JOURNAL OF COMPLEXITY **13,** 509–544 (1998)
ARTICLE NO. CM970459

On the Value of Partial Information for Learning from Examples

Joel Ratsaby*

Department of Electrical Engineering, Technion, Haifa, 32000 Israel

and

Vitaly Maiorov†

Department of Mathematics, Technion, Haifa, 32000 Israel

Received August 15, 1996

The PAC model of learning and its extension to real valued function classes provides a well-accepted theoretical framework for representing the problem of learning a target function $g(x)$ using a random sample $\{(x_i,\ g(x_i))\}_{i=1}^m$. Based on the uniform strong law of large numbers the PAC model establishes the sample complexity, i.e., the sample size m which is sufficient for accurately estimating the target function to within high confidence. Often, in addition to a random sample, some form of prior knowledge is available about the target. It is intuitive that increasing the amount of information should have the same effect on the error as increasing the sample size. But quantitatively how does the rate of error with respect to increasing information compare to the rate of error with increasing sample size? To answer this we consider a new approach based on a combination of information-based complexity of Traub *et al.* and Vapnik–Chervonenkis (VC) theory. In contrast to VC-theory where function classes of finite pseudo-dimension are used only for statistical-based estimation, we let such classes play a dual role of functional estimation as well as approximation. This is captured in a newly introduced quantity, $\rho_d(\mathcal{F})$, which represents a nonlinear width of a function class $\mathcal{F}$. We then extend the notion of the nth minimal radius of information and define a quantity $I_{n,d}(\mathcal{F})$ which measures the minimal approximation error of the worst-case target $g \in \mathcal{F}$ by the family of function classes having pseudo-dimension d given partial information on g consisting of values taken by n linear operators. The error rates are calculated which leads to a quantitative notion of the value of partial information for the paradigm of learning from examples.

*E-mail: jer@ee.technion.ac.il. All correspondence should be mailed to this author.
†E-mail: maiorov@tx.technion.ac.il.

0885-064X/97 $25.00

1. INTRODUCTION

The problem of machine learning using randomly drawn examples has received in recent years a significant amount of attention while serving as the basis of research in what is known as the field of computational learning theory. Valiant [35] introduced a learning model based on which many interesting theoretical results pertaining to a variety of learning paradigms have been established. The theory is based on the pioneering work of Vapnik and Chervonenkis [36–38] on finite sample convergence rates of the uniform strong law of large numbers (SLN) over classes of functions. In its basic form it sets a framework known as the probably approximately correct (PAC) learning model. In this model an abstract teacher provides the learner a finite number m of i.i.d. examples $\{(x_i, g(x_i))\}_{i=1}^{m}$ randomly drawn according to an unknown underlying distribution P over X, where g is the target function to be learnt to some prespecified arbitrary accuracy $\epsilon > 0$ (with respect to the $L_1(P)$-norm) and confidence $1 - \delta$, where $\delta > 0$. The learner has at his discretion a functional class referred to as the hypothesis class $\mathcal{H}$ from which he is to determine a function $\hat{h}$, sample-dependent, which estimates the unknown target g to within the prespecified accuracy and confidence levels.

There have been numerous studies and applications of this learning framework to different learning problems (Kearns and Vazirani [18], Hanson *et al.* [15]). The two main variables of interest in this framework are the sample complexity which is the sample size sufficient for guaranteeing the prespecified performance and the computational complexity of the method used to produce the estimator hypothesis $\hat{h}$.

The bulk of the work in computational learning theory and, similarly, in the classical field of pattern recognition, treats the scenario in which the learner has access *only* to randomly drawn samples. It is often the case, however, that some additional knowledge about the target is available through some form of *a priori* constraints on the target function g. In many areas where machine learning may be applied there is a source of information, sometimes referred to as an oracle or an expert, which supplies random examples and even more complex forms of partial information about the target. A few instances of such learning problems include: (1) *pattern classification.* Credit card fraud detection where a tree classifier (Devroye *et al.* [12]) is built from a training sample consisting of patterns of credit card usage in order to learn to detect transactions that are potentially fraudulent. Partial information may be represented by an existing tree which is based on human-expert knowledge. (2) *prediction and financial analysis.* Financial forecasting and portfolio management where an artificial neural network learns from time-series data and is given rule-based partial knowledge translated into constraints on the weights of the neuron elements. (3) *control and optimization.* Learning a control process for industrial manufacturing where partial information represents quantitative physical constraints on the various machines and their operation.

For some specific learning problems the theory predicts that partial knowledge is very significant, for instance, in statistical pattern classification or in density estimation, having some knowledge about the underlying probability distributions may crucially influence the complexity of the learning problem (cf. Devroye [11]). If the distributions are known to be of a certain parametric form an exponentially large savings in sample size may be obtained (cf. Ratsaby [28], Ratsaby and Venkatesh [30, 31]). In general, partial information may appear as knowledge about certain properties of the target function. In parametric-based estimation or prediction problems, e.g., maximum likelihood estimation, knowledge concerning the unknown target may appear in terms of a geometric constraint on the Euclidean subset that contains the true unknown parameter. In problems of pattern recognition and statistical regression estimation, often some form of a criterion functional over the hypothesis space is defined. For instance, in artificial neural networks, the widely used back-propagation algorithm (cf. Ripley [32]) implements a least-squared-error criterion defined over a finite-dimensional manifold spanned by ridge-functions of the form $\sigma(a^T x + b)$, where $\sigma(y) = 1/(1 + e^{-y})$. Here prior knowledge can take the form of a constraint added on to the minimization of the criterion. In Section 3 we provide further examples where partial information is used in practice.

It is intuitive that general forms of prior partial knowledge about the target and random sample data are both useful. PAC provides the complexity of learning in terms of the sample sizes that are sufficient to obtain accurate estimation of g. Our motive in this paper is to study the complexity of learning from examples while being given prior partial information about the target. We seek the value of partial information in the PAC learning paradigm. The approach taken here is based on combining frameworks of two fields in computer science, the first being information-based complexity (cf. Traub *et al.* [34]) which provides a representation of partial information while the second, computational learning theory, furnishes the framework for learning from random samples.

The remainder of this paper is organized as follows: In Section 2 we briefly review the PAC learning model and Vapnik–Chervonenkis theory. In Section 3 we provide motivation for the work. In Section 4 we introduce a new approximation width which measures the degree of nonlinear approximation of a functional class. It joins elementary concepts from Vapnik–Chervonenkis theory and classical approximation theory. In Section 5 we briefly review some of the definitions of information-based complexity and then introduce the minimal information-error $I_{n,d}(\mathcal{F})$. In Section 6 we combine the PAC learning error with the minimal partial information error to obtain a unified upper bound on the error. In Section 7 we compute this upper bound for the case of learning a Sobolev target class. This yields a quantitative trade-off between partial information and sample size. We then compute a lower bound on the minimal partial information error for the Sobolev class which yields an almost optimal information operator. The Appendix includes the proofs of all theorems in the paper.

2. OVERVIEW OF THE PROBABLY APPROXIMATELY CORRECT LEARNING MODEL

Valiant [35] introduced a new complexity-based model of learning from examples and illustrated this model for problems of learning indicator functions over the boolean cube $\{0, 1\}^n$. The model is based on a probabilistic framework which has become known as the *probably approximately correct*, or PAC, model of learning. Blumer *et al.* [6] extended this basic PAC model to learning indicator functions of sets in Euclidean $\mathbb{R}^n$. Their methods are based on the pioneering work of Vapnik and Chervonenkis [36] on finite sample convergence rates of empirical probability estimates, independent of the underlying probability distribution. Haussler [16] has further extended the PAC model to real and vector-valued functions which is applicable to general statistical regression, density estimation and classification learning problems. We start with a description of the basic PAC model and some of the relevant results concerning the complexity of learning.

A *target class* $\mathcal{F}$ is a class of Borel measurable functions over a domain X containing a *target function* g which is to be learnt from a *sample* $z^m = \{(x_i, g(x_i))\}_{i=1}^m$ of m examples that are randomly drawn i.i.d. according to *any* fixed probability distribution P on X. Define by $S_{\mathcal{F}}$ the *sample space* for $\mathcal{F}$ which is the set of all samples of size m over all functions $f \in \mathcal{F}$ for all $m \geq 1$. Fix a *hypothesis class* $\mathcal{H}$ of functions on X which need not be equal nor contained in $\mathcal{F}$. A *learning algorithm* ϕ: $S_{\mathcal{F}} \to \mathcal{H}$ is a function that, given a large enough randomly drawn sample of any target in $\mathcal{F}$, returns a Borel measurable function h (a *hypothesis*) which is with high probability a good approximation of the target function g.

Associated with each hypothesis h, is a nonnegative *error* value $L(h)$, which measures its disagreement with the target function g on a randomly drawn example and an *empirical error* $L_m(h)$, which measures the disagreement of h with g averaged over the observed m examples. Note that the notation of $L(h)$ and $L_m(h)$ leaves the dependence on g and P implicit.

For the special case of $\mathcal{F}$ and $\mathcal{H}$ being classes of *indicator functions* over sets of $X = \mathbb{R}^n$ the *error* of a hypothesis h is defined to be the probability (according to P) of its symmetric difference with the target g; i.e.,

$$L(h) = P(\{x \in \mathbb{R}^n: g(x) \neq h(x)\}). \tag{1}$$

Correspondingly, the *empirical error* of h is defined as

$$L_m(h) = \frac{1}{m} \sum_{i=1}^{m} 1_{\{g(x_i) \neq h(x_i)\}}, \tag{2}$$

where $1_{\{x \in A\}}$ stands for the indicator function of the set A. For *real-valued function* classes $\mathcal{F}$ and $\mathcal{H}$ the error of a hypothesis h is taken as the expectation $\mathrm{E}l(h, g)$ (with respect to P) of some positive real-valued *loss* function $l(h, g)$, e.g., quadratic loss $l(h, g) = (h(x) - g(x))^2$ in regression estimation, or the log likelihood loss $l(h, g) = \ln(g(x)/h(x))$ for density estimation. Similarly, the empirical error now becomes the average loss over the sample, i.e., $L_m(h) = (1/m)\sum_{i=1}^{m} l(h(x_i), g(x_i))$.

We now state a formal definition of a learning algorithm which is an extension of a definition in Blumer *et al.* [6].

DEFINITION 1 (PAC-learning algorithm). Fix a target class $\mathcal{F}$, a hypothesis class $\mathcal{H}$, a loss function $l(\cdot, \cdot)$, and any probability distribution P on X. Denote by P^m the m-fold joint probability distribution on X^m. A function ϕ is a *learning algorithm* for $\mathcal{F}$ with respect to P with sample size $m \equiv m(\epsilon, \delta)$ if for all $\epsilon > 0$, $0 < \delta < 1$, for any fixed target $g \in \mathcal{F}$, with probability $1 - \delta$, based on a randomly drawn sample z^m, the hypothesis $\hat{h} = \phi(z^m)$ has an error $L(\hat{h}) \leq L(h^*) + \epsilon$, where h^* is an optimal hypothesis; i.e., $L(h^*) = \inf_{h \in \mathcal{H}} L(h)$. Formally, this is stated as: $P^m(z^m \in X^m\colon L(\hat{h}) > L(h^*) + \epsilon) \leq \delta$.

The smallest sample size $m(\epsilon, \delta)$ such that there exists a learning algorithm ϕ for $\mathcal{F}$ with respect to *all* probability distributions is called the *sample complexity* of ϕ or simply the sample complexity for learning $\mathcal{F}$ by $\mathcal{H}$. If such a ϕ exists then $\mathcal{F}$ is said to be uniformly learnable by $\mathcal{H}$. We note that in the case of real-valued function classes the sample complexity depends on the error function through the particular loss function used.

Algorithms ϕ which output a hypothesis $\hat{h}$ that minimizes $L_m(h)$ over all $h \in \mathcal{H}$ are called *empirical risk minimization* (ERM) algorithms (cf. Vapnik [38]). The theory of uniform learnability for ERM algorithms forms the basis for the majority of the works in the field of computational learning theory, primarily for the reason that the sample complexity is directly related to a capacity quantity called the *Vapnik–Chervonenkis dimension* of $\mathcal{F}$ for the case of an indicator function class $\mathcal{F}$, or to the *pseudo-dimension* in case of a real-valued function class $\mathcal{F}$. These two quantities are defined and discussed below. Essentially the theory says that if the capacity of $\mathcal{F}$ is finite then $\mathcal{F}$ is uniformally learnable. We note that there are some pedagogic instances of functional classes, even of infinite pseudo-dimension, for which any target function can be exactly learnt by a *single* example of the form $(x, g(x))$ (cf. Bartlett *et al.*, p. 299). For such target classes the sample complexity of learning by ERM is significantly greater than one so ERM is not an efficient form of learning. Henceforth all the results are limited to ERM learning algorithms.

We start with the following definition.

DEFINITION 2 (Vapnik–Chervonenkis dimension). Given a class $\mathcal{H}$ of indicator functions of sets in X the Vapnik–Chervonenkis dimension of $\mathcal{H}$, denoted as $\mathrm{VC}(\mathcal{H})$, is defined as the largest integer m such that there exists a sample $x^m = \{x_1, \ldots, x_m\}$ of points in X such that the cardinality of the set of boolean vectors $S_{x^m}(\mathcal{H}) = \{[h(x_1), \ldots, h(x_m)]\colon h \in \mathcal{H}\}$ satisfies $|S_{x^m}(\mathcal{H})| = 2^m$. If m is arbitrarily large then the VC-dimension of $\mathcal{H}$ is infinite.

Remark. The quantity $\max_{x^m} |S_{x^m}(\mathcal{H})|$, where the maximum is taken over all possible m-samples, is called the *growth function* of $\mathcal{H}$.

EXAMPLE. Let $\mathcal{H}$ be the class of indicator functions of interval sets on $X = \mathbb{R}$. With a single point $x_1 \in X$ we have $|\{[h(x_1)]\colon h \in \mathcal{H}\}| = 2$. For two points $x_1, x_2 \in X$ we have $|\{[h(x_1), h(x_2)]\colon h \in \mathcal{H}\}| = 4$. When $m = 3$, for any points $x_1, x_2, x_3 \in X$ we have $|\{[h(x_1), h(x_2), h(x_3)]\colon h \in \mathcal{H}\}| < 2^3$ thus $\mathrm{VC}(\mathcal{H}) = 2$.

