Journal of Computer and System Sciences **58**, 183–192 (1999)
Article ID jcss.1998.1604, available online at http://www.idealibrary.com on IDEAL®

On the Learnability of Rich Function Classes

Joel Ratsaby

Manna Network Technologies, Tel Aviv 61574, Israel

and

Vitaly Maiorov

Department of Mathematics, Technion, Haifa 32000, Israel

The probably approximately correct (PAC) model of learning and its extension to real-valued function classes sets a rigorous framework based upon which the complexity of learning a target from a function class using a finite sample can be computed. There is one main restriction, however, that the function class have a finite VC-dimension or scale-sensitive pseudo-dimension. In this paper we present an extension of the PAC framework with which rich function classes with possibly infinite pseudo-dimension may be learned with a finite number of examples and a finite amount of partial information. As an example we consider learning a family of infinite dimensional Sobolev classes.

Key Words: PAC learning; computational learning theory; information-based complexity; VC-theory; approximation theory; partial information.

1. INTRODUCTION

Valiant [31] and Blumer, Ehrenfeucht, Haussler and Warmuth [8] introduced the probably approximately correct (PAC) learning model. In its basic form, there is an abstract teacher providing the learner a finite number of examples of an unknown target function $g(x)$ which is a set-indicator function over some domain $\mathscr{X}$. Based on a sequence of examples $\{(x_i, y_i)\}_{i=1}^m$, $x_i \in \mathscr{X}$, $y_i \in \mathscr{Y}$, where $y_i = g(x_i)$, $1 \leqslant i \leqslant m$, and $\mathscr{Y} = \{0, 1\}$, which are randomly drawn according to an unknown underlying distribution over $\mathscr{X}$, the aim is to learn or estimate the target to within a prespecified arbitrary accuracy $\varepsilon > 0$ and confidence $1-\delta$.

In this paper we adopt the extension of the PAC model to *real* valued functions (cf. Haussler [13]). We assume an unknown probability distribution P on $\mathscr{X} \times \mathscr{Y}$, where $\mathscr{Y} = \mathbb{R}$ and the notation P is used for all joint, marginal and conditional probability distributions. In the general case, the *target* is defined as a deterministic function $g(x) = \mathrm{E}(Y \mid X = x)$ and the data sample consists of $\{(x_i, y_i)\}_{i=1}^m$, y_i are possibly noisy versions of $g(x_i)$, where (x_i, y_i), $1 < i < m$, are drawn i.i.d. according to P over $\mathscr{X} \times \mathscr{Y}$. The learner has a known *hypothesis* class $\mathscr{H}$ which does not necessarily contain the target g. A *loss* function $L_q(h)$ measures the expected dissimilarity between the random variable Y and $h(X)$ and is defined as $L_q(h) = (\mathrm{E}\,|Y - h(X)|^q)^{1/q}$ for any $q \geqslant 1$, where expectation is taken with respect to P. Based on the sample the *empirical loss* is defined as $\hat{L}_q(h) = ((1/m) \sum_{i=1}^m (y_i - h(x_i))^q)^{1/q}$. The specific noise-free case amounts to $Y = g(X)$, the loss being $L_q(h) = (\mathrm{E}\,|g(X) - h(X)|^q)^{1/q}$ and the sample consists of $(x_i, g(x_i))$, $1 \leqslant i \leqslant m$. We will also denote $L_q(h) = \|h - g\|_{L_q(P)}$ which is the L_q-norm with respect the probability distribution P.

We limit our results here to the noise-free case since learning with noise amounts to adding a constant variance-like term to the loss of any hypothesis, in particular, to the one which minimizes the empirical loss. We will assume a compact domain $\mathscr{X}$ and a probability distribution P on $\mathscr{X}$ with a density function $dP(x)$ bounded over $\mathscr{X}$. This allows a direct upper bound on the loss $L_q(h)$ in terms of a constant multiple of $\|g - h\|_{L_q}$, where the latter denotes the L_q norm with respect to the uniform probability distribution over $\mathscr{X}$. When referring to the classical PAC learning model we adhere to the following definition.

