the previous section. The case $k = 0$ in which the sample is compressed to $\ln(|Q|)$ bits was studied in [BEHW87] Our bounds always contain the bounds of [BEHW87] as a subcase.

Theorem 3.1:

For a compression scheme with kernel size k and additional information Q the error is larger than ϵ with probability less than $|Q|\binom{m}{k}(1-\epsilon)^{m-k}$ when given a sample of size $m \geq k$.

Proof: This proof is an extension of the proof of Theorem 2.1. Again we want to bound $P^m(E)$ where

$$E = \{\bar{x} \in X^m : P(\{y : \rho(\kappa(L_\xi(\bar{x})), y) = \xi(y)\}) < 1 - \epsilon\}.$$

The index set T is extended with Q. The sequences of T now consist of an element of Q followed by an k-element subsequence of $(1, 2, ..., m)$. We adapt the definitions of $A_{\bar{t}}, E_{\bar{t}}, U_{\bar{t}}$ and $B_{\bar{t}}$. For any $\bar{t} = (t_0, t_1, ..., t_k) \in T$,

$$A_{\bar{t}} = \{\bar{x} \in X^m : \kappa(L_\xi(\bar{x})) = (t_0, L_\xi(x_{t_1}), ..., L_\xi(x_{t_k}))\}$$

$$E_{\bar{t}} = \{\bar{x} \in A_{\bar{t}} : P(\{y : \rho(t_0, L_\xi(x_{t_1}), ..., L_\xi(x_{t_k}), y) = \xi(y)\}) < 1 - \epsilon\}.$$

$$U_{\bar{t}} = \{\bar{x} \in X^m : P(\{y : \rho(t_0, L_\xi(x_{t_1}), ..., L_\xi(x_{t_k}), y) = \xi(y)\}) < 1 - \epsilon\}.$$

$$B_{\bar{t}} = \{\bar{x} \in X^m : \rho(t_0, L_\xi(x_{t_1}), ..., L_\xi(x_{t_k}), x_i) = \xi(x_i),$$
$$\text{for all } x_i \text{ with } i \notin \{t_i : 1 \leq i \leq k\}\}.$$

Again we have

$$P^m(E_{\bar{t}}) \leq P^m(B_{\bar{t}} \cap U_{\bar{t}}).$$

and we rearrange the coordinates. For any permutation $\pi_{\bar{t}}$ of $1, 2, ..., m$ which sends i to ti, for $i = 1, ..., k$, let $\phi_{\bar{t}} : Q \times X^m \to Q \times X^m$ send $(t_0, x_1, x_2, ..., x_m)$ to $(t_0, x_{\pi_{\bar{t}}(1)}, x_{\pi_{\bar{t}}(2)}, ..., x_{\pi_{\bar{t}}(m)})$. phi rearranges $U_{\bar{t}}$ and $B_{\bar{t}}$:

$$\phi_{\bar{t}}(U_{\bar{t}}) = \{\bar{x} \in X^m : P(\{y : \rho(t_0, L_\xi(x_1), ..., L_\xi(x_k), y) = \xi(y)\}) < 1 - \epsilon\},$$

$$\phi_{\bar{t}}(B_{\bar{t}}) = \{\bar{x} \in X^m : \rho(t_0, L_\xi(x_1), ..., L_\xi(_k), x_i) = \xi(x_i) for i = k+1, ..., m\}.$$

Again

$$P^m(B_{\bar{t}} \cap U_{\bar{t}}) = P^m(\phi_{\bar{t}}(B_{\bar{t}}) \cap \phi_{\bar{t}}(U_{\bar{t}}) = \int_{\phi_{\bar{t}}(U_{\bar{t}})} I_{\phi_{\bar{t}}(B_t)} dP^m = \int_{V_{\bar{t}}} dP^k \int_{X^{m-k}} I_{\phi_{\bar{t}}(B_t)} dP^{m-k},$$

where $V_{\bar{t}} \subset L(X^k)$ such that $\phi_{\bar{t}}(U_{\bar{t}}) = V_{\bar{t}} \times X^{m-k}$. We now describe $\phi_{\bar{t}}(B_{\bar{t}})$ using

$$W_{t_0, x_1, ..., x_k} = \{y \in X : \rho(t_0, L_\xi(x_1), ..., L_\xi(x_k), y) = \xi(y)\}$$