The main interest in the VC-dimension quantity is due to the following result on a uniform strong law of large numbers which is a variant of Theorem 6.7 in Vapnik [38].

LEMMA 1 (Uniform SLN for the indicator function class). *Let g be any fixed target indicator function and let $\mathcal{H}$ be a class of indicator functions of sets in X with* $\mathrm{VC}(\mathcal{H}) = d < \infty$. *Let $z^m = \{(x_i, g(x_i))\}_{i=1}^m$ be a sample of size $m > d$ consisting of randomly drawn examples according to any fixed probability distribution P on X. Let $L_m(h)$ denote the empirical error for h based on z^m and g as defined in* (2). *Then for arbitrary confidence parameter* $0 < \delta < 1$, *the deviation between the empirical error and the true error uniformly over $\mathcal{H}$ is bounded as*

$$\sup_{h\in\mathcal{H}} |L(h) - L_m(h)| \le 4\sqrt{\frac{d(\ln(2m/d) + 1) + \ln(9/\delta)}{m}}$$

with probability $1 - \delta$.

Remark. The result actually holds more generally for a boolean random variable $y \in Y = \{0, 1\}$ replacing the deterministic target function $g(x)$. In such a case the sample consists of random pairs $\{(x_i, y_i)\}_{i=1}^m$ distributed according to any fixed joint probability distribution P over $X \times Y$.

Thus a function class of finite VC-dimension possesses a certain statistical smoothness property which permits simultaneous error estimation over all hypotheses in $\mathcal{H}$ using the empirical error estimate. We note in passing that there is an interesting generalization (cf. Buescher and Kumar [7], Devroye *et al.* [12]) of the empirical error estimate to other smooth estimators based on the idea of empirical coverings which removes the condition of needing a finite VC-dimension.

As a direct consequence of Lemma 1 we obtain the necessary and sufficient conditions for a target class $\mathcal{F}$ of indicator functions to be uniformly learnable by a hypothesis class $\mathcal{H}$. This is stated next and is a slight variation of Theorem 2.1 in Blumer *et al.* [6].

LEMMA 2 (Uniform learnability of indicator function class). *Let $\mathcal{F}$ and $\mathcal{H}$ be a target class and a hypothesis class, respectively, of* indicator functions *of sets in X. Then $\mathcal{F}$ is uniformly learnable by $\mathcal{H}$ if and only if the* $\mathrm{VC}(\mathcal{H}) < \infty$. *Moreover, if* $\mathrm{VC}(\mathcal{H}) = d$, *where $d < \infty$, then for any* $0 < \epsilon$, $\delta < 1$, *the sample complexity of an algorithm ϕ is bounded from above by* $c((d/\epsilon)\log(1/\delta))$, *for some absolute constant $c > 0$.*

We proceed now to the case of real-valued functions. The next definition which generalizes the VC-dimension is taken from Haussler [16] and is based on the work of Pollard [27]. Let sgn(y) be defined as 1 for $y > 0$ and −1 for $y \leq 0$. For a Euclidean vector $v \in \mathbb{R}^m$ denote by $\mathrm{sgn}(v) = [\mathrm{sgn}(v_1), \ldots, \mathrm{sgn}(v_m)]$.

DEFINITION 3 (Pseudo-dimension). Given a class $\mathcal{H}$ of real-valued functions defined on X. The pseudo-dimension of $\mathcal{H}$, denoted as $\dim_p(\mathcal{H})$, is defined as the largest integer m such that there exists $\{x_1, \ldots, x_m\} \in X$ and a vector $v \in \mathbb{R}^m$ such that the cardinality of the set of boolean vectors satisfies $|\{\mathrm{sgn}[h(x_1)+v_1, \ldots, h(x_m)+v_m]: h \in \mathcal{H}\}| = 2^m$. If m is arbitrarily large then the $\dim_p(\mathcal{H}) = \infty$.

The next lemma appears as Theorem 4 in Haussler [16] and states that for the case of finite-dimensional vector spaces of functions the pseudo-dimension equals its dimension.

LEMMA 3. *Let $\mathcal{F}$ be a d-dimensional vector space of functions from a set X into $\mathbb{R}$. Then* $\dim_p(\mathcal{F}) = d$.

For several useful invariance properties of the pseudo-dimension cf. Pollard [27] and Haussler [16, Theorem 5].

The main interest in the pseudo-dimension arises from having the SLN hold uniformly over a real-valued function class if it has a finite pseudo-dimension. In order to apply this to the PAC-framework we need a uniform SLN result not for the hypothesis class $\mathcal{H}$ but for a class defined by $\mathcal{L}_{\mathcal{H}} = \{l(h(x), y): h \in \mathcal{H}, x \in X, y \in \mathbb{R}\}$ for some fixed loss function l, since an ERM-based algorithm minimizes the empirical error, i.e., $L_m(h)$, over $\mathcal{H}$. While the theory presented in this paper applies to general loss functions we restrict here to the absolute-loss $l(h(x), g(x)) = |h(x) - g(x)|$. The next lemma is a variant of Theorem 7.3 of Vapnik [38].

THEOREM 1. *Let P be any probability distribution on X and let $g \in \mathcal{F}$ be a fixed target function. Let $\mathcal{H}$ be a class of functions from X to $\mathbb{R}$ which has a pseudo-dimension $d \geq 1$ and for any $h \in \mathcal{H}$ denote by $L(h) = E|h(x) - g(x)|$ and assume $L(h) \leq M$ for some absolute constant $M > 0$. Let* $\{(x_i, g(x_i))\}_{i=1}^{m}$,

$x_i \in X$, be an i.i.d. sample of size $m > 16(d+1)\log^2 4(d+1)$ drawn according to P. Then for arbitrary $0 < \delta < 1$, simultaneously for every function $h \in \mathcal{H}$, the inequality

$$|L(h) - L_m(h)| \leq 4M\sqrt{\frac{16(d+1)\ \log^2\ 4(d+1)(\ln(2m)+1) + \ln(9/\delta)}{m}} \qquad (3)$$

holds with probability $1 - \delta$.

The theorem is proved in Section A.1.

Remark. For uniform SLN results based on other loss functions see Theorem 8 of Haussler [16].

We may take twice the right-hand side of (3) to be bounded from above by the simpler expression

$$\epsilon(m,\ d,\ \delta) \equiv c_1\sqrt{\frac{d\ \log^2\ d\ \ln m + \ln(1/\delta)}{m}} \qquad (4)$$

for some absolute constant $c_1 > 0$. Being that an ERM algorithm picks a hypothesis $\hat{h}$ whose empirical error satisfies $L_m(\hat{h}) = \inf_{h\in\mathcal{H}} L_m(h)$ and by Definition 1, $L(h^*) = \inf_{h\in\mathcal{H}} L(h)$, it follows that

$$\begin{aligned} L(\hat{h}) &\leq L_m(\hat{h}) + \frac{\epsilon(m,\ d,\ \delta)}{2} \\ &\leq L_m(h^*) + \frac{\epsilon(m,\ d,\ \delta)}{2} \\ &\leq L(h^*) + \epsilon(m,\ d,\ \delta). \end{aligned} \qquad (5)$$

By (5) and according to Definition 1 it is immediate that ERM may be considered as a PAC learning algorithm for $\mathcal{F}$. Thus we have the following lemma concerning the *sufficient* condition for uniform learnability of a real-valued function class.

LEMMA 4 (Uniform learnability of real-valued function class). *Let $\mathcal{F}$ and $\mathcal{H}$ be the target and hypothesis classes of real-valued functions, respectively, and let P be any fixed probability distribution on X. Let the loss function $l(g(x),\ h(x)) = |g(x) - h(x)|$ and assume $L(h) \leq M$ for all $h \in \mathcal{H}$, and $g \in \mathcal{F}$, for some absolute constant $M > 0$. If $\dim_p(\mathcal{H}) < \infty$ then $\mathcal{F}$ is uniformly learnable by $\mathcal{H}$. Moreover, if $\dim_p(\mathcal{H}) = d < \infty$ then for any $\epsilon > 0$, $0 < \delta < 1$, the sample complexity of learning $\mathcal{F}$ by $\mathcal{H}$ is bounded from above by $(cM^2 d\ \ln^2(d)/\epsilon^2)(\ln(dM/\epsilon) + \ln(1/\delta))$, for some absolute constant $c > 0$.*

Remarks. As in the last remark above, this result can be extended to other loss functions *l*. In addition, Alon *et al.* [4] recently showed that a quantity called the *scale-sensitive dimension* which is a generalization of the pseudo-dimension, determines the *necessary and sufficient* condition for uniform learnability.

It is also worth noting that there have been several works related to the pseudo-dimension but which are used for mathematical analysis other than learning theory. As far as we are aware, Warren [39] was the earliest who considered a quantity called the number of connected components of a nonlinear manifold of real-valued functions, which closely resembles the growth function of Vapnik and Chervonenkis for set-indicator functions, see Definition 2. Using this he determined lower bounds on the degree of approximation by certain nonlinear manifolds. Maiorov [20] calculated this quantity and determined the degree of approximation for the nonlinear manifold of ridge functions which include the manifold of functions represented by artificial neural networks with one hidden layer. Maiorov, Meir, and Ratsaby [21], extended his result to the degree of approximation measured by a probabilistic (n, δ)-width with respect to a uniform measure over the target class and determined finite sample complexity bounds for model selection using neural networks [29]. For more works concerning probabilistic widths of classes see Traub *et al.* [34], Maiorov and Wasilkowski [22].

Throughout the remainder of the paper we will deal with learning real-valued functions while denoting explicitly a hypothesis class $\mathcal{H}^d$ as one which has $\dim_p(\mathcal{H}^d) = d$. For any probability distribution P and target function g, the error and empirical error of a hypothesis h are defined by the $L_1(P)$-metric as

$$L(h) = E|h(x) - g(x)|, \qquad L_m(h) = \frac{1}{m}\sum_{i=1}^{m} |h(x_i) - g(x_i)|, \tag{6}$$

respectively.

We discuss next some practical motivation for our work.

3. MOTIVATION FOR A THEORY OF LEARNING WITH PARTIAL INFORMATION

It was mentioned in Section 1 that the notion of having partial knowledge about a solution to a problem, or more specifically about a target function, is often encountered in practice. Starting from the most elementary instances of learning in humans it is almost always the case that a learner begins with some partial information about the problem. For instance, in learning cancer diagnosis, a teacher not only provides examples of pictures of healthy cells and benign cells but also descriptive partial information such as “a benign cell has color black and elongated shape,” or “benign cells usually appear in clusters.” Similarly, for machine learning it is intuitive that partial information must be useful.

While much of the classical theory of pattern recognition (Duda and Hart [13], Fukunaga [14]) and the more recent theory of computational learning (Kearns and Vazirani [18]) and neural networks (Ripley [32]) focus on learning from randomnly drawn data, there has been an emergence of interest in nonclassical forms of learning, some of which indicates that partial information in various forms which depend on the specific application is useful in practice. This is related to the substream known as *active learning*, where the learner participates actively by various forms of querying to obtain information from the teacher. For instance, the notion of selective sampling (cf. Cohn *et al.* [8]) permits the learner to query for samples from domain-regions having high classification uncertainty. Cohn [9] uses methods based on the theory of optimal experiment design to select data in an on-line fashion with the aim of decreasing the variance of an estimate. Abu-Mostafa [1–3] refers to partial information as *hints* and considers them for financial prediction problems. He shows that certain types of hints which reflect invariance properties of the target function g, for instance saying that $g(x) = g(x')$, at some points x, x' in the domain, may be incorporated into a learning error criterion.

In this paper we adopt the framework of information-based complexity (cf. Traub *et al.* [34]) to represent partial information. In the framework whose basic definitions are reviewed in Section 5, we limit to linear information comprised of n linear functionals $L_i(g)$, $1 \leq i \leq n$, operating on the target function g. In order to motivate the interest in partial information as being given by such n-dimensional linear operators we give the following example of learning pattern classification using a classical nonparametric discriminant analysis method (cf. Fukunaga [14]).

The field of pattern recognition treats a wide range of practical problems where an accurate decision is to be made concerning a stochastic pattern which is in the form of a multidimensional vector of features of an underlying stochastic information source, for instance, deciding which of a finite number of types of stars corresponds to given image data taken by an exploratory spacecraft, or deciding which of the words in a finite dictionary correspond to given speech data which consist of spectral analysis information on a sound signal. Such problems have been classically modeled according to a statistical framework where the input data are stochastic and are represented as random variables with a probability distribution over the data space. The most widely used criterion for learning pattern recognition (or classification) is the misclassification probability on randomly chosen data which have not been seen during the training stage of learning. In order to ensure an accurate decision it is necessary to minimize this criterion. The optimal decision rule is one which achieves the minimum possible misclassification probability and has been classically referred to as Baye’s decision rule.

We now consider an example of learning pattern recognition using randomly drawn examples, where partial information takes the form of feature extraction.

EXAMPLE (Learning pattern classification). The setting consists of M pattern classes represented by *unknown* nonparametric class conditional probability density functions $f(x|j)$ over $X = \mathbb{R}^l$ with known correpsonding *a priori* class probabilities p_j, $1 \leq j \leq M$. It is well known that the optimal Bayes classifier which has the minimal misclassification probability is defined as follows: $g(x) = \mathrm{argmax}_{1 \leq j \leq M}\{p_j f(x|j)\}$, where $\mathrm{argmax}_{j \in A} B_j$ denotes any element j in A such that $B_j \geq B_i$, $j \neq i$. Its misclassification probability is called the *Bayes error*. For instance, suppose that $M = 2$ and $f(x/j)$, $j = 1, 2$, are both l-dimensional Gaussian probability density functions. Here the two pattern classes clearly overlap as their corresponding functions $f(x|1)$ and $f(x|2)$ have an overlapping probability-1 support; thus the optimal Bayes misclassification probability must be greater than zero. The Bayes classifier in this case is an indicator function over a set $A = \{x \in \mathbb{R}^l : q(x) > 0\}$, where $q(x)$ is a second degree polynomial over $\mathbb{R}^l$. We henceforth let the *target* function, denoted by $g(x)$, be the Bayes classifier and note that it may not be unique.