Definition 1 (PAC-Learnability). For arbitrary $\varepsilon > 0$, $0 < \delta < 1$, a hypothesis class $\mathscr{H}$ is (ε, δ)-*learnable* (or PAC-learnable) if for any target function g and any probability distribution P on $\mathscr{X}$ there exists a learning algorithm which can find based on a finite number $m(\varepsilon, \delta)$ of i.i.d. examples $\{(x_i, g(x_i))\}_{i=1}^m$, $x_i \in \mathscr{X}$, a hypothesis $\hat{h} \in \mathscr{H}$ whose loss satisfies

$$L_q(\hat{h}) \leqslant L^*_{g,q} + \varepsilon$$

with probability greater than $1-\delta$, where $L^*_{g,q} = \inf_{h \in \mathscr{H}} \|g - h\|_{L_q(P)}$ denotes the loss of an optimal hypothesis in $\mathscr{H}$. We call $m(\varepsilon, \delta)$ the sample *complexity* of learning $\mathscr{H}$.

Remark. Note that if $g \in \mathscr{H}$ then $L^*_{g,q} = 0$. As in the PAC-model we henceforth restrict our framework to algorithms which learn by minimization of the empirical loss,

0022-0000/99 $30.00

that is, $\hat{h}$ is obtained by picking any hypothesis h in $\mathscr{H}$ which has a minimal $\hat{L}_q(h)$.

The PAC-model applies to many learning problems, for instance, to the classical problem of learning pattern recognition, where the target function is a classifier and is represented by a set-indicator function $g(x) = 1_{\{x \in A_g\}}$, $A_g \subset \mathscr{X}$. Here the loss of an indicator function hypothesis h is taken as $L_1(h) = \mathrm{E}\,|h(x) - g(x)|$ which equals the probability of the symmetric difference between the set A_h and A_g. In learning regression the learner has access to noisy examples $\{(x_i, y_i)\}_{i=1}^m$ of the target g, where $g(x) = \mathrm{E}(Y \mid X = x)$, $y_i = g(x_i) + v_i$, and v_i, $1 \leqslant i \leqslant m$, are independent noise-random variables with zero mean. Here the quadratic loss functional $L_2(h) = \mathrm{E}\,|Y - h(X)|^2$ is used to measure the discrepancy of h. For other classical learning problems which can be modeled by the PAC framework see Haussler [13].

The primary contribution of the PAC model to the theory of pattern recognition and machine learning arises from stating a condition under which PAC-learnability may be attained, i.e, a guarantee that a target g can be learned to an accuracy which is arbitrarily close to the optimal loss $L^*_{g,q}$, based on a *finite* sample. This guarantee depends on whether the hypothesis class has a finite complexity which is measured by the quantity called *pseudo-dimension*. Pollard [22] and, later, Haussler [13] extended the well-known Vapnik–Chervonenkis dimension to the real-valued function class case, calling it the pseudo-dimension, $\dim_p(\mathscr{H})$, of a class $\mathscr{H}$. It is defined as follows: Let $\mathrm{sgn}(y)$ be defined as 1 for $y > 0$ and -1 for $y \leqslant 0$. For a Euclidean vector $v \in \mathbb{R}^m$ denote by $\mathrm{sgn}(v) = [\mathrm{sgn}(v_1), ..., \mathrm{sgn}(v_m)]$.

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

The importance of this quantity arises from the fact that a class $\mathscr{H}$ which has a finite pseudo-dimension is PAC-learnable.

The PAC model comes short of guaranteeing in general that a hypothesis $\hat{h}$ with a loss arbitrarily close to zero can be produced. It essentially ignores the value of the optimal loss $L^*_{g,q}$ and assumes it is either zero (as in the original PAC work [8]), where the target is a member of the hypothesis class, or that it is uncontrollable. In many cases the first assumption is too artificial while the second assumption is too strict, One way to decrease $L^*_{g,q}$ is to use a "smarter" learner, i.e., a hypothesis class $\mathscr{H}^d$ with a large pseudo-dimension d. Theoretically one could use an infinite-pseudo-dimensional class and reduce $L^*_{g,q}$ to zero; however, in general, such a class will not be PAC-learnable, (Another extension of the VC-dimension named the *scale-sensitive dimension*, (cf. Alon, Ben-David, Cesa-Bianchi, and Haussler [5]), which is a parameterized version of the pseudo-dimension, guarantees necessary and sufficient condition for PAC-learnability.)