Clearly

$$(x_1, ..., x_k) \times X^{m-k} \cap \phi_{\bar{t}}(B_{\bar{t}}) = (x_1, ..., x_k) \times W^m_{t_0, x_1, ..., x_k} - k.$$

The inner integral equals $P^{m-k}(W^{m-k}_{t_0, x_1, ..., x_k})$ which is less than $(1-\epsilon)^{m-k}$, since $\{x_1, ..., x_k\} \in V_{\bar{t}}\}$. The entire integral and therefore $E_{\bar{t}}$ are bounded in the

same way. In the case of the compressions scheme with additional information $|T| = |Q|\binom{m}{k}$. Thus

$$P^m(E) < |Q|\binom{m}{k}(1-\epsilon)^{m-k}$$

Note that for the basic compression scheme (Theorem 2.1) $|T| = \binom{m}{k}$. □

As in Theorem 2.2 we get bounds on the sample size that guarantee learnability. Note that the first bound in the max expression is exactly the bound proven in [BEHW87] which is the case where the kernel is empty.

Theorem 3.2:

Any compression scheme with kernel-size k and additional information Q produces with probability at least $1-\delta$ a hypothesis with error at most ϵ when given a sample of size

$$m \geq \max(\frac{2}{\epsilon}(\ln(\frac{1}{\delta}) + \ln(|Q|)), \frac{4k}{\epsilon}\ln(\frac{4k}{\epsilon}) + 2k).$$

This holds for arbitrary ϵ and δ. □

In the example of learning n orthogonal rectangles in E^2 with kernel size $4n$ the set Q has cardinality $4n!$ and the above bound is at most $\max(\frac{2}{\epsilon}(\ln(\frac{1}{\delta}) + 4n\ln(4n)), \frac{4k}{\epsilon}\ln(\frac{4k}{\epsilon}) + 2k)$. Again our bound compares favorably to the bounds of [BEHW86]: $m \geq \max\left(\frac{4}{\epsilon}\log\frac{2}{\delta}, \frac{8d}{\epsilon}\log\frac{8d}{\epsilon}\right)$, where $d = 8n\log(4n)$ for this example (see [HW87] for how to estimate the dimension).

4 DEPENDENCE OF THE KERNEL SIZE ON THE CONCEPT

To keep the notation simple we assumed that the kernel always has fixed size. In many cases however the kernel size might depend on the sample size and on the concept that is learned. It can be verified easily that our proofs hold for that case as well.

We present an example of ([BEHW86]) in which the kernel size depends on the concept and an improved learning algorithm for the example in which the kernel size also depends on the sample size. Earlier it was mentioned that the union of n orthogonal rectangles can be represented with a kernel of size $4n$ plus some finite information, thus demonstrating the learnability of such a concept. If our concept class consists of arbitrary unions of rectangles, then no bounded kernel size will suffice for all concepts in the class. But by allowing the kernel size to depend on the concept (the number of rectangles in the union), we can find a data compression scheme for this class. In this case, this is a demonstration of learnability but not of polynomial learnability, since it is NP-hard to find the smallest number of rectangles which interprets a sample [M78]. To get polynomial learnability, we can use a polynomial approximation ([J74], [N69]) to the minimum cover. The approximation algorithm finds a consistent hypothesis using a union of $n\log m$ rectangles in polynomial time. A kernel of size $4n\log m$ plus some finite information suffices to represent this hypothesis.

The kernel size now depends on the sample size. The appropriate polynomial bounds on the sample size follow from Theorem 3.1.