The *target class* $\mathcal{F}$ is defined as a rich class of classifiers each of which maps X to $\{1, \ldots, M\}$. The training *sample* consists of m i.i.d. pairs $\{(x_i, y_i)\}_{i=1}^m$, where $y_i \in \{1, 2, \ldots, M\}$ takes the value j with probability p_j, and x_i is drawn according to the probability distribution corresponding to $f(x|y_i)$, $1 \leq i \leq m$. The learner has a *hypothesis* class $\mathcal{H}$ of classifier functions mapping X to $\{1, \ldots, M\}$ which has a finite pseudo-dimension d.

Formally, the *learning problem* is to approximate g by a hypothesis h in $\mathcal{H}$. The *error* of h is defined as $L(h) = \|h - g\|_{L_1(P)}$, where P is some fixed probability distribution over $\mathcal{X}$. Stated in the PAC-framework, a target class $\mathcal{F}$ is to be uniformly learned by $\mathcal{H}$; i.e., for any fixed target $g \in \mathcal{F}$ and any probability distribution P on X, find an $\hat{h} \in \mathcal{H}$ which depends on g and whose error $L(\hat{h}) \leq L(h^*) + \epsilon$ with probability $1 - \delta$, where $L(h^*) = \inf_{h \in \mathcal{H}} \|g - h\|_{L_1(P)}$.

As *partial information* consider the ubiquitous method of *feature extraction* which is described next. In the pattern classification paradigm it is often the case that, based on a given sample $\{(x_i, y_i)\}_{i=1}^m$ which consists of feature vectors $x_i \in \mathbb{R}^l$, $1 \leq i \leq m$, one obtains a hypothesis classifer $\hat{h}$ which incurs a large misclassification probability. A natural remedy in such situations is to try to improve the set of features by generating a new feature vector $y \in Y = \mathbb{R}^k$, $k \leq l$, which depends on x, with the aim of finding a better representation for a pattern which leads to larger separation between the different pattern-classes. This in turn leads to a simpler classifier $\tilde{g}$ which can now be better approximated by a hypothesis $\tilde{h}^*$ in the same class $\mathcal{H}$ of pseudo-dimension d, the latter having not been rich enough before for approximating the original target g. Consequently with the same sample complexity one obtains via ERM a hypothesis $\hat{h}$ which estimates $\tilde{g}$ better and therefore having a misclassification probability closer to the optimal Bayes misclassification probability.

Restricting to linear mappings A: $X \to Y$, classical discriminant analysis methods (cf. Fukunaga [14, Section 9.2]; Duda and Hart [13, Chap. 4]) calculate the optimal new feature vector y by determining the best linear map A* which, according to one of the widely used criteria, maximizes the pattern class separability. Such criteria are defined by the known class probabilities p_j, the class conditional means $\mu_j = E(X|j)$, and the class conditional covariance matrices $C_j = E((X - \mu_j)(X - \mu_j)^T|j)$, $1 \leq j \leq M$, where expectation $E(\cdot|j)$ is taken with respect to the jth class conditional probability distribution corresponding to $f(x|j)$. In reality the empirical average over the sample is used instead of taking expectation, since the underlying probability distributions corresponding to $f(x|j)$, $1 \leq j \leq M$, are unknown. Theoretically, the quantities μ_j, C_j, may be viewed as partial indirect information about the target Bayes classifier g. Such information can be represented by an n-dimensional vector of *linear* functionals acting on $f(x|j)$, $1 \leq j \leq M$, i.e., $N([f(x|1), \ldots, f(x|M)]) = [\{\mu_{j,s}\}_{j=1,}^{M}{}_{s=1}^{l}, \{\sigma_{s,r}^{j}\}_{j=1,}^{M}{}_{s\leq r=1}^{l}]$, where $\mu_{j,s} = \int_X x_s f(x|j)\, dx$, and $\sigma_{s,r}^{j} = \int_X x_s x_r f(x|j)\, dx$, where x_r, x_s, $1 \leq r, s \leq l$, are elements of x. The dimensionality of the information vector is $n = (Ml/2)(l + 3)$.

We have so far presented the theory for learning from examples and introduced the importance of partial information from a practical perspective. Before we proceed with a theoretical treatment of learning with partial information we digress momentarily to introduce a new quant ity which is defined in the context of the mathematical field of approximation theory which plays an important part in our learning framework.

4. A NEW NONLINEAR APPROXIMATION WIDTH

The large mathematical field of approximation theory is primarily involved in problems of existence, uniqueness, and characterization of the best approximation to elements of a normed linear space $\mathcal{F}$ by various types of finite-dimensional subspaces $\mathcal{F}_n$ of $\mathcal{F}$ (cf. Pinkus [25]). Approximation of an element $f \in \mathcal{F}$ is measured by the distance of the finite-dimensional subspace $\mathcal{F}_n$ to f where distance is usually defined as $\inf_{g \in \mathcal{F}_n} \|f - g\|$, where throughout this discussion $\|\cdot\|$ is any well-defined norm over $\mathcal{F}$. The degree of approximation of a subset (possibly a nonlinear manifold) $F \subset \mathcal{F}$ by $\mathcal{F}_n$ is defined by the distance between F and $\mathcal{F}_n$ which is usually taken as $\sup_{f \in F} \inf_{g \in \mathcal{F}_n} \|f - g\|$. The Kolmogorov n-width is the classical distance definition when one allows the approximating set $\mathcal{F}_n$ to vary over all possible linear subspaces of $\mathcal{F}$. It is defined as $K_n(F; \mathcal{F}) = \inf_{\mathcal{F}_n \subset \mathcal{F}} \sup_{f \in F} \inf_{g \in \mathcal{F}_n} \|f - g\|$. This definition leads to the notion of the best approximating subspace $\mathcal{F}_n$, i.e., the one whose distance from F equals $K_n(F; \mathcal{F})$.

While linear approximation, e.g., using finite dimensional subspaces of polynomials, is important and useful, there are many known spaces $\mathcal{F}$ which can be approximated better by *nonlinear* subspaces, for instance, by the span of a neural-network basis $\mathcal{H} = \{h(x) = \sum_{i=1}^{n} c_i \sigma(w_i^T x - b_i)\colon\ w_i \in \mathbb{R}^l,\ c_i,\ b_i \in \mathbb{R},\ 1 \le i \le n\}$, where $\sigma(y) = 1/(1+e^{-y})$. In this brief overview we will follow the notation and definitions of Devore [10]. Let M_n be a mapping from $\mathbb{R}^n$ into the Banach space $\mathcal{F}$ which associates each $a \in \mathbb{R}^n$ the element $M_n(a) \in \mathcal{F}$. Functions $f \in \mathcal{F}$ are approximated by functions in the manifold $\mathcal{M}_n = \{M_n(a)\colon a \in \mathbb{R}^n\}$. The measure of approximation of f by $\mathcal{M}_n$ is naturally defined as the distance $\inf_{a\in\mathbb{R}^n} \|f - M_n(a)\|$. As above, the degree of approximation of a subset F of $\mathcal{F}$ by $\mathcal{M}_n$ is defined as $\sup_{f\in F} \inf_{a\in\mathbb{R}^n} \|f - M_n(a)\|$.

In analogy to the Kolmogorov n-width, it would be tempting to define the optimal approximation error of F by manifolds of finite dimension n as $\inf_{\mathcal{M}_n} \sup_{f\in F} \inf_{a\in\mathbb{R}^n} \|f - M_n(a)\|$. However, as pointed out in [10], this width is zero for all subsets F in every separable class $\mathcal{F}$. To see this, consider the following example which describes a space filling manifold: let $\{f_k\}_{k=-\infty}^{\infty}$ be dense in $\mathcal{F}$ and define $M_1(a) = (a-k) f_{k+1} + (k+1-a) f_k$ for $k \le a \le k+1$. The mapping $M_1\colon \mathbb{R}^1 \to \mathcal{F}$, is continuous with a corresponding one-dimensional manifold $\mathcal{M}_1 \subset \mathcal{F}$ satisfying $\sup_{f\in F} \inf_{a\in\mathbb{R}^1} \|f - M_1(a)\| = 0$.

Thus this measure of width of F is not natural. One possible alternative used in approximation theory is to impose a smoothness constraint on the nonlinear manifolds $\mathcal{M}_n$ that are allowed in the outermost infimum. However, this excludes some interesting manifolds such as splines with free knots. A more useful constraint is to limit the selection operator r, which takes an element $f \in F$ to $\mathbb{R}^n$, to be continuous. Given such operator r then the approximation of f by a manifold $\mathcal{M}_n$ is $M_n(r(f))$. The distance between the set F and the manifold $\mathcal{M}_n$ is then defined as $\sup_{f\in F} \|f - M_n(r(f))\|$. The continuous *nonlinear* n-width of F is then defined as $D_n(F;\ \mathcal{F}) = \inf_{r\colon \mathrm{cont.},\ \mathcal{M}_\mathrm{n}} \sup_{f\in F} \|f - M_n(r(f))\|$, where the infimum is taken over all continuous selection operators r and all manifolds $\mathcal{M}_n$. This width is considered by Alexandrov [33] and Devore [10] and is determined for various F and $\mathcal{F}$ in [10].

The Alexandrov nonlinear width does not in general reflect the degree of approximation of the more natural selection operator r which chooses the best approximation for an $f \in F$ as its closest element in $\mathcal{M}_n$, i.e., that whose distance from f equals $\inf_{g\in\mathcal{M}_n} \|f - g\|$, the reason being that such r is not necessarily continuous. In this paper we consider an interesting alternate definition for a nonlinear width of a function class which does not have this deficiency.

Based on the pseudo-dimension (Definition 3 in Section 2) we define the nonlinear width

$$\rho_d(F) \equiv \inf_{\mathcal{H}^d} \sup_{f\in F} \inf_{h\in\mathcal{H}^d} \|f - h\|, \tag{7}$$

where $\mathcal{H}^d$ runs over all classes (not necessarily in $\mathcal{F}$) having pseudo-dimension d.

Now the natural selection operator is used, namely, the one which approximates f by an element $h(f)$, where $\|f - h(f)\| = \inf_{h\in\mathcal{H}^d} \|f - h\|$. The constraint of using finite pseudo-dimensional approximation manifolds allows dropping the smoothness constraint on the manifold $\mathcal{H}^d$ and the continuity constraint on the selection operator. The width ρ_d expresses the ability of manifolds to approximate according to their pseudo-dimension d as opposed to their dimensionality as in some of the classical widths.

The reason that ρ_d is interesting from a learning theoretic aspect is that the constraint on the approximation manifold $\mathcal{H}^d$ involves the pseudo-dimension $\dim_p(\mathcal{H}^d)$ which was shown in Section 2 to have a direct effect on uniform learnability, namely, a finite pseudo-dimension guarantees consistent estimation. Thus ρ_d involves two independent mathematical notions, namely, the approximation ability and the statistical estimation ability of $\mathcal{H}^d$. As will be shown in the next sections, joining both notions in one quantity enables us to quantify the trade-off between information and sample complexity as applied to the learning paradigm.

We halt the discussion about ρ_d and refer the interested reader to [23] where we estimate it for a standard Sobolev class $W_p^{r,l}$, $1 \leq p, q \leq \infty$.

5. THE MINIMAL PARTIAL INFORMATION ERROR

In this section we review some basic concepts in the field of information-based complexity and then extend these to define a new quantity called the minimal partial information error which is later used in the learning framework. Throughout this section, $\|\cdot\|$ denotes any function norm and the distance between two function classes $\mathcal{A}$ and $\mathcal{B}$ is denoted as $\operatorname{dist}(\mathcal{A}, \mathcal{B}, L_q) = \sup_{a\in\mathcal{A}} \inf_{b\in\mathcal{B}} \|a - b\|_{L_q}$, $q \geq 1$.

The following formulation of partial information is taken from Traub *et al.* [34]. While we limit here to the case of approximating functions $f \in \mathcal{F}$ we note that the theory is suitable for problems of approximating general functionals $S(f)$.

Let $N_n \colon \mathcal{F} \to N_n(\mathcal{F}) \subseteq \mathbb{R}^n$ denote a general information operator. The information $N_n(g)$ consists of n measurements taken on the target function g, or in general, any function $f \in \mathcal{F}$; i.e.,

$$N_n(f) = [L_1(f), \ldots, L_n(f)]$$

where L_i, $1 \leq i \leq n$, denote any functionals. We call n the *cardinality* of information and we sometimes omit n and write $N(f)$. The variable y denotes an element in $N_n(\mathcal{F})$. The subset $N_n^{-1}(y) \subset \mathcal{F}$ denotes all functions $f \in \mathcal{F}$ which share the same information vector y, i.e.,

$$N_n^{-1}(y) = \{f \in \mathcal{F} \colon N_n(f) = y\}.$$

We denote by $N_n^{-1}(N_n(g))$ the *solution set* which may also be written as $\{f \in \mathcal{F} \colon N_n(f) = N_n(g)\}$, which consists of all indistinguishable functions

$f \in \mathcal{F}$ having the same information vector as the target g. Given $y \in \mathbb{R}^n$, it is assumed that a single element denoted as $g_y \in N_n^{-1}(y)$ can be constructed.

In this model information effectively partitions the target class $\mathcal{F}$ into infinitely many subsets $N_n^{-1}(y)$, $y \in \mathbb{R}^n$, each having a *single* representative g_y which forms the approximation for any $f \in N^{-1}(y)$. Denote the radius of $N^{-1}(y)$ by

$$r(N, y) = \inf_{f' \in \mathcal{F}} \sup_{f \in N^{-1}(y)} \| f - f' \| \tag{8}$$

and call it the *local radius of information N* at y. The *global radius of information N* at y is defined as the local radius for a worst y, i.e.,

$$r(N) = \sup_{y \in N(\mathcal{F})} r(N, y).$$

This quantity measures the intrinsic uncertainty or error which is associated with a fixed information operator N. Note that in both of these definitions the dependence on $\mathcal{F}$ is implicit.