Due to this it is common to use a nested structure of finite-pseudo-dimensional hypothesis classes $\{\mathscr{H}^d\}_{d=1}^{\infty}$ and then balance the trade-off between the learning accuracy $\varepsilon(m, d)$ and the optimal loss $L^*_{g,d,q}$, both of which depend, but in opposite directions, on the pseudo-dimension d. Such balancing leads to an optimal value d^* or an optimally complex hypothesis class. The theory of many statistical estimation methods is based on this idea. Methods such as Vapnik's structural risk minimization (Vapnik [32], Shawe-Taylor, Bartlett, Williamson, and Anthony [28]), regularization in statistical estimation (White [33], Lugosi and Zeger [18], Grenander [12]), and model selection (Barron [6], Lugosi and Nobel [17], Ratsaby, Meir, and Maiorov [24]) all consider a learner with a potentially infinite amount of resources which is optimally balanced against the variability introduced by the finite sample.

When g is assumed to be a member of a known rich target class $\mathscr{F}$ then a measure of the learner' s optimal learning ability in the infinite sample limit (assuming for the moment that the learner is limited to an hypothesis class $\mathscr{H}^d$ of pseudo-dimension d) is captured in the quantity $\sup_{g \in \mathscr{F}} L^*_{g,d,q}$ which is usually referred to as the approximation error of $\mathscr{F}$ by $\mathscr{H}^d$. The larger d, the richer the class $\mathscr{H}^d$, and the lower the error $\sup_{g \in \mathscr{F}} L^*_{g,d,q}$ The classical field of approximation theory (cf. Lorentz, Golitschek, and, Makovoz [16], Pinkus [21]) has many established results on the estimation of this error for numerous combinations of target classes $\mathscr{F}$ and hypothesis classes $\mathscr{H}^d$.

In this paper we consider another alternative for controlling $L^*_{g,d,q}$ which is conceptually different from the one above. Instead of enriching the learner by an infinite pseudo-dimensional hypothesis class structure we consider having a more helpful teacher. This is related to the notion of active-learning which studies the complexity of learning while possessing some knowledge about the target obtained through means which are more general than classical random sampling.

That partial information about a problem to be learnt is important can often be seen in humans, as well as machine learning. For instance, when learning using artificial neural networks one form of useful partial information about the target function is a good starting point in the error-surface descent. Several experimental results which demonstrate that partial knowledge helps include the work of Abu-Mostafa [1–3] who refers to partial knowledge as hints which are found to be useful, for instance, in financial prediction problems. Roscheisen, Hofmann, and Tresp [27] consider ways of incorporating partial knowledge into a system that learns by examples by resorting to a Bayesian model. A prior probability density over the target class is defined and a portion of the training sample is artificially generated using this prior knowledge. Towell and Shavlik [29] show how rule-based prior knowledge can be incorporated into neural networks consisting of sigmoidal units.

As part of the motivation for the interest in a theoretical learning framework which takes into account available partial information, we now consider several instances of learning problems with different types of partial information. The first two examples are classical problems from the field of pattern recognition.

EXAMPLE 1 (Parametric classification). The setting consists of two pattern classes with unknown class conditional probability densities $f_1(x)$, $f_2(x)$ over $\mathscr{X} = \mathbb{R}^l$ and *a priori* probabilities p_1, p_2. The *data* sample consists of labeled examples $\{(x_i, y_i)\}_{i=1}^m$, where y_i is first drawn from $\{1, 2\}$, taking the value 1 with probability p_1, and x_i is drawn with respect to $f_{y_i}(x_i)$, $1 \leqslant i \leqslant m$. The *target* $g(x)$ is the discriminate function corresponding to the Bayes optimal classifier which classifies the region $R_g = \{x \in \mathscr{X}: g(x) = \ln(p_1 f_1(x)/p_2 f_2(x)) \geqslant 0\}$ by "1" and the region $\mathscr{X} \backslash R_g$ by "2."