Theorem 4.1:

For a compression scheme with kernel size pm^α $(0 \leq \alpha < 1)$ and additional information Q, where $|Q| \leq qm^\gamma$, the error is at most ϵ with probability less than $qm^\gamma \binom{m}{pm^\alpha}(1-\epsilon)^{m-pm^\alpha}$ when given a sample of size $m \geq pm^\alpha$. Here p, α, q, and γ are fixed for the concept class C.

5 RELAXING THE CONSTISTENCY CONSTRAINT

We will now relax Condition (C2) which asserts that ρ must be consistent with the m-sample. In practice it might be hard to find polynomial algorithms ρ that are consistent with all m samples. But there might be polynomial algorithms that are consistent with a large portion of the samples. The question is how many samples may be missed (we denote this number by l) and still assure learnability.

(C2') For any $\xi \in C$, any m, any $\bar{x} \in X^m$, and any point x_i of $\bar{x}$, $\rho(\kappa(L_\xi(\bar{x})), x_i) = \xi(x_i)$ holds *for all except for* l of the m points x_i.

Theorem 4.1:

For a compression scheme with kernal size k that misses at most l points the error is at most ϵ with probability less than $\binom{k+l}{k}\binom{m}{k+l}(1-\epsilon)^{m-k-l}$ when given a sample of size $m \geq k+l$.

Proof: In the proof of theorems 2.1 and 3.1 we did not use the fact that $\rho(\kappa(L_\xi(\bar{x}), y) = \xi(y)$ if y is in the kernal $\kappa(L_\xi(\bar{x}))$. We only needed that ρ is consistent with ξ for all points outside of the kernal. See definition of $B_{\bar{t}}$ in both proofs. We will use this by encorporating the l inconsistent samples into the kernal.

Let ρ, κ be a compression scheme with kernal size k that is inconsistent with at most l elements. From this we construct a related compression scheme with additional information ρ', κ', Q with kernal size $k+l$ that is consistent with all points of the sample except with some of the points of the kernal. κ' simply applies κ to $\bar{x}$ and then scans the sample with ρ to determine which of the m samples are not predicted correctly. The new kernal of κ' will be the kernal of κ plus some l samples on which ρ might not predict correctly. The additional information Q is used to specify the original kernal of size k among the new kernal of size $k+l$. Q consists of all bitmasks of length $k+l$ in which exactly k bits are one. ρ' scans the kernal of size $k+l$ removes the l points that were not in the original kernal. ρ' them applies ρ.

It is easy to see that the two compresssion schemes predict the same function values. In particular their error is the same. We thus can apply the modified proof of Theorem 2.1 to the $\rho', \kappa', k+l$ scheme and the bound of the theorem follows. Note that $|Q| = \binom{k+l}{k}$. $\square$

6 ERRORS IN THE SAMPLES

7 INTRODUCING COMPLEXITY

8 DISCUSSION AND OPEN PROBLEMS

Our proof of learnability for compression schemes with fixed kernel size is much shorter than the proof in [BEHW86]. On the other hand there they are able to show that their condition of fixed Vapnik-Chervonenkis dimension is also necessary for learnability. For our scheme we show sufficiency, but necessity remains an open question. Are there concept classes with finite dimension for which there is no scheme with bounded kernel size and bounded additional information?

Our compression can be compared with compression implicit in [BEHW87] There one compresses the sample to a fixed number of bits which encode a consistent hypothesis. In contrast in our scheme we compress to a fixed size subsample, the points of which might be given with arbitrarily high precision. One way to gain an understanding of the relation between these two approaches is to compare the bounds produced for a case to which both apply. For example, suppose the class of concepts to be learned is subintervals $[0, c]$ of the interval $[0, 1)$. Suppose further that our domain contains only the finite subset of $[0, 1)$ which can be represented with binary fractions of b-bits, for some b. Then there are 2^b possible concepts. We can represent any hypothesis with b bits, and a single b-bit number will give us enough information to reconstruct any sample. This is sufficient, using the argument of [BEHW 86a] (or Theorem 3.1 for $k = 0$), to guarantee that we can learn with sample size