Let Λ be a family of functionals and consider the family Λ_n which consists of all information $N = [L_1, \ldots, L_k]$ of cardinality $k \leq n$ with $L_i \in \Lambda$, $1 \leq i \leq n$. Then

$$r(n, \Lambda) = \inf_{N \in \Lambda_n} r(N)$$

is called the nth *minimal radius of information* in the family Λ and $N_n^* = [L_1^*, \ldots, L_n^*]$ is called the nth *optimal information* in the class Λ iff $L_i^* \in \Lambda$ and $r(N_n^*) = r(n, \Lambda)$.

When Λ is the family of all *linear* functionals then $r(n, \Lambda)$ becomes a slight generality of the well-known Gelfand-width of the class $\mathcal{F}$ whose classical definition is $d^n(\mathcal{F}) = \inf_{A^n} \sup_{f \in \mathcal{F} \cap A^n} \| f \|$, where A^n is any linear subspace of codimension n. In this paper we restrict to the family Λ of linear functionals and for notational simplicity we will henceforth take the information space $N_n(\mathcal{F}) = \mathbb{R}^n$.

As already mentioned in the definition of $r(N, y)$ there is a single element $g_y \in \mathcal{F}$ not necessarily in $N^{-1}(y)$ which is selected as an approximator for all functions $f \in N^{-1}(y)$. Such a definition is useful for the problem of information-based complexity since all that one is concerned with is to produce an ϵ-approximation based on information alone. In the PAC framework, however, a major significance is placed on providing an approximator to a target g which is an element not necessarily of the target class $\mathcal{F}$ but of some hypothesis class $\mathcal{H}$ of finite pseudo dimension by which $\mathcal{F}$ is uniformly learnable.

We therefore replace the single-representative of the subset $N^{-1}(y)$ by a whole approximation class of functions $\mathcal{H}_y^d$ of pseudo-dimension d. Note that now information alone does not "point" to a single ϵ-approximation element, but rather to a manifold $\mathcal{H}_y^d$, possibly nonlinear, which for *any* $f \in N^{-1}(y)$, in particular the target g, contains an element h^*, dependent on g, such that the distance $\|g - h^*\| \leq \epsilon$. Having a pseudo-dimension d implies that with a finite random sample $\{(x_i,\ g(x_i))\}_{i=1}^m$, an ERM learning algorithm (after being shown partial information and hence pointed to the class $\mathcal{H}_y^d$) can determine a function $\hat{h} \in \mathcal{H}_y^d$ whose distance from g is no farther than ϵ from the distance between h^* and g with confidence $1 - \delta$. Thus based on n units of information about g and m labeled examples $\{(x_i,\ g(x_i))\}_{i=1}^m$, an element $\hat{h}$ can be found such that $\|g - \hat{h}\| \leq 2\epsilon$ with probability $1 - \delta$.

The sample complexity m does not depend on the type of hypothesis class but only on its pseudo-dimension d. Thus the above construction is true for any hypothesis class (or manifold) of pseudo-dimension d. Hence we may permit *any* hypothesis class of pseudo-dimension d to play the role of the approximation manifold $\mathcal{H}_y^d$ of the subset $N^{-1}(y)$. This amounts to replacing the infimum in the definition (8) of $r(N, y)$ by $\inf_{\mathcal{H}^d}$ and replacing $\| f - f'\|$ by $\text{dist}(f, \mathcal{H}^d) = \inf_{h \in \mathcal{H}^d} \| f - h\|$, yielding the quantity $\rho_d(N^{-1}(y))$ as a new definition for a local "radius" and a new quantity $I_{n,\,d}(\mathcal{F})$ (to be defined later) which replaces $r(n, \Lambda)$.

We next formalize these ideas through a sequence of definitions. We use $\rho_d(K, L_q)$ to explicitly denote the norm L_q used in the definition of (7). We now define three optimal quantities, N_n^*, $\mathcal{H}_{N_n^*}^d$, and h^*, all of which implicitly depend on the unknown distribution P while h^* depends also on the unknown target g.

Definition 4. Let the optimal linear information operator N_n^* of cardinality n be one which minimizes the approximation error of the solution set $N_n^{-1}(y)$ (in the worst case over $y \in \mathbb{R}^n$) over all linear operators N_n of cardinality n and manifolds of pseudo-dimension d. Formally, it is defined as one which satisfies

$$\sup_{y \in \mathbb{R}^n} \rho_d(N_n^{*-1}(y),\ L_1(P)) = \inf_{N_n} \sup_{y \in \mathbb{R}^n} \rho_d(N_n^{-1}(y),\ L_1(P)).$$

Definition 5. For a fixed optimal linear information operator N_n^* of cardinality n define the optimal hypothesis class $\mathcal{H}_y^d$ of pseudo-dimension d (which depends implicitly on N_n^* through y) as one which minimizes the approximation error of the solution set $N_n^{*-1}(y)$ over all manifolds of pseudo-dimension d. Formally, it is defined as one which satisfies

$$\text{dist}(N_n^{*-1}(y),\ \mathcal{H}_y^d,\ L_1(P)) = \rho_d(N_n^{*-1}(y),\ L_1(P)).$$

Definition 6. For a fixed target $g \in \mathcal{F}$, optimal linear information operator N_n^* and optimal hypothesis class $\mathcal{H}_{N_n^*(g)}^d$ define the optimal hypothesis $h^* \in \mathcal{H}_{N_n^*(g)}^d$ to be any function which minimizes the error over $\mathcal{H}_{N_n^*(g)}^d$, namely,

$$L(h^*) = \inf_{h \in \mathcal{H}_{N_n^*(g)}^d} L(h). \tag{9}$$

As mentioned earlier, the main motive of the paper is to compute the value of partial information for learning in the PAC sense. We will assume that the teacher has access to unlimited (linear) information which is represented by him *knowing* the optimal linear information operator N_n^* and optimal hypothesis class $\mathcal{H}_y^d$ for every $y \in \mathbb{R}^n$. Thus in this ideal setting providing partial information amounts to pointing to the optimal hypothesis class $\mathcal{H}_{N_n^*(g)}^d$ which contains an optimal hypothesis h^*. We again note that information alone does not point to h^* but it is the role of learning from examples to complete the process through estimating h^* using a hypothesis $\hat{h}$.

The error of h^* is important in its own right. It represents the minimal error for learning a particular target g given optimal information of cardinality n. In line with the notion of uniform learnability (see Section 2) we define a variant of this optimal quantity which is *independent* of the target g and probability distribution P; i.e., instead of a specific target $g \in \mathcal{F}$, we consider the worst target in $\mathcal{F}$ and we use the L_∞ norm for approximation. This yields the following definition.

DEFINITION 7 (Minimal partial information error). For any target class $\mathcal{F}$ and any integers $n, d \geq 1$, let

$$I_{n,\,d}(\mathcal{F}) \equiv \inf_{N_n} \sup_{y \in \mathbb{R}^n} \rho_d(N_n^{-1}(y),\ L_\infty),$$

where N_n runs over all linear information operators.

$I_{n,\,d}(\mathcal{F})$ represents the minimal error for learning the worst-case target in the PAC sense (i.e., assuming an unknown underlying probability distribution) while given optimal information of cardinality n and using an optimal hypothesis class of pseudo-dimension d.

We proceed next to unify the theory of Section 2 with the concepts introduced in the current section.

6. LEARNING FROM EXAMPLES WITH OPTIMAL PARTIAL INFORMATION

In Section 2 we reviewed the notion of uniform learnability of a target class $\mathcal{F}$ by a hypothesis class $\mathcal{H}^d$ of pseudo-dimension $d < \infty$. By minimizing an empirical error based on the random sample, a learner obtains a hypothesis $\hat{h}$ which provides a close approximation of the optimal hypothesis h^* to within ϵ accuracy with confidence $1 - \delta$.

Suppose that prior to learning the learner obtains optimal information $N_n^*(g)$ about g. This effectively points the learner to a class $\mathcal{H}_{N_n^*(g)}^d$ which contains a hypothesis h^* as defined in (9). The error of h^* is bounded from above as

$$
\begin{aligned}
L(h^*) &= \inf_{h \in \mathcal{H}^d_{N_n^*(g)}} L(h) && (10)\\
&= \inf_{h \in \mathcal{H}^d_{N_n^*(g)}} \|g - h\|_{L_1(P)} && (11)\\
&\leq \sup_{\{f \in \mathcal{F}:\, N_n^*(f) = N_n^*(g)\}} \inf_{h \in \mathcal{H}^d_{N_n^*(g)}} \|f - h\|_{L_1(P)} && (12)\\
&= \operatorname{dist}\left(N_n^{*-1}(N_n^*(g)),\ \mathcal{H}^d_{N_n^*(g)},\ L_1(P)\right). && (13)
\end{aligned}
$$

By Definition 5 this equals $\rho_d(N_n^{*-1}(N_n^*(g)),\ L_1(P))$ and is bounded from above by

$$\sup_{y \in \mathbb{R}^n} \rho_d\left(N_n^{*-1}(y),\ L_1(P)\right).$$

The latter equals

$$\inf_{N_n} \sup_{y \in \mathbb{R}^n} \rho_d(N_n^{-1}(y),\ L_1(P))$$

by Definition 4. This is bounded from above by $\inf_{N_n} \sup_{y \in \mathbb{R}^n} \rho_d(N_n^{-1}(y),\ L_\infty)$ which from Definition 7 is $I_{n,d}(\mathcal{F})$. Subsequently, the teacher provides m i.i.d. examples $\{(x_i,\ g(x_i))\}_{i=1}^m$ randomly drawn according to any probability distribution P on X. Armed with prior knowledge and a random sample the learner then minimizes the empirical error $L_m(h)$ over all $h \in \mathcal{H}^d_{N_n^*(g)}$, yielding an estimate $\hat{h}$ of h^*. We may break up the error $L(\hat{h})$ into a *learning error* and a *minimal partial information error* components

$$
\begin{aligned}
L(\hat{h}) &= \left(L(\hat{h}) - L(h^*)\right) + L(h^*)\\
&\leq \overbrace{\epsilon(m,\ d,\ \delta)}^{\text{“learning error”}} + \overbrace{I_{n,d}(\mathcal{F})}^{\text{“minimal partial information error”}}, && (14)
\end{aligned}
$$

where the learning error, defined in (4), measures the extra error incurred by using $\hat{h}$ as opposed to the optimal hypothesis h^*.

The important difference from the PAC model can be seen in comparing the upper bound of (14) with that of (5). The former depends not only on the sample size m and pseudo-dimension d but also on the amount n of partial information. To see how m, n, and d influence the performance, i.e., the error of $\hat{h}$, we will next particularize to a specific target class.

7. SOBOLEV TARGET CLASS

The preceding theory is now applied to the problem of learning a target in a Sobolev class $\mathcal{F} = W_\infty^{r,l}(M)$, for $r,\ l \in \mathbb{Z}_+,\ M > 0$, which is defined as all functions over $X = [0, 1]^l$ having all partial derivatives up to order r bounded in the L_∞ norm by M. Formally, let $k = [k_1, \ldots, k_l] \in \mathbb{Z}_+^l$, $\|k\| = \sum_{i=1}^l k_i$, and denote by $D^k f = (\partial^{k_1} + \cdots + k_l)/(\partial x_1^{k_1} \ldots \partial x_l^{k_l}) f$, then

$$W_\infty^{r,l}(M) = \{f\colon \sup_{x \in [0,1]^t} |D^k f(x)| \le M,\ \|k\| \le r\}$$

which henceforth is referred to as $W_\infty^{r,l}$ or $\mathcal{F}$. We now state the main results and their implications.

THEOREM 2. *Let $\mathcal{F} = W_\infty^{r,l}$, $n \ge 1$, $d \ge 1$, be given integers and $c_2 > 0$ a constant independent of n and d. Then*

$$I_{n,d}(\mathcal{F}) \le \frac{c_2}{(n+d)^{r/l}}.$$

The proof of the theorem is in Section A.2.

THEOREM 3. *Let the target class $\mathcal{F} = W_\infty^{r,l}$ and $g \in \mathcal{F}$ be the unknown target function. Given an i.i.d. random sample $\{(x_i, g(x_i))\}_{i=1}^m$ of size m drawn according to any unknown distribution P on X. Given an optimal partial information vector $N_n^*(g)$ consisting of n linear operations on g. For any $d \ge 1$, let $\mathcal{H}_{N_n^*(g)}^d$ be the optimal hypothesis class of pseudo-dimension d. Let $\hat{h}$ be the output hypothesis obtained from running empirical error minimization over $\mathcal{H}_{N_n^*(g)}^d$. Then for an arbitrary $0 < \delta < 1$, the error of $\hat{h}$ is bounded as*

$$L(\hat{h}) \le c_1 \sqrt{\frac{d \log^2 d \ln m + \ln(1/\delta)}{m}} + \frac{c_2}{(n+d)^{r/l}}, \tag{15}$$

where $c_1, c_2 > 0$ are constants independent of m, n, and d.

The proof of Theorem 3 is based on Theorem 1 and Theorem 2, both of which are proved in the Appendix.

We now discuss several dependences and trade-offs between the three complexity variables m, n, and d. First, for a fixed sample size m and fixed information cardinality n there is an optimal class complexity

$$d^* \le c_3 \left(\left\{ \frac{rm}{l\sqrt{\ln m}} \right\}^{2l/(l+2r)} - n \right), \tag{16}$$

which minimizes the upper bound on the error where $c_3 > 0$ is an absolute constant. The complexity d is a free parameter in our learning setting and is proportional to the degree in which the estimator $\hat{h}$ fits the data while estimating the optimal hypothesis h^*. The result suggests that for a given sample size

m and partial information cardinality n, there is an optimal estimator (or model) complexity d^* which minimizes the error rate. Thus if a structure of hypothesis classes $\{\mathcal{H}^d\}_{d=1}^{\infty}$ is available in the learning problem, then based on fixed m and n the best choice of a hypothesis class over which the learner should run empirical error minimization is $\mathcal{H}^{d^*}$ with d^* as in (16).