Assume that the target is contained in a parametric *hypothesis* class $\mathscr{H}^d$ which is a d-dimensional linear space of r-degree polynomials over $\mathscr{X}$, i.e., $\mathscr{H}^d = \{h_a(x) = \sum_{i:\, \|i\| \leqslant r} a_i x_1^{i_1} \cdots x_l^{i_l},\ a_i \in \mathbb{R}\}$, where for a nonnegative multiinteger $i \in \mathbb{Z}^l_+$ the norm $\|i\| = \sum_{j=1}^l |i_j|$. Here $d < \infty$ is the number of vectors i whose norm is less than or equal to r. The classifier corresponding to a hypothesis h is defined as labeling the region $R_h = \{x \in \mathscr{X} : h(x) \geqslant 0\}$ by "1" and the region $\mathscr{X} \backslash R_h$ by "2." For the target, use the notation g_a with parameter a as it is contained in $\mathscr{H}^d$. The *loss* of h is defined as $L_l(h) = \|1_{R_g} - 1_{R_h}\|_{L_1(P)}$, where 1_{R_h} is the indicator function for the set R_h.

The *problem* is to (ε, δ)-learn the class $\mathscr{H}^d$, i.e., output a hypothesis $\hat{h} \in \mathscr{H}^d$ such that $L_1(\hat{h}) \leqslant \varepsilon$, with probability $1 - \delta$ which implies that the probability of error of the classifier based on $\hat{h}$ is no more than ε from the Bayes error. (Note that the optimal loss $L^*_{g,1} = 0$ since the target $g \in \mathscr{H}^d$.)

Partial information: *feature selection*. An expert points to *interesting* feature components x_j, $j \in J$, where $J \subset \{1, 2, ..., l\}$, which are the most significant as far as the separability of the pattern classes is concerned. Denote by $I = \{i = [i_1, ..., i_l] \in \mathbb{Z}^l_+ : \|i\| \leqslant r,\ \exists j \notin J,\ i_j \neq 0\}$, let $n = |I|$ and I^c denotes the complement of I, i.e. $I^c = \{i = [i_1, ..., i_l] \in \mathbb{Z}^l_+ : \|i\| \leqslant r, i_j = 0, j \notin J\}$. Effectively, the expert information reduces the hypothesis class $\mathscr{H}^d$ to a subset $\mathscr{H}^{d-n} = \{h_a(x) = \sum_{i:\, i \in I^c} a_i x_1^{i_1} \cdots x_l^{i_l},\ a_i \in \mathbb{R}\}$, of dimensionality, or equivalently of pseudo-dimension, $d - n$, since the hypothesis class is a vector space of functions (cf. Theorem 4 of Haussler [31). Thus, given such partial information the learner knows that the Bayes optimal hypothesis g_a is contained in a smaller subset $\mathscr{H}^{d-n}$ of the hypothesis class $\mathscr{H}^d$. As such, the sample complexity of PAC-learning $\mathscr{H}^{d-n}$ is smaller than that of learning $\mathscr{H}^d$.

Note that partial information about g_a can be expressed as the value taken by a *linear* projection operator $N_I: \mathscr{H}^d \to \mathbb{R}^n$, where for a function $h_a \in \mathscr{H}^d$ we have $N_I(h_a) = [a_{k_1}, a_{k_2}, ..., a_{k_n}]$, $k_j \in I$, $1 \leqslant j \leqslant n$, and the information vector corresponding to the target is $N_I(g_a) = [0, ..., 0]$.

If we denote by $m(d, \varepsilon)$ the sample complexity of learning $\mathscr{H}^d$ then for information of size n (i.e., an information vector of dimensionality n), the reduction in the sample complexity is $m(d, \varepsilon) - m(d-n, \varepsilon)$. This represents the number of examples that information of size n is worth.