$$m > \frac{1}{\epsilon}(b + \ln(\frac{1}{\delta})).$$

With our data compression scheme, this concept class can be learned with a kernel of size one. By Theorem 2.1 it suffices to take a sample of size

$$m \geq \frac{1}{\epsilon}(4 \ln(\frac{4}{\epsilon}) + \ln(\frac{1}{\delta})) + 2.$$

Note that, unlike the first bound, this bound is independent of the precision b with which we represent the points of the interval. Clearly the combinatorial complexity of this example is captured by the fact that one can compress down to one point. The precision of the point is a side issue. Note that the Vapnik-Chervonenkis dimension of the concept class of the example is one. Thus classifying learnability with this dimension also avoids the issue of the precision.

The paradigm of the compression scheme is simple enough that it can be extended in various ways. It is the aim of this paper to introduce the basic scheme. In our further research we first relax Condition C2 which required that ρ be consistent with the sample when given the kernel as an input. In practice, it might be hard to find compression schemes that guarantee consistency with

the whole sample. We explore bounds on how much of the sample can be missed by ρ for the class to remain learnable.

Secondly we address the case where the sample is not reliable. We study the relation between the amount of error and the speed and confidence with which we can learn.

If errors are modeled probabilistically, this leads one toward considering the case where the learnability or the speed of learning depends on the underlying probability distribution. One step in doing this is to relax the requirement that the bounds on the sample size be independent of the underlying probability distribution. Certain concept classes, which do not necessarily have finite Vapnik-Chervonenkis dimension, become learnable under this broader definition of learnability. The data compression scheme can still be applied if one uses the extended scheme which requires only partial consistency with the sample. Even if we require that concepts remain learnable for arbitrary distributions, the speed of learning may now be heavily dependent on the distribution. Since the underlying distribution is unknown, our bounds would no longer give practical *apriori* information on the sample size needed to learn with a desired degree of confidence. Thus in this case one might wish to use empirical tests to estimate the required sample size.

The function value of some observations might have been changed. We consider the case where these changes are made probabilistically or by an adversary and determine how much of the sample can be changed and still have learnability.

REFERENCES

[BEHW86] Blumer, A., A. Ehrenfeucht, D. Haussler and M. Warmuth, "Classifying Learnable Geometric Concepts with the Vapnik-Chervonenkis Dimension," Preceedings of the *Eighteenth Annual ACM Symposium on Theory of Computing,* Berkeley, May 28-30, 1986, pp. 273-282.

[BEHW87] Blumer, A., A. Ehrenfeucht, D. Haussler and M. Warmuth, "Occam's Razor," *Information Processing Letters 24*, 1987, pp. 377-380.

[BL86] Blumer, A., and N. Littlestone, "Learning Faster Than Promised by the Vapnik-Chervonenkis Dimension," unpublished manuscript.

[HW87] Haussler, D. and E. Welzl, "Epsilon-nets and range queries," *Discrete Computational Geometry* 2, 1987, pp. 373-395.

[J74] Johnson, D.S., "Approximation Algorithms for combinatorial problems," Journal of Computer and Systems Sciences, Vol. 9, 1974.

[M78] Masek, W.J., "Some NP-Complete Set Cover Problems," MIT Laboratory for Computer Science, unpublished manuscript.

[N69] Nigmatullin, R.G., "The Fastest Descent Method for Covering Problems (in Russian)," Proceedings of a Symposium on Questions of Precision and Efficiency of Computer Algorithms," Book 5, Kiev, 1969, pp. 116-126.

[R74] Rudin, W., Real and Complex Analysis, McGraw-Hill Series in Higher Mathematics, 1974.

[V84] 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. 264-80.