The notion of having an optimal complexity d^* is closely related to statistical model selection (cf. Linhart and Zucchini [19], Devroye *et al.* [12], Ratsaby *et al.* [29]). For instance, in Vapnik's structural risk minimization criterion (SRM) [38] the trade-off is between m and d. For a fixed m, it is possible to calculate the optimal complexity d^* of a hypothesis class in a nested class structure, $\mathcal{H}^1 \subset \mathcal{H}^2 \ldots$, by minimizing an upper bound on the error $L(\hat{h}) \leq L_m(\hat{h}) + \epsilon(m, d, \delta)$, over all $d \geq 1$. The second term $\epsilon(m, d, \delta)$ is commonly referred to as the penalty for data-overfitting which one wants to balance against the empirical error. Similarly, in our result, the upper bound on the learning error reflects the cost or penalty of overfitting the data—the larger d, the higher the degree of data fit and the larger the penalty.

However, here, as opposed to SRM, the bound is independent of the random sample and there is an extra parameter n that affects how m and d trade off. As seen from (16), for a fixed sample size m it follows that the larger n the smaller d^*. This is intuitive since the more partial information, the smaller the solution set $N_n^{-1}(N_n(g))$ and the lower the complexity of a hypothesis class needed to approximate it. Consequently, the optimal estimator $\hat{h}$ belongs to a simpler hypothesis class and does not overfit the data as much.

We next compute the trade-off between n and m. Assuming d is fixed (not necessarily at the optimal value d^*) and fixing the total available information and sample size, $m + n$, at some constant value while minimizing the upper bound on $L(\hat{h})$ over m and n, we obtain $m \leq c_5 n^{(l+2r)/2l}\sqrt{\ln n}$ for a constant $c_5 > 0$ which depends polynomially only on l and r. We conclude that when the dimensionality l of X is smaller than twice the smoothness parameter r, the sample size m grows polynomially in n at a rate no larger than $n^{(1+r)/l}$; i.e., partial information about the target g is worth approximately a polynomial number of examples. For $l > 2r$, n grows polynomially in m at a rate no larger than $m^2/\ln m$; i.e., information obtained from examples is worth a polynomial amount of partial information.

We have focused so far on dealing with the ideal learning scenario in which the teacher has access to the optimal information operator N_n^* and optimal hypothesis class $\mathcal{H}_{N_n^*}^d$. The use of such optimally efficient information was required from an information theoretic point of view in order to calculate the trade-off between the sample complexity m and information cardinality n. But we have not specified the form of such optimal information and hypothesis class.

In the next result we state a lower bound on the minimal partial information error $I_{n,d}(\mathcal{F})$ and subsequently show that there exists an operator and a hypothesis class which almost achieve this lower bound.

THEOREM 4. *Let $\mathcal{F} = W_\infty^{r,l}$ and $n \geq 20$, $d \geq 1$ be given integers. Then*

$$I_{n,d}(\mathcal{F}) \geq \frac{1}{(1280n \ln n + 128d \ln d)^{r/l}}.$$

The proof is in Section A.3.

Our next result shows that there exists an operator $\hat{N}_n$ and a *linear* manifold $\mathcal{H}^d_{N_n}$ which together achieve the upper bound on $I_{n,d}$ stated in Theorem 2.

First we note several definitions and facts. For a multi-integer $k \in \mathbb{Z}^l_+$ denote by $\|k\| = \sum_{i=1}^l k_i$. Let $\Delta_j \subset [0, 1]^l$ be an l-dimensional cube. Denote by $1_{\Delta_j}(x)$ the indicator function of Δ_j. Denote by $\alpha_{r,l}$ the number of vectors k for which $\|k\| \leq r - 1$, where r is the smoothness parameter of $W^{r,l}_\infty$. Let R_q be a partition of the domain $X = [0, 1]^l$ which is uniform in every variable x_i, $1 \leq i \leq l$, and consists of a total of q identical cubes Δ_j, $1 \leq j \leq q$. Let $S_{q,r} = \{p_j(x) = \sum_{k:\,\|k\| \leq r-1} a_k x_1^{k_1} \ldots x_l^{k_l} 1_{\Delta_j}(x)\colon 1 \leq j \leq q\}$ be a linear subspace of all piecewise polynomials of degree at most $r - 1$ in the variables x_i, $1 \leq i \leq l$, with a support being a cube Δ_j, $1 \leq j \leq q$. The dimension of $S_{q,r}$ equals $q\alpha_{r,l}$. There exists a linear operator $T_{q,r}\colon W^{r,l}_\infty \to S_{q,r}$ which maps an $f \in W^{r,l}_\infty$ to an element of $S_{q,r}$.

THEOREM 5. *Given integers n and $d \geq 1$, choose q such that the dimension of $S_{q,r}$ is $q\alpha_{r,l} = n + d$. Consider the target class $\mathcal{F} = W^{r,l}_\infty$. Denote by $\phi_1, \ldots, \phi_{n+d}$ a basis in $S_{q,r}$. Then for any $f \in W^{r,l}_\infty$, we have $T_{q,r}(f) = \sum_{i=1}^{n+d} L_i(f)\phi_i(x)$ for some linear functionals L_i, $1 \leq i \leq n + d$. Define the information operator $\hat{N}_n(f) = [L_1(f), \ldots, L_n(f)]$ and the approximating class to be a linear subspace*

$$H^d_y \equiv \mathcal{H}^d_{\hat{N}_n(f)} = \left\{ \sum_{i=1}^{n} y_i\phi_i(x) + \sum_{i=n+1}^{n+d} c_i\phi_i(x)\colon c_i \in \mathbb{R} \right\}.$$

Then the specific combination of information operator $\hat{N}_n$ and hypothesis classes $\{H^d_y\}_{y\in\mathbb{R}^n}$ achieve a partial information error which is bounded from above as

$$\sup_{y\in\mathbb{R}^n} \sup_{f\in\mathcal{F}\cap\hat{N}_n^{-1}(y)} \inf_{h\in H^d_y} \|f - h\|_{L_\infty} \leq \frac{c_6}{(n + d)^{r/l}}$$

for some constant $c_6 > 0$ independent of n and d.

From Theorems 4 and 5 it follows that $\hat{N}_n$ and $\mathcal{H}^d_{\hat{N}_n(g)}$ incur an error which to within a logarithmic factor in n and d is close to the minimal partial information error $I_{n,d}(\mathcal{F})$. Thus they come close to being the optimal combination N^*_n and

$\mathcal{H}^d_{N^*_n}$. Hence for learning a target g in a Sobolev class using examples with partial information, the operator $\hat{N}_n$ and the linear hypothesis class $\mathcal{H}^d_{\hat{N}_n(g)}$ guarantee an almost optimal performance; i.e., the upper bound on the error $L(\hat{h})$ is almost minimal, where $\hat{h}$ is taken as the empirical error minimizer over $\mathcal{H}^d_{\hat{N}_n(g)}$.

An additional comment is due. The fact that a *linear* manifold $\mathcal{H}^d_{\hat{N}_n(g)}$ achieves an almost optimal upper bound among *all* possible manifolds of pseudo-dimension d is a consequence of the choice of the target class $W^{r,l}_\infty$ and the norm L_∞ used for approximation. Suppose we consider, instead, another classical Sobolev class defined for fixed $1 \leq p \leq 2$ by $W^{r,l}_p = \{f : \|D^k f\|_{L_p} \leq M, \|k\| \leq r\}$. From classical results on the estimation of the Kolmogorov width of $W^{r,l}_p$, denoted here as $K_d(W^{r,l}_p, L_\infty)$, it can be shown that when using the L_∞-norm for approximation the optimal d-dimensional linear manifold has a worst-case approximation error which is lower bounded by $c_7/d^{r/l-1/p}$ for some constant $c_7 > 0$ independent of d. Whereas doing approximation by linear combinations of d piecewise polynomial-splines of degree r (but allowing the spline basis to depend on the target function which implies nonlinear approximation) leads to the ρ_d-width satisfying $\rho_d(W^{r,l}_p, L_\infty) \leq c_8(1/d)^{r/l}$. Thus $\rho_d(W^{r,l}_p, L_\infty) \ll K_d(W^{r,l}_p, L_\infty)$, where $a_d \ll b_d$ means $a_d/b_d \to 0$ as $d \to \infty$. Thus ρ_d is a genuine *nonlinear* width since there are target classes for which it is less than the Kolmogorov linear width in a strong sense.

8. CONCLUSIONS

We introduced a theoretical framework for representing the problem of learning a target function $g \in \mathcal{F}$ from examples by an empirical error minimization algorithm with partial information. Having defined a new information quantity $I_{n,d}(\mathcal{F})$ leads to an upper bound on the error of the estimator $\hat{h}$ which depends on the sample size m, the information cardinality n, and the pseudo-dimension of the approximating class $\mathcal{H}^d$. For a specific Sobolev target class $W^{r,l}_\infty$ one immediate consequence is a clear trade-off between m and n which suggests that the ratio between the smoothness parameter r and the dimensionality l of the domain X is crucial in determining which of the two types of information, namely information obtained from examples versus partial information obtained by a linear operator, is worth more. Roughly speaking, partial information is worth polynomially more than information from examples when $l < 2r$ while the opposite holds when $l > 2r$. Moreover, for the Sobolev class we obtained an information operator $\hat{N}_n$ which yields an almost optimal partial information error rate.

APPENDIX: PROOFS OF RESULTS

In this section we prove lower and upper bounds on $I_{n,d}(\mathcal{F})$. The method of proof for the lower bound (Theorem 1) is interesting in its own right as it combines the well-known property of a finite pseudo-dimensional manifold for the purpose of showing the existence of at least one bad target function which the manifold does not approximate well enough. For proving the upper bound (Theorem 2) we use the fact that a linear manifold of dimension d has a pseudo-dimension d (Lemma 3 in Section 2) which allows us to linearly approximate $\mathcal{F}$.

We first introduce the notation. Let $\mathbb{Z}_+$ denote the set of nonnegative integers, and $\mathbb{Z}_+^l$ denotes all l-dimensional vectors whose components are in $\mathbb{Z}_+$. Unless otherwise mentioned it will be implicit that the domain $X = [0, 1]^l$ and we write x for the vector $[x_1, \ldots, x_l]$, $\int f(x)\,dx$ represents $\int_X f(x_1, \ldots, x_l)\,dx_1 \ldots dx_l$. The notation $\mathcal{H}^d$ represents any manifold with $\dim_p(\mathcal{H}^d) = d$ and for all $h \in \mathcal{H}^d$, $g \in \mathcal{F}$, $L(h) = E|h - g| \le M$. For any vector $v \in \mathbb{R}^m$ and function f, we use the standard norms, $\|v\|_{l_p^m} \equiv (\sum_{j=1}^m v_j^p)^{1/p}$ and $\|f\|_{L_p} \equiv (\int_X |f(x)|^p\,dx)^{1/p}$, respectively. When the dimensionality of the vector is clear, we write $\|v\|_{l_p}$. We use the standard notation for a ball $B_p^m(r)$ of radius r in $\mathbb{R}^m$, where distance is measured in the l_p-norm, $1 \le p \le \infty$ (if $p = \infty$ then $B_\infty^m(1)$ is a cube of side 2). We define for any $a \in A$, $\mathrm{dist}(a, B, l_2) \equiv \inf_{b\in B} \|a - b\|_{l_2}$. The distance between two Euclidean sets $A, B \subset \mathbb{R}^m$ is defined as $\mathrm{dist}(A, B, l_2) \equiv \sup_{a\in A} \mathrm{dist}(a, B, l_2)$.

Sometimes we underline a symbol to explicitly indicate that it is a Euclidean vector or a set of vectors. For $x \in \mathbb{R}$ the function $\mathrm{sgn}(x)$ is defined as $\mathrm{sgn}(x) = 1$ if $x \ge 0$ and $\mathrm{sgn}(x) = -1$ if $x < 0$. For a vector $x \in \mathbb{R}^m$, define $\mathrm{sgn}(x) = [\mathrm{sgn}(x_1), \ldots, \mathrm{sgn}(x_m)]$. Let $\underline{i} \in \{-1, 1\}^m$. An *orthant* $Q_{\underline{i}}$ in $\mathbb{R}^m$ is an extension of the definition of a quadrant in $\mathbb{R}^2$; namely, there are 2^m orthants in $\mathbb{R}^m$ and $Q_{\underline{i}} = \{x \in \mathbb{R}^m : \mathrm{sgn}(x) = \underline{i}\}$.

We start with the proof of Theorem 1.

A.1. *Proof of Theorem* 1

We follow Vapnik's proof of Theorem 7.3 in [38], but where the complexity measure of $\mathcal{H}$ is the pseudo-dimension instead of his capacity measure defined on page 189. It is given that $\dim_p(\mathcal{H}^d) = d$. Let $y \in \mathbb{R}$ and $x \in X = [0, 1]^l$. First, we have

CLAIM 1. *The set of indicator functions*

$$A \equiv \{1_{\{(x,y):|h(x)-y)|>\beta\}} : h \in \mathcal{H},\ \beta \in \mathbb{R}_+\}$$

has $\mathrm{VC}(A) \le 16(d + 1)\log^2 4(d + 1)$.

Proof of Claim 1. Define the class of functions

$$\mathcal{G} = \{g_h(x, y) = h(x) - y : h \in \mathcal{H}\},$$

where we will also refer to $z \equiv (x, y) \in \mathbb{R}^{l+1}$. Consider the corresponding set

$$B = \{1_{\{g_h(z)>0\}}: h \in \mathcal{H}\}$$

of indicator functions. It is easily seen that $\mathrm{VC}(B) = \dim_p(\mathcal{H}) = d$ as we now show: If $\mathrm{VC}(B) > d$ then there exists a set of points $\{z_i\}_{i=1}^m$, where $z_i \equiv (x_i, y_i)$, $1 \le i \le m$, $m > d$, which is shattered by B. This implies that there exist pairs $\{(x_i, y_i)\}_{i=1}^m$, such that the set of binary vectors

$$\tilde{B} = \{[\mathrm{sgn}(h(x_1) - y_1), \mathrm{sgn}(h(x_2) - y_2), \ldots, \mathrm{sgn}(h(x_m) - y_m)]: h \in \mathcal{H}\}$$

equals $\{-1, 1\}^m$. The latter implies that $\dim_p(\mathcal{H}) > d$ and leads to a contradiction. For the other direction, suppose $\mathrm{VC}(B) < d$ then there does not exist a set of points $z_i \equiv (x_i, y_i)$, $1 \le i \le m$, $m = d$, which can be shattered by B. This implies there do not exist pairs $\{(x_i, y_i)\}_{i=1}^m$, such that the set of binary vectors $\{[\mathrm{sgn}(h(x_1) - y_1), \mathrm{sgn}(h(x_2) - y_2), \ldots, \mathrm{sgn}(h(x_m) - y_m)]: h \in \mathcal{H}\}$ equals $\{-1, 1\}^m$. This contradicts the fact that $\mathcal{H}$ has a pseudo-dimension d which proves that $\mathrm{VC}(B) = d$.