EXAMPLE 2 (Nonparametric classification). The setting consists of M unknown nonparametric class conditional probability distributions $f_j(x)$ over $\mathscr{X} = \mathbb{R}^l$ with their corresponding known *a priori* class probabilities p_j, $1 \leqslant j \leqslant M$. Denote the mixture probability density $dP(x) = \sum_{i=j}^M p_j f_j(x)$. The *target* is defined as the vector-valued function $g(x) = [f_1(x), ..., f_M(x)]$. The optimal Bayes classifier is defined as: Classify an x by "j^*" where $j^* = \operatorname{argmax}_{1 \leqslant j \leqslant M}\{p_j f_j(x)\}$. The *sample* consists of m i.i.d. pairs $\{(x_i, y_i)\}_{i=1}^m$, where $y_i \in \{1, 2, ..., M\}$ takes the value j with probability p_j, and x_i is drawn according to $f_{y_i}(x)$, $1 \leqslant i \leqslant m$. The learner has an *hypothesis* class $\mathscr{H} = \mathscr{H}_1 \times \cdots \times \mathscr{H}_M$ of vector-valued functions $h(x) = [h_1(x), ..., h_M(x)]$, where $\mathscr{H}_j$, $1 \leqslant j \leqslant M$, have finite pseudo-dimension.

The *problem* is to estimate each of the M probability densities g_j by h_j, $1 \leqslant j \leqslant M$. The *loss* of a hypothesis $h \in \mathscr{H}$ is defined as $L_q(h) = \sum_{i=1}^M p_j \|g_j - h_j\|_{L_q(P)}$ for any fixed $q \geqslant 1$, where $\|f\|_{L_q(P)} = (\int_{\mathscr{X}} |f(x)|^q \, dP(x))^{1/q}$. Stated in the PAC-framework, the problem is to (ε, δ)-learn $\mathscr{H}$, i.e., to find an $\hat{h} \in \mathscr{H}$ whose loss $L_q(\hat{h}) \leqslant L^*_{g,q} + \varepsilon$, with probability $1 - \delta$, where $L^*_{g,q} = \inf_{h \in \mathscr{H}} \|g - h\|_{L_q(P)}$.

Partial information: *feature extraction*. The learner is told which of the fewest k of the l feature components of x are the most significant as far as classification is concerned.

Using classical discriminate analysis methods (cf. Fukunaga [11], Duda and Hart [10]) such information about the target g may be expressed in terms of the class probabilities p_j, the class conditional means $\mu_j = E(X \mid j)$ and the class conditional covariance matrices $E((X - \mu_j)(X - \mu_j)^T \mid j)$, $1 \leqslant j \leqslant M$. This information is sufficient for defining one of the standard criterion functionals which relate the within-class and the between-class variability. One may then compute the optimal linear discriminate matrix A, of size $k \times l$, which maps a feature vector x to a lower dimensional $y \in \mathbb{R}^k$. The components of y are the k features which best preserve the pattern class separability in the lowest possible k-dimensional feature space. Such information can be represented by an n-dimensional vector of *linear* functionals acting on g, i.e., $N(g) = [\{\mu_{j,s}\}_{j=1,}^{M}{}_{s=1}^{l}, \{\sigma_{s,r}^{j}\}_{j=1,}^{M}{}_{s \leqslant r=1}^{l}]$ where $\mu_{j,s} = \int_{\mathscr{X}} x_s f_j(x)\,dx$, and $\sigma_{s,r}^{j} = \int_{\mathscr{X}} x_s x_r f_j(x)\,dx$. The dimensionality of the information vector is $n = (Ml/2)(l+3)$.

Example 3 (Nonparametric density estimation). The setting consists of an underlying probability density function $g(x) = dP(x)$ over $\mathscr{X}$, which is the *target* to be learned. The target is assumed to be a member of a rich nonparametric probability density class $\mathscr{F}$. The *data* sample consists of unlabeled examples $x_i \in \mathscr{X}$, $1 \leqslant i \leqslant m$, drawn according to $P(x)$. The *hypothesis* class $\mathscr{H}^d$ is a probability density function class of pseudo-dimension $d < \infty$. The *loss* is defined as $L_2(h) = \|g - h\|_{L_2(P)}$. The problem is to (ε, δ)-learn the class $\mathscr{H}^d$, i.e., output a hypothesis $\hat{h} \in \mathscr{H}^d$ such that $L_2(\hat{h}) \leqslant L_{g,2}^* + \varepsilon$, with probability $1 - \delta$, where $L_{g,2}^* = \inf_{h \in \mathscr{H}^d} \|g - h\|_{L_2(P)}$.