Next, denote by $g_{h,\beta}(x, y) \equiv g_h(z) - \beta = h(x) - y - \beta$ and let

$$C = \{1_{\{g_{h,\beta}(x,y)>0\}}: h \in \mathcal{H}, \beta \in \mathbb{R}_+\}.$$

Claim 2. *The* VC *dimension of the set* C *is upper bounded by* $2(d + 1) \log e(d + 1)$.

Proof of Claim 2. To see this, fix any sample $\{z_i\}_{i=1}^m$ and any function $g_h(z) \in \mathcal{G}$. Consider the class of functions that are translates of this fixed function g_h, i.e., $\{g_h - \beta: \beta \in \mathbb{R}_+\}$, together with its corresponding class of indicator functions,

$$C_{g_h} \equiv \{1_{\{g_h(z)-\beta>0\}}: \beta \in \mathbb{R}_+\},$$

where h is fixed. We claim that the $\mathrm{VC}(C_{g_h}) = 1$ as is now shown. The dichotomies that C_{g_h} picks on a sample $\{z_i\}_{i=1}^m$ are precisely the set of dichotomies on the points $(z_i, g_h(z_i)) \in \mathbb{R}^{l+2}$, $1 \le i \le m$, that are picked by the class of half-spaces of the form $H_\beta = \{(z, r) \in \mathbb{R}^{l+2}: r > \beta\}$, $\beta \in \mathbb{R}$. It is easy to see that the VC-dimension of the class of such half spaces is 1. Thus, for any fixed $g_h \in \mathcal{G}$, $\mathrm{VC}(C_{g_h}) = 1$. Hence from Proposition A2.1(iii) of [6], the number of dichotomies picked on $\{z_i\}_{i=1}^m$ by C_{g_h} is no more than em, for $m \ge 1$. Therefore for any $g_h \in \mathcal{G}$, no more than em new dichotomies which differ from its current dichotomy $[\mathrm{sgn}(g_h(z_1)), \mathrm{sgn}(g_h(z_2)), \ldots, \mathrm{sgn}(g_h(z_m))]$, are produced by adding arbitrary $\beta \in \mathbb{R}$ to g_h. Now from the proof of Claim 1, $\mathrm{VC}(B) = d \ge 1$. Hence, from [6] it follows that for any $m \ge d$ the number of dichotomies picked by the class B on any sample $\{z_i\}_{i=1}^m$ is at most $(em/d)^d$. It follows that for all $m \geq d+1$ the set of dichotomies

$$\tilde{C} = \{[\mathrm{sgn}(g_h(z_1)-\beta),\ \mathrm{sgn}(g_h(z_2)-\beta),\ \ldots,\ \mathrm{sgn}(g_h(z_m)-\beta)]\colon h \in \mathcal{H},\ \beta \in \mathbb{R}\},$$

picked by C on a sample $\{z_i\}_{i=1}^m$ has cardinality no more than $(em)(em/d)^d \leq (em)^{d+1}$. To find an upper bound on VC(C) we solve for the largest m for which $(em)^{d+1} \leq 2^m$, and obtain VC(C) $\leq 2(d+1)\log e(d+1)$ which proves Claim 2.

Continuing, consider the class

$$D = \{1_{\{g_{h,-\beta}(x,y)<0\}}\colon h \in \mathcal{H},\ \beta \in \mathbb{R}_+\}.$$

CLAIM 3. *The* VC-*dimension of D is upper bounded by* $2(d+1)\log e(d+1)$.

Proof. The proof follows from the proof of Claim 2, except that now we consider dichotomies on the set of points $\{(z_i, -g_h(z_i))\}_{i=1}^m$ picked by half spaces H_β as above.

We now continue with the proof of Claim 1. For any $h \in \mathcal{H}$, $\beta \in \mathbb{R}_+$,

$$\{(x,y)\colon |h(x)-y| > \beta\} = \{(x,y)\colon h(x)-y-\beta > 0\} \cup \{(x,y)\colon h(x)-y+\beta < 0\},$$

and since both VC(B) and VC(D) are at most $2(d+1)\log e(d+1)$ then by Proposition A2.1(ii) of [6], for any sample $\{z_i\}_{i=1}^m$ of size m, where $m > 2(d+1)\log e(d+1)$ the number of dichotomies picked on $\{z_i\}_{i=1}^m$ by the class of indicator functions

$$\{1_{\{\{(x,y)\colon h(x)-y-\beta>0\}\cup\{(x,y)\colon h(x)-y+\beta<0\}\}}\colon h \in \mathcal{H},\ \beta \in \mathbb{R}_+\} \tag{17}$$

is no more than $m^{2(d+1)\log e(d+1)} m^{2(d+1)\log e(d+1)} = m^{4(d+1)\log e(d+1)}$ dichotomies. The class in (17) is precisely the class A. To find an upper bound on VC(A) it suffices to solve for the largest m for which $m^{4(d+1)\log e(d+1)} \leq 2^m$. Solving for m yields that VC(A) $\leq 16(d+1)\log^2 4(d+1)$ for all $d \geq 1$, which proves Claim 1.

It only remains to follow Vapnik's proof of Theorem 7.3 in [38], where, instead of using a class of indicator functions of the form $\{1_{\{(x,y)\colon (h(x)-y)^2>\beta\}}\colon h \in \mathcal{H}^d,\ \beta \in \mathbb{R}_+\}$, we use our set A and the fact that VC(A) $\leq 16(d+1)\log^2 4(d+1) \equiv d'$ to imply that the growth function of A is bounded by $1.5m^{d'}/d'!$. The statement of the theorem then follows from Vapnik's proof. ■

A.2. *Proof of Theorem 2*

To establish an upper bound on $I_{n,d}(\mathcal{F})$ it suffices to choose a particular information operator $\hat{N}_n$ and a particular manifold $\hat{\mathcal{H}}^d$ since

$$I_{n,d}(\mathcal{F}) = \inf_{N_n} \sup_{y\in N_n(\mathcal{F})} \inf_{\mathcal{H}^d} \sup_{f\in\mathcal{F}\cap N_n^{-1}(y)} \inf_{h\in\mathcal{H}^d} \|f-h\|_{L_\infty} \tag{18}$$

$$\leq \sup_{y\in\hat{N}_n(\mathcal{F})} \sup_{f\in\mathcal{F}\cap\hat{N}_n^{-1}(y)} \inf_{h\in\hat{\mathcal{H}}^d} \|f-h\|_{L_\infty}. \tag{19}$$

We next describe the particular choice of $\hat{\mathcal{H}}^d$ followed by the choice of $\hat{N}_n$.

We will take $\hat{\mathcal{H}}^d$ to be a linear manifold H_y^d of dimension d which from Haussler [16] has a pseudo-dimension d, where y shows the dependence of the manifold on the information vector y. Specifically, let H_y^d be the space of piecewise polynomial functions,

$$H_y^d \equiv \left\{ \sum_{i=1}^{n} y_i\phi_i(x) + \sum_{i=n+1}^{n+d} c_i\phi_i(x) \colon c_i \in \mathbb{R} \right\},$$

where n, d are any given positive integers and the functions $\phi_i(x)$ may also be indexed by a vector index $[j, k]$ and written $\phi_{[j,k]}$. They are defined as

$$\phi_{[j,k]}(x) \equiv x_1^{k_1} \ldots x_l^{k_l} 1_{\Delta_j}(x) = x^k 1_{\Delta_j}(x),$$

where the set of mutually disjoint cubes Δ_j of equal volumes $|\Delta|$ forms a partition of $X = [0, 1]^l$, $1_{\Delta_j}(x)$ denotes the indicator function for the set Δ_j, and $k = [k_1, \ldots, k_l] \in \mathbb{Z}_+^l$ satisfies $|k| \equiv \sum_{i=1}^l k_i \leq r-1$, where r is the smoothness parameter in the definition of the target class $\mathcal{F} = W_\infty^{r,l}$. The volume $|\Delta|$ of every cube Δ_j is chosen such that the total number q of basis functions $\phi_{[j,k]}$ equals $n + d$; i.e., $q \equiv (1/|\Delta|)\alpha_{r,l} = n + d$, where $\alpha_{r,l}$ is the number of vectors $k \in \mathbb{Z}_+^l$ whose $|k| \leq r - 1$.

Define the linear operator $P_{\Delta_j} f$ as

$$\begin{aligned} P_{\Delta_j} f &= \sum_{k\colon |k|\leq r-1} b_{[j,k]}\phi_{[j,k]} \\ &= \sum_{k\colon |k|\leq r-1} b_{[j,k]} x^k 1_{\Delta_j}(x) \end{aligned}$$

and where the coefficients $b_{[j,k]}$ which depend on f are obtained by solving the following set of equations for the coefficients $b_{[j,k]}$:

$$\int_{\Delta_j} x^k \left(\sum_{k'\colon |k'|\leq r-1} b_{[j,k']} x^{k'} \right) dx = \int_{\Delta_j} x^k f(x)\, dx$$
$$\forall [j, k],\ 1 \leq j \leq q,\ |k| \leq r-1. \tag{20}$$

There are a total of $q\alpha_{r,l} = n + d$ such coefficients. Reindex these coefficients and their associated basis functions by an integer scalar and let $b(f) = [b_1(f), \ldots, b_{n+d}(f)]$ be the coefficient vector. We have a polynomial

$$\begin{aligned} p_f(x) &= \sum_{j=1}^{q} P_{\Delta_j} f(x) \\ &= \sum_{[j,k]} b_{[j,k]} x^k 1_{\Delta_j}(x) \\ &= \sum_{i=1}^{n+d} b_i(f)\phi_i(x), \end{aligned}$$

where $b_i(f)$ is the coefficient of the ith term. Define the information operator

$$\hat{N}_n(f) \equiv [b_1(f), \ldots, b_n(f)].$$

Continuing from (19) we have

$$\begin{aligned} &\sup_{y\in\hat{N}_n(\mathcal{F})} \sup_{f\in\mathcal{F}\cap\hat{N}_n^{-1}(y)} \inf_{h\in H_y^d} \|f-h\|_{L_\infty} && (21) \\ &\leq \sup_{y\in\hat{N}_n(\mathcal{F})} \sup_{f\in\mathcal{F}\cap\hat{N}_n^{-1}(y)} \left\| f - \sum_{i=1}^{n} y_i\phi_i(x) - \sum_{i=n+1}^{n+d} b_i(f)\phi_i(x) \right\|_{L_\infty} && (22) \\ &= \sup_{f\in\mathcal{F}} \left\| f - \sum_{i=1}^{n+d} b_i(f)\phi_i(x) \right\|_{L_\infty}; && (23) \end{aligned}$$

the last equality follows since for all $f \in \mathcal{F} \cap \hat{N}_n^{-1}(y)$, $\hat{N}_n(f) = [b_1(f), \ldots, b_n(f)] = y$. Now from Birman and Solomjak [5, Lemma 3.1] for every $f \in W_\infty^{r,l}$

$$\|f - P_{\Delta_j} f\| L_{\infty(\Delta_j)} \leq c_9 |\Delta|^{r/l} \|f\|_{W_\infty^{r,l}(\Delta_j)}$$

for a constant c_9, independent of j, and in our case, $\|f\|_{W_\infty^{r,l}(\Delta_j)} \equiv \sup_{x\in\Delta_j} |f^{(k)}(x)| \leq M$ for all k, $|k| \leq r$. Hence,

$$\begin{aligned} \left\| f - \sum_{i=1}^{n+d} b_i(f)\phi_i(x) \right\|_{L_\infty} &= \left\| f - \sum_{1\leq j\leq q} \sum_{k:\,|k|\leq r-1} b_{[j,k]} x^k 1_{\Delta_j}(x) \right\|_{L_\infty} && (24) \\ &= \max_j \sup_{x\in\Delta_j} \left| f(x) - \sum_k b_{[j,k]} x^k \right| && (25) \\ &= \max_j \sup_{x\in\Delta_j} |f(x) - P_{\Delta_j} f(x)| && (26) \\ &\leq c_9 |\Delta|^{r/l} M && (27) \\ &= c_{10} \frac{\alpha_{r,l}}{(n+d)^{r/l}} && (28) \\ &= \frac{c_{11}}{(n+d)^{r/l}} && (29) \end{aligned}$$

for positive constants c_9, c_{10}, and c_{11} independent of n and d. Hence, continuing from (23) we have

$$\sup_{f\in\mathcal{F}}\left\| f-\sum_{i=1}^{n+d} c_i(f)\phi_i(x)\right\|_{L_\infty} \le \frac{c_{11}}{(n+d)^{r/l}}$$

which proves the theorem. ■

A.3. *Proof of Theorem 4*

We first state several auxiliary lemmas. The following lemma is a consequence of Chebychev's inequality applied to a weighted sum of i.i.d. random variables; see, for instance, Petrov [24].

LEMMA 5. *For* $1 \le i \le m$, *let the independent random variables* x_i *be binomial on* $\{-1, +1\}$ *with probability* $\frac{1}{2}$ *and let* a_i *be constants such that* $\sum_{i=1}^m a_i^2 = 1$. *Then*

$$\mathbf{P}\left(\left|\sum_{i=1}^m a_i x_i\right| > \epsilon\sqrt{m}\right) \le 2e^{-m\epsilon^2/4}.$$

We now state a lemma concerning a lower bound on the distance

$$\operatorname{dist}(B_\infty^m(1)\cap L^n, \mathcal{H}^d, l_2)$$

between the intersection of a cube $B_\infty^m(1)$ and any fixed subspace of codimension n to any manifold $\underline{\mathcal{H}}^d \equiv \{\underline{h} = [h(x_1), \ldots, h(x_m)]: h \in \mathcal{H}^d\}$ in $\mathbb{R}^m$.

LEMMA 6. *Given integers* $m > \max\{320n \ln n, 32d \ln d\}$, $n \ge 20$, *and* $d \ge 1$. *Given a set of points* $\{x_1, \ldots, x_m\}$, *where* $x_i \in X$, $1 \le i \le m$. *Given a cube* $B_\infty^m(1) = [-1, 1]^m$ *and a subspace* L^n *of codimension n, both in* $\mathbb{R}^m$, *and a manifold* $\underline{\mathcal{H}}^d \subset \mathbb{R}^m$ *defined as* $\{\underline{h} = [h(x_1) \ldots h(x_m)]: h \in \mathcal{H}^d\}$, *where* $\mathcal{H}^d$ *is a class of functions with a pseudo-dimension d. Then*

$$\operatorname{dist}(B_\infty^m(1)\cap L^n, \underline{\mathcal{H}}^d, l_2) \ge \sqrt{m}/4.$$

Proof. Define by $V \equiv \{-1, 1\}^m$ the set of vertices of the cube $[-1, 1]^m$. Clearly $V \subset B_\infty^m(1)$. The subspace L^n in $\mathbb{R}^m$ is an $(m-n)$-dimensional subspace denoted as L_{m-n}. We first sketch the major steps in the proof. Begin by showing that there exists a vertex $v^* \in V$ which is $c_{12}\sqrt{m}$-close to L_{m-n} but is $c_{13}\sqrt{m}$-far from $\underline{\mathcal{H}}^d$ for some absolute constants $c_{13} > c_{12} > 0$. Denoting by y^* the point in L_{m-n} which is closest to v^* it then follows that

$\text{dist}(y^*, \underline{\mathcal{H}}^d, l_2) \geq c_{13}\sqrt{m} - c_{12}\sqrt{m}$. The proof is completed after showing that there exists a $\hat{y} \in B_\infty^m(1) \cap L_{m-n}$ which is close enough to y^*.

First we show that there are exponentially many vertices which are close to L_{m-n}. Fix any such subspace L_{m-n}. We have $L_{m-n} = \{x: w_1^T x = 0, \ldots, w_n^T x = 0\}$. For any vertex $v \in V$, we have

$$\text{dist}^2(v, L_{m-n}, l_2) = |P_{L_n} v|^2, \tag{30}$$

where L_n denotes the subspace which is orthogonal to L_{m-n}. Now let $u_1, \ldots, u_n$ be an orthonormal basis of L_n. Then the right-hand side of (30) becomes simply

$$|P_{L_n} v|^2 = \sum_{i=1}^n |(v, u_i)|^2.$$

We use a probabilistic argument to calculate the number of vertices that are $c_{12}\sqrt{m}$-close to L_{m-n}.

Draw uniformly a vertex v from V, i.e., pick its ith elements from $\{-1, +1\}$ with probability $\frac{1}{2}$, and repeat this for all $1 \leq i \leq m$ independently. Clearly, the number of vertices whose distance from L_{m-n} is greater than $c_{12}\sqrt{m}$ equals $\mathbf{P}(\{v \in V: \text{dist}(v, L_{m-n}, l_2) > c_{12}\sqrt{m}\})2^m$, where $\mathbf{P}$ is the uniform distribution over V. We can therefore upper bound this number by finding an upper bound on the probability $\mathbf{P}(\{v \in V: \text{dist}(v, L_{m-n}, l_2) > c_{12}\sqrt{m}\})$. We have

$$\begin{aligned}
&\mathbf{P}(\{v \in V: \text{dist}(v, L_{m-n}, l_2) > c_{12}\sqrt{m}\}) && (31)\\
&\quad = \mathbf{P}(\{v \in V: \text{dist}^2(v, L_{m-n}, l_2) > (c_{12}\sqrt{m})^2\}) && (32)\\
&\quad = \mathbf{P}\left(\sum_{i=1}^n |(v, u_i)|^2 > (c_{12}\sqrt{m})^2\right) && (33)\\
&\quad \leq \sum_{i=1}^n \mathbf{P}\left(|(v, u_i)|^2 > \frac{(c_{12}\sqrt{m})^2}{n}\right) && (34)\\
&\quad = \sum_{i=1}^n \mathbf{P}\left(|(v, u_i)| > \frac{c_{12}\sqrt{m}}{\sqrt{n}}\right) && (35)\\
&\quad \leq 2ne^{-mc_{12}^2/4n}, && (36)
\end{aligned}$$

where (36) follows from Lemma 5. As $c_{12} > 0$ is arbitrary we may choose $c_{12} = \frac{1}{8}$. (This particular choice is used further below.) Choose $m \geq 320n \ln n$. Then

$$2ne^{-mc_{12}^2/4n} \leq 2ne^{-320n \ln n/256n} = 2ne^{-80 \ln n/64} = 2\left(\frac{1}{n}\right)^{1/4}.$$

Hence, the number of vertices $v \in V$ such that $\text{dist}(v, L_{m-n}, l_2) \leq c_{12}\sqrt{m}$ is at least

$$2^m\left(1-2\left(\frac{1}{n}\right)^{1/4}\right) \tag{37}$$

under the constraint that $m \geq 320n \ln n$.

Consider the manifold $\underline{\mathcal{H}}^d \subset \mathbb{R}^m$ defined as

$$\underline{\mathcal{H}}^d \equiv \{\underline{h} = [h(x_1)\ \ldots\ h(x_m)] \colon h \in \mathcal{H}^d\},$$

where $\mathcal{H}^d$ is the class of functions with pseudo-dim d. For a vector $\underline{h}$ let $\operatorname{sgn}(\underline{h}) = [\operatorname{sgn} h(x_1) \ldots \operatorname{sgn} h(x_m)]$. Then since $d = \dim_p(\mathcal{H}^d)$ and from Sauer's lemma (cf. Haussler, [16, Lemma 3]) it follows that $\{\operatorname{sgn}(\underline{h}) \colon \underline{h} \in \underline{\mathcal{H}}^d\}$ has cardinality

$$\sum_{k=0}^{d}\binom{m}{k} \leq \left(\frac{em}{d}\right)^d. \tag{38}$$

This clearly implies that the manifold $\mathcal{H}^d$ intersects $q_m \leq (em/d)^d$ orthants. Every vertex corresponds to a unique orthant. Denote by $A = \{Q_i\}_{i=1}^{q_m}$ and $B = \{v_i\}_{i=1}^{q_m}$, the orthants which are intersected by the manifold and their corresponding vertices, respectively.

Denote by C the set of vertices $v \in V$ such that for each $v \in C$ there exists some $v_i \in B$ such that $\|v_i - v\|_1 \leq 2k$; i.e., v and v_i differ on at most k vector elements. We also have

$$|C| \leq \left(\frac{em}{d}\right)^d \sum_{i=0}^{k}\binom{m}{i}.$$

To simplify this expression we may choose $k = m/4$ and, using a bound on the tails of the binomial distribution (see, for example, Hoeffding [17]), the number of vertices in C may be bounded from above by $(em/d)^d\, 2^m \mathrm{e}^{-m/8}$.

From (37) the number of vertices $v \in V$ that have the property $\operatorname{dist}(v, L_{m-n}, l_2) \leq c_{12}\sqrt{m}$ is at least $2^m(1 - 2(1/n)^{1/4})$. Even if all the vertices in C have this property we are still left with at least

$$\begin{aligned}2^m\left(1-2\left(\frac{1}{n}\right)^{1/4}\right) - |C| &\geq 2^m\left(1-2\left(\frac{1}{n}\right)^{1/4}\right) - \left(\frac{em}{d}\right)^d 2^m e^{-m/8}\\ &= 2^m\left(1-2\left(\frac{1}{n}\right)^{1/4} - \left(\frac{em}{d}\right)^d e^{-m/8}\right)\end{aligned} \tag{39}$$

vertices which are not in C and which satisfy this property. For all $m \geq 64d \ln d$, $(em/d)^d e^{-m/8} \leq e^{-m/16}$. Thus (39) is lower bounded by

$$2^m \left(1 - 2\left(\frac{1}{n}\right)^{1/4} - e^{-m/16}\right).$$

Taking $m \geq \max\{320n \ln n, 64d \ln d\}$ then $(1 - 2(1/n)^{1/4} - e^{-m/16})$ is greater than 1 for all $n \geq 20$; thus, the total expression is larger than 1 for $n \geq 20$. Thus there exists at least one vertex which is not in C and which satisfies the property above. Moreover, for any such vertex $v \notin C$ and for any point z on the manifold $\underline{\mathcal{H}}^d$ the l_2 Euclidean distance

$$\|z - v\|_{l_2}^2 = \sum_{i=1}^{m} |z_i - v_i|^2 > k \cdot 1^2 = \frac{m}{4}.$$

Thus we have proved that there is at least one vertex $v^* \in V$ such that $\text{dist}(v^*, L_{m-n}, l_2) \leq c_{12}\sqrt{m}$ and that $\text{dist}(v^*, \underline{\mathcal{H}}^d, l_2) \geq \sqrt{m}/2$. The constant c_{13} mentioned earlier is $\frac{1}{2}$.

Finally, we wish to show that there exists a point $\hat{y}$ in the intersection $B_\infty^m(1) \cap L_{m-n}$ between the cube of side 2 and the linear space L_{m-n} such that $\text{dist}(\hat{y}, \underline{\mathcal{H}}^d, l_2) \geq c_{14}\sqrt{m}$ for some $c_{14} > 0$. For this we first show that there is a y^* in the intersection $B_2^m(\sqrt{m}) \cap L_{m-n}$ of the ball of radius $\sqrt{m}$ and the subspace L_{m-n} such that $\text{dist}(y^*, \mathcal{H}^d, l_2) \geq c_{14}\sqrt{m}$. Consider the point on L_{m-n} closest to the vertex v^*. Clearly, $\|v^* - y^*\|_{l_2} \leq c_{12}\sqrt{m}$. Moreover, y^* is simply the projection of v^* on L_{m-n}. As L_{m-n} goes through the origin and as $v^* \in B_2^m(\sqrt{m})$ it follows that y^* must be contained in $B_2^m(\sqrt{m}) \cap L_{m-n}$ (but not necessarily in $B_\infty^m(1)$). By a geometric argument one can show that there exists a point $\hat{y} \in B_\infty^m(1) \cap L_{m-n}$ which is no farther than $c_{12}\sqrt{m}$ from y^*. Also, we have for any $z \in \underline{\mathcal{H}}^d$

$$\|\hat{y} - z\|_{l_2} \geq \|v^* - z\|_{l_2} - \|y^* - v^*\|_{l_2} - \|\hat{y} - y^*\|_{l_2};$$

thus,

$$\begin{aligned}\inf_{z \in \underline{\mathcal{H}}^d} \|\hat{y} - z\|_{l_2} &\geq \inf_{z \in \underline{\mathcal{H}}^d} \|v^* - z\|_{l_2} - \|y^* - v^*\|_{l_2} - \|\hat{y} - y^*\|_{l_2} \\ &\geq \frac{\sqrt{m}}{2} - c_{12}\sqrt{m} - c_{12}\sqrt{m} \\ &\geq \frac{\sqrt{m}}{4}\end{aligned}$$

by the previous choice of $c_{12} = \frac{1}{8}$.

We have proved that there exists a point $\hat{y} \in B_\infty^m(1) \cap L_{m-n}$ such that $\inf_{z \in \underline{\mathcal{H}}^d} \|\hat{y} - z\|_{l_2} \geq \sqrt{m}/4$. Finally, we therefore conclude that

$$\sup_{y \in B_\infty^m(1) \cap L_{m-n}} \inf_{z \in \underline{\mathcal{H}}^d} \|y - z\|_{l_2} \geq \frac{\sqrt{m}}{4};$$

i.e.,

$$\operatorname{dist}(B_{\infty}^{m}(1) \cap L_{m-n},\ \underline{\mathcal{H}}^{d},\ l_2) \geq \frac{\sqrt{m}}{4}. \quad \blacksquare$$

COROLLARY 1. *For the same setting as in Lemma* 6, *the distance measured in the l_∞-norm is lower bounded as*

$$\operatorname{dist}(B_{\infty}^{m}(1) \cap L^{n},\ \underline{\mathcal{H}}^{d},\ l_\infty) \geq \tfrac{1}{4}.$$

Proof. For any vectors $a,\ b \in \mathbb{R}^m$, if $\|a-b\|_{l_2}^2 \geq m/16$ then at least one component $|a_j - b_j|^2 \geq \frac{1}{16}$ which implies that $\|a-b\|_{l_\infty} \geq \frac{1}{4}$. $\blacksquare$

We now prove Theorem 4.

Proof. We have

$$\inf_{N_n} \sup_{y \in \mathbb{R}^n} \inf_{\mathcal{H}^d} \sup_{\mathcal{F} \cap N^{-1}(y)} \inf_{h \in \mathcal{H}^d} \|f-h\|_{L_\infty} \tag{40}$$

$$\geq \inf_{N_n} \inf_{\mathcal{H}^d} \sup_{\mathcal{F} \cap N^{-1}(0)} \inf_{h \in \mathcal{H}^d} \|f-h\|_{L_\infty} \tag{41}$$

$$\geq \inf_{N_n} \inf_{\mathcal{H}^d} \sup_{\mathcal{F} \cap N^{-1}(0)} \inf_{h \in \mathcal{H}^d} \max_{1 \leq j \leq m} |f(x_j) - h(x_j)|, \tag{42}$$

where the set of m points x_j uniformly partition the domain X and we may use an integer vector to index a point as $x_{\underline{j}} = [x_{1,\,j_1}, \ldots, x_{l,\,j_l}]$, where $x_{i,\,j_i} = j_i/m^{1/l} + 1/2m^{1/l}$, $0 \leq j_i \leq m^{1/l} - 1$; $1 \leq i \leq l$.