Linear partial information. A histogram density estimate (cf, Devroye, Gyorfi, and Lugosi [9]) based on a uniform partition π_n of $\mathscr{X}$ with equal-volume cells c_i, $1 \leqslant i \leqslant n$, of side s, i.e., of the type $[k_1 s, (k_1+1)\,s) \times \cdots \times [k_l s, (k_l+1)\,s)$, for a multiinteger $k \in \mathbb{Z}_+^l$. The learner is given partial information consisting of $P(c_i)$, $1 \leqslant i \leqslant n$. This partial information can be represented by a *linear* operator $N\colon \mathscr{F} \to \mathbb{R}^n$ taking the value $N(g) = [P(c_1), ..., P(c_n)]$, where $P(c_i) = \int_{c_i} g(x)\,dx$ is a linear functional of g.

In the previous example, a nonparametric histogram density estimator represented linear partial information about the target probability density. In general, nonparametric density estimation can be used in conjunction with parametric estimation algorithms as a basis for learning with partial information. Consider, for instance, the problem of adaptive equalization in data communication. Here the user's communication device is given only several hundreds of milliseconds to learn the equalizer parameters for the particular channel. It needs to relearn these parameters at the beginning of every communication session as they depend on the channel which in turn depends on the particular host computer's communication device, the condition of the communication link, etc. Thus training needs to be quick which restricts the algorithm to be of a parametric estimation type. For supervised learning, a parametric estimation method such as the least-mean-square (LMS) algorithm which does gradient descent on a quadratic loss surface is often used. The target to be learnt corresponds to some parametric estimator of the optimal equalizer. Partial information about a *typical* channel may help the parametric training procedure in learning the *particular* target equalizer corresponding to a specific session. Protocol handshaking between the devices on both ends provides the supervised sample which is restricted to be small due to the learning time limitation. In this problem, partial information can be injected through running a nonparametric density estimation algorithm in the background. This long-term learning algorithm has access to a much larger unsupervised random sample which is accumulated over numerous communication sessions during the operational history of the user's device. The session-dependent parametric learning algorithm can peek any time at the current version of this nonparametric estimate (for instance, a histogram estimate) and use it as partial information. This means that the sample size and, hence the training time needed by the parametric algorithm may be significantly reduced.

The previous examples included finite dimensional parametric target classes, as well as infinite dimensional nonparametric target classes, while the hypothesis classes were all of finite pseudo-dimension. This ensured that a hypothesis $\hat{h}$ with a loss close to the optimal may be learnt, based on a finite sample. It is intuitive that increasing the amount of partial information, as given in the various forms above, should make the loss of $\hat{h}$ be closer to zero (the absolute optimal) regardless of the richness of the target class $\mathscr{F}$.

The focus of this paper is to show that partial information allows extending the notion of PAC-learnability to target classes which are richer than the ones considered by the classical PAC framework, in particular, target classes having infinite pseudo-dimension. With the framework to be introduced in the next section, learnability is no longer restricted to a *hypothesis* class $\mathscr{H}$ as in Definition 1 but holds for the actual *target* class $\mathscr{F}$. A learner of finite capacity (i.e., possessing a hypothesis class of finite pseudo-dimension) is able to learn rich function classes (of infinite pseudo-dimension) using a finite sample and a finite amount of partial information. He may determine a hypothesis $\hat{h}$ with a loss $L_q(\hat{h})$ arbitrarily close to zero with arbitrarily high confidence. Based on this theoretical framework it is possible to quantitatively compare the value of partial information versus that of information contained in the random sample.

2. THE FRAMEWORK

The branch of the field of computational complexity known as information based complexity cf. Traub, Wasilkowski, and Wozniakowski [30] deals with the intrinsic difficulty of providing approximate solutions to problems for which information is partial, noisy, or costly. We borrow some of their definitions as applied to problems of approximating a general target function class $\mathscr{F}$. Let $N_n\colon \mathscr{F} \to \mathbb{R}^n$ denote a general information operator. The information $N_n(g)$ consists of n real-valued measurements taken on the target function g, or in general, any function $f \in \mathscr{F}$, i.e.,

$$N_n(f) = [Q_1(f), ..., Q_n(f)],$$

where $Q_i(\ )$, $1 \leqslant i \leqslant n$, are functionals over $\mathscr{F}$. We call n the *cardinality* of information and sometimes omit n and write $N(f)$. The variable y denotes an element in $N_n(\mathscr{F})$. The subset $\mathscr{F}_y = N_n^{-1}(y) \subset \mathscr{F}$ denotes all functions $f \in \mathscr{F}$ which share the same information vector y, i.e.

$$\mathscr{F}_y = \{ f \in \mathscr{F} : N_n(f) = y \}.$$

We denote by $N_n^{-1}(N_n(g))$ the *solution set* which may also be written as $\{ f \in \mathscr{F} : N_n(f) = N_n(g) \}$, which consists of all indistinguishable functions $f \in \mathscr{F}$ sharing the same information vector as the target g.

Information about g is represented by a partition $\Pi_{N_n}(\mathscr{F})$ of $\mathscr{F}$ as

$$\Pi_{N_n}(\mathscr{F}) = \bigcup_{y \in \mathbb{R}^n} \mathscr{F}_y.$$

Denote the distance between two function classes $\mathscr{A}$ and $\mathscr{B}$ by $\operatorname{dist}(\mathscr{A}, \mathscr{B}, L_q) = \sup_{\{a \in \mathscr{A}\}} \inf_{\{b \in \mathscr{B}\}} \|a - b\|_{L_q}$ for any $q \geqslant 1$.

Following the discussion in Section 1, we now present a sequence of definitions of entities which are analogous to those used in Definition 1, the main aim being to control the optimal loss $L^*_{g,q}$ in the PAC-learning setting to an extent which it can be arbitrarily small, regardless of whether the target class $\mathscr{F}$ has finite or infinite pseudo-dimension.

We start with the analog of an hypothesis class $\mathscr{H}$ which we define as a family $\mathscr{G}^d = \{ \mathscr{H}^d : \dim_p(\mathscr{H}^d) = d \}$ of all hypothesis classes of pseudo-dimension d. We define a *target subset* $\mathscr{F}_y \in \Pi_N(\mathscr{F})$ which is analogous to the fixed target g in Definition 1. The loss $L_q(h)$ of a hypothesis h has the following analog:

Definition 3 (Loss of a class $\mathscr{H}^d$). Fix a target subset $\mathscr{F}_y \subset \mathscr{F}$. The loss of $\mathscr{H}^d$ is

$$L_q(\mathscr{H}^d) = \operatorname{dist}(\mathscr{F}_y, \mathscr{H}^d, L_q).$$

Remark. As was the case for the loss of a hypothesis function whose notation depended implicitly on the target g, here in the case of the loss of a class $\mathscr{H}^d$ we also suppress the dependency on the target subset $\mathscr{F}_y$.

The next definition is analogous to the optimal hypothesis loss $L^*_{g,q}$ for a fixed target g.

Definition 4 (Optimal class-loss). For a fixed target subset $\mathscr{F}_y \subset \mathscr{F}$ and fixed positive integer $d < \infty$, define the optimal class loss as

$$L^*_{y,d,q} = \inf_{\{\mathscr{H}^d \in \mathscr{G}^d\}} L_q(\mathscr{H}^d).$$

Definition 5 (Loss of a partition). For a fixed positive integer $d < \infty$, a fixed information operator N, define the loss of a partition $\Pi_N(\mathscr{F})$ as the loss of the worst subset in the partition, i.e.,

$$L^d_q(\Pi_N) = \sup_y L^*_{y,d,q}.$$

Remark. We will also refer to $L^d_q(\Pi_N)$ as the loss of the information operator N.

Following the framework of information-based complexity (cf Traub *et. al*, [30])), we henceforth limit to linear information operators. A *linear* information operator $N: \mathscr{F} \to \mathbb{R}^n$ is a linear mapping satisfying $N(\alpha f + \beta f') = \alpha N(f) + \beta N(f')$ for any $f, f' \in \mathscr{F}$. The notion defined next is analogous to the condition for (ε, δ)-learnability of $\mathscr{H}$ in Definition 1.

Definition 6 ((d, ε)-approximability). A target class $\mathscr{F}$ is (d, ε) -approximable if for any fixed integer $d < \infty$, for all $\varepsilon > 0$, and for some $n(\varepsilon) < \infty$ there exists a linear information operator $\hat{N}_{n(\varepsilon)}$ such that the partition $\Pi_{\hat{N}_n}(\mathscr{F})$ has a loss,

$$L^d_q(\Pi_{\hat{N}_n}) \leqslant \varepsilon.$$

It may seem at first that Definition 6 reduces the complexity or richness of the target class $\mathscr{F}$. This is not the case since $\mathscr{F}$ can still be rich with possibly an infinite pseudo-dimension. It is a partition $\Pi_{\hat{N}_n}$ with large enough, but finite, n which makes $\mathscr{F}$ be (d, ε)-approximable with a finite d.

The next theorem follows directly from the above definitions and from the fact that a class $\mathscr{H}^d$ of finite pseudo-dimension d is PAC-learnable.

Theorem 1 (PAC-learnability of $\mathscr{F}$). *If a target class $\mathscr{F}$ is (d, ε)-approximable then for any fixed integer $d > 0$, for all accuracy $\varepsilon > 0$, and for confidence parameter $0 < \delta < 1$, there exists a continuous information operator $\hat{N}_n$ such that for any target function $g \in \mathscr{F}$, based on partial information $\hat{N}_n(g)$ of finite cardinality $n(\varepsilon)$ and on a finite i.i.d. sample $\{(x_i, g(x_i))\}_{i=1}^{m(\varepsilon,\delta)}$, an algorithm can determine an hypothesis $\hat{h}$ in some class $\hat{\mathscr{H}}^d$ of pseudo-dimension d such that*

$$L_q(\hat{h}) \leqslant \varepsilon$$

with probability $1 - \delta$.

Remark. Compare the statement of the theorem to the classical PAC-model's guarantee of $L_q(\hat{h}) \leqslant L^*_{g,q} + \varepsilon$ (see Definition 1). In Theorem 1, hypothesis $\hat{h}$ has a loss which is arbitrarily close to zero.

Thus while the classical notion of PAC-learnability applies only to hypothesis classes $\mathscr{H}$ of finite pseudo-dimension, the new notion of PAC-learnability applies to any general target class $\mathscr{F}$ as long as $\mathscr{F}$ is (d, ε)-approximable. The elegance comes from the fact that the basis for the information-based simplification of $\mathscr{F}$ (into any subset $\mathscr{F}_y \in \Pi_N(\mathscr{F})$), which allows a class of pseudo-dimension d to approximate it arbitrarily well, is a *finite* amount of partial information represented by a form general enough to fit many types of learning settings. One can conceive other means of simplification, for instance, decomposing $\mathscr{F}$ into an infinite structure $\bigcup_{d \leqslant 1} \mathscr{H}^d$ of hypothesis classes of finite pseudo-dimensions but here the notion of partial information would have to be in terms of a teacher pointing to one such class. This is too strong of a constraint on the type of information that may be used. In contrast, by appealing to the theory of information-based complexity, one obtains a quantitative information representation which adheres to a wide array of learning settings.

Note that in all the above definitions the variable d appears as an arbitrary finite parameter. Depending on the particular learning setting, the pseudo-dimension of the hypothesis class is either fixed or allowed to vary as in the case of a structure $\{\mathscr{H}^d\}_{d=1}^{\infty}$. In either