We now define a subset $F_m \subset \mathcal{F} \equiv W_{\infty}^{r,\,l}$ such that the set of vectors

$$\{\underline{f} = [f(x_1), \ldots, f(x_m)]\colon f \in F_m\}$$

maps onto the cube $B_{\infty}^{m}(1/m^{r/l}) = [-1/m^{r/l},\ 1/m^{r/l}]^m$. For this, fix any function $\phi \in W_{\infty}^{r,\,1}(M)$ with support on $[0,\ 1]$ which satisfies $\phi(0) = \phi(1) = 0$, and $\phi\left(\frac{1}{2}\right) = 1$. Let $m' = m^{1/l}$, $E = \{0,\ 1, \ldots, m'-1\}^l$, $\phi_{i_j}(y) \equiv \phi(m'y - i_j)$, $0 \leq i_j \leq m' - 1$, $1 \leq j \leq 1$, and

$$\phi_{\underline{i}}(x) \equiv \phi_{i_1}(x_1) \ldots \phi_{i_l}(x_l) = \phi(m'x_1 - i_1)\phi(m'x_2 - i_2) \ldots \phi(m'x_l - i_l).$$

We define

$$F_m \equiv \left\{ f_{\underline{a}}(x) = \frac{1}{m^{r/l}} \sum_{\underline{i} \in E} a_{\underline{i}} \phi_{\underline{i}}(x)\colon a_{\underline{i}} \in [-1,\ 1] \right\}.$$

We will sometimes index the elements by a scalar integer and write for the vector $\underline{a} \equiv [a_1, \dots, a_m]$. First it is shown that for any $\underline{a} \in [-1, 1]^m$, $f_{\underline{a}} \in W_\infty^{r,l}(M)$. For this it suffices to upper bound $\sup_x |f_{\underline{a}}^{(\alpha)}(x)|$ by M, for $\alpha = [\alpha_1, \dots, \alpha_l]$, $\alpha_i \in \mathbb{Z}_+$, $\sum_{i=1}^{l} \alpha_i = r$. We have

$$\begin{aligned}
\sup_{x\in[0,1]^l} |f_{\underline{a}}^{(\alpha)}(x)| &= \frac{1}{m^{r/l}} \sup_{x\in[0,1]^l} \left| \sum_{\underline{i}\in E}^{m} a_{\underline{i}} \phi_{\underline{i}}^{(\alpha)}(x) \right| \\
&= \frac{1}{m^{r/l}} \max_{\underline{j}\in E} \sup_{x\in\Delta_{\underline{j}}} \left| \sum_{\underline{i}\in E} a_{\underline{i}} \phi_{\underline{i}}^{(\alpha)}(x) \right| \\
&= \frac{1}{m^{r/l}} \max_{\underline{j}\in E} \sup_{x\in\Delta_{\underline{j}}} |a_{\underline{j}} \phi_{\underline{j}}^{(\alpha)}(x)| \\
&= \frac{1}{m^{r/l}} \max_{\underline{j}\in E} |a_{\underline{j}}| \sup_{x\in\Delta_{\underline{j}}} |\phi_{\underline{j}}^{(\alpha)}(x)| \\
&= \frac{1}{m^{r/l}} \max_{\underline{j}\in E} |a_{\underline{j}}| \sup_{x\in\Delta_{\underline{j}}} |\phi^{(\alpha_1)}(m'x_1 - j_1)\phi^{(\alpha_2)} \\
&\qquad \times (m'x_2 - j_2) \dots \phi^{(\alpha_1)}(m'x_l - j_l)| \\
&= \frac{1}{m^{r/l}} \max_{\underline{j}} |a_{\underline{j}}| m'^r \sup_{x\in[0,1]^l} |\phi^{(\alpha)}(x)| \\
&\le \sup_{x\in[0,1]^l} |\phi^{(\alpha)}(x)| \le M;
\end{aligned}$$

the last line follows since by assumption $\phi \in W_\infty^{r,l}(M)$.

By a similar argument it may be shown that

$$\| f_{\underline{a}} \|_{L_\infty} = \sup_{x\in[0,1]^l} |f_{\underline{a}}(x)| = \frac{1}{m^{r/l}} \max_{\underline{j}\in E} |a_{\underline{j}}| = \frac{1}{m^{r/l}} \|\underline{a}\|_{l_\infty} = \| \underline{f}_{\underline{a}} \|_{l_\infty},$$

where we used the fact that

$$f_{\underline{a}}(x_{\underline{j}}) = \frac{1}{m^{r/l}} \sum_{\underline{i}\in E} a_{\underline{i}} \phi_{\underline{i}}(x_{\underline{j}}) = \frac{1}{m^{r/l}} a_{\underline{j}}.$$

Continuing from (42) we will drop the subscript $\underline{a}$ and just write f for any function in F_m. Using a scalar index j for the points x_j we have

$$\begin{aligned}
&\inf_{N_n} \inf_{\mathcal{H}^d} \sup_{f\in\mathcal{F}\cap N^{-1}(0)} \inf_{h\in\mathcal{H}^d} \max_{1\le j\le m} f(x_j) - h(x_j)| && (43)\\
&\quad \ge \inf_{N_n} \inf_{\mathcal{H}^d} \sup_{f\in F_m\cap N^{-1}(0)} \inf_{h\in\mathcal{H}^d} \max_{1\le j\le m} |f(x_j) - h(x_j)| && (44)\\
&\quad = \inf_{N_n} \inf_{\mathcal{H}^d} \sup_{\underline{f}\in B_\infty^m(1/m^{r/l})\cap L^n} \inf_{\underline{h}\in H^d} \|\underline{f} - \underline{h}\|_{l_\infty}, && (45)
\end{aligned}$$

where L^n is a subspace in $\mathbb{R}^m$ of codimension n and we used the fact that the set of vectors

$$\begin{aligned}
&\{\underline{f}\colon f \in F_m \cap N^{-1}(0)\} \\
&= \left\{\underline{f}\colon f(x) = \frac{1}{m^{r/l}} \sum_{\underline{j} \in E} a_{\underline{j}} \phi_{\underline{j}}(x),\ L_1(f) = 0, \ldots, L_n(f) = 0,\ \underline{a} \in [-1, 1]^m\right\} \\
&= \left\{\underline{f}\colon \underline{f} = \frac{1}{m^{r/l}} \underline{a},\ w_1^T \underline{a} = 0, \ldots, w_n^T \underline{a} = 0,\ \underline{a} \in B_\infty^m(1)\right\} \\
&= B_\infty^m\left(\frac{1}{m^{r/l}}\right) \cap L^n,
\end{aligned}$$

where by definition

$$w_i = [L_i(\phi_1), \ldots, L_i(\phi_m)], \qquad 1 \le i \le n.$$

Using Corollary 1 and continuing from (45)

$$\inf_{N_n} \inf_{\mathcal{H}^d} \sup_{\underline{f} \in B_\infty^m(1/m^{r/l}) \cap L^n} \inf_{\underline{h} \in H^d} \|\underline{f} - \underline{h}\|_{l_\infty} \ge \frac{1}{4m^{r/l}}.$$

We may substitute $m = 320n \ln n + 32d \ln d$ and still satisfy the conditions of Lemma 6 and Corollary 1. Thus we conclude that

$$\inf_{N_n} \sup_{y \in \mathbb{R}^n} \inf_{\mathcal{H}^d} \sup_{\mathcal{F} \cap N^{-1}(y)} \inf_{h \in \mathcal{H}^d} \|f - h\|_{L_\infty} \ge \frac{1}{4(320n \ln n + 32d \ln d)^{r/l}}. \quad \blacksquare$$

ACKNOWLEDGMENTS

J. Ratsaby acknowledges the support of the Ministry of the Sciences and Arts of Israel and the Ollendorff center of the Faculty of Electrical Engineering at the Technion. V. Maiorov acknowledges the Ministry of Absorption of Israel.

REFERENCES

1. Abu-Mostafa, Y. S. (1990), Learning from hints in neural networks, *J. Complexity* **6,** 192–198.
2. Abu-Mostafa, Y. S. (1993), Hints and the VC dimension, *Neural Comput.* **5,** 278–288.
3. Abu-Mostafa, Y. S. (1995), Machines that learn from hints, *Sci. Am.* **272,** No. 4.

4. Alon, N., Ben-David, S., Cesa-Bianchi, N., and Haussler, D. (1993), Scale-sensitive dimensions, uniform convergence, and learnability, in "Proceedings, 34th Annual Symposium on Foundations of Computer Science," pp. 292–301, IEEE Comput. Soc. Press, Los Alamitos, CA.

4a. Bartlett, P. L., Long, P. M., and Williamson, R. C., (1994), Fat-shattering and the learnability of real-valued functions, in "Proceedings of the 7th Annual Conference on Computational Learning Theory," p. 299, ACM, New York, NY.

5. Birman, M. S., and Solomjak, M. Z. (1967), Piecewise-polynomial approximations of functions of the classes, W_p^α, *Math. USSR-Sb.* **2,** No. 3, 295–317.

6. Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. (1989), Learnability and the Vapnik–Chervonenkis dimension, *J. Assoc. Comput. Mach.* **36,** No. 4, 929–965.

7. Buescher, K. L., and Kumar, P. R. (1996), Learning by canonical smooth estimation, Part I: Simultaneous estimation, *IEEE Trans. Automat. Control* **41,** No. 4, 545.

8. Cohn, D. (1996), Neural network exploitation using optimal experiment design, *Neural Networks* **9,** No. 6, 1071–1083.

9. Cohn, D., Atlas, L., and Ladner, R. (1994), Improving generalization with active learning, *Mach. Learning* **15,** 201–221.

10. Devore, R. A. (1989), Degree of nonlinear approximation, in "Approximation Theory VI," Vol. 1, Chui, C. K., Schumaker, L. L., and Ward, J. D. (Eds.), pp. 175–201, Academic Press, San Diego.

11. Devroye, L. (1987), A course in density estimation, in "Progress in Probability and Statistics," Vol. 14, Birkhauser, Boston.

12. Devroye, L., Gyorfi, L., and Lugosi, G. (1996), "A Probabilistic Theory of Pattern Recognition," Springer-Verlag, New York/Berlin.

13. Duda, R. O., and Hart, P. E. (1973), "Pattern Classification and Scene Analysis," Wiley, New York.

14. Fukunaga, K. (1990), "Introduction to Statistical Pattern Recognition," Academic Press, Boston.

15. Hanson, S., Petsche, T., Kearns, M., and Rivest, R. (1996), "Computational Learning Theory and Natural Learning Systems," MIT Press, Cambridge, MA.

16. Haussler, D. (1992), Decision theoretic generalizations of the PAC model for neural net and other learning applications, *Inform. Comput.* **100,** No. 1, 78–150.

17. Hoeffding, W. (1963), Probability inequalities for sums of bounded random variables, *J. Amer. Statist. Assoc.* **58,** No. 301, 13–30.

18. Kearns, M. J., and Vazirani, U. (1997), "An Introduction to Computational Learning Theory," MIT Press.

19. Linhart, H., and Zucchini, W. (1986), "Model Selection," Wiley Series in Probability and Mathematical Statistics, Wiley, New York.

20. Maiorov, V. E. (1996), On best approximation by ridge functions, *J. Approx. Theory,* in press.

21. Maiorov, V. E., Meir, R., and Ratsaby, J. (1996), On the approximation of functional classes equipped with a uniform measure using ridge functions, *J. Approx. Theory*, in press.

22. Maiorov, V. E., and Wasilkowski, G. W. (1996), Probabilistic and average linear widths in L_∞-norm with respect to r-fold Wiener measure, *J. Approx. Theory* **84,** No. 1.

23. Maiorov, V., and Ratsaby, J. (1997), On the degree of approximation using manifolds of finite pseudo-dimension, *J. Constr. Approx.,* in press.

24. Petrov, V. V. (1975), "Sums of Independent Random Variables," Springer-Verlag, Berlin.

25. Pinkus, A. (1985), "*n*-widths in Approximation Theory," Springer-Verlag, New York.

26. Pollard, D. (1984), “Convergence of Stochastic Processes,” Springer Series in Statistics, Springer-Verlag, Berlin/New York.

27. Pollard, D. (1989), “Empirical Processes: Theory and Applications,” NSF-CAMS Regional Conference Series in Probability and Statistics, Vol. 2, Inst. Math. Stat. Am. Stat. Assoc., Hayward, CA.

28. Ratsaby, J. (1994), “The Complexity of Learning from a Mixture of Labeled and Unlabeled Examples,” Ph.D. thesis, University of Pennsylvania.

29. Ratsaby, J., Meir, R., and Maiorov, V. (1996), Towards robust model selection using estimation and approximation error bounds, in “Proc. 9th Annual Conference on Computational Learning Theory,” p. 57, ACM, New York, NY.

30. Ratsaby, J., and Venkatesh, S. S. (1995), Learning from a mixture of labeled and unlabeled examples, invited paper, in “Proc. 33rd Allerton Conferene on Communication Control and Computing,” (pp. 1002–1009).

31. Ratsaby, J., and Venkatesh, S. S. (1995), Learning from a mixture of labeled and unlabeled examples with parametric side information, in “Proc. 8th Annual Conference on Computational Learning Theory,” p. 412, Morgan Kaufmann, San Maeto, CA.

32. Ripley, B. D. (1996), “Pattern Recognition and Neural Networks,” Cambridge University Press, Cambridge, UK.

33. Tikhomirov, V. M. (1976), “Some Problems in Approximation Theory,” Moscow State University, Moscow. [Russian].

34. Traub, J. F., Wasilkowski, G. W., and Wozniakowski, H. (1988), “Information-Based Complexity,” Academic Press, San Diego.

35. Valiant, L. G. (1984), A theory of the learnable, *Comm. ACM* **27,** No. 11, 1134–1142.

36. Vapnik, V. N., and Chervonenkis, A. Ya. (1971), On the uniform convergence of relative frequencies of events to their probabilities, *Theoret. Probl. Appl.* **16,** No. 2, 264–280.

37. Vapnik, V. N., and Chervonenkis, A. Ya. (1981), Necessary and sufficient conditions for the uniform convergence of means to their expectations, *Theoret. Probl. Appl.* **26,** No. 3, 532–553.

38. Vapnik, V. N. (1982), “Estimation of Dependences Based on Empirical Data,” Springer-Verlag, Berlin.

39. Warren, H. E. (1968), Lower bounds for approximation by non-linear manifolds, *Trans. Amer. Math. Soc.* **133,** 167–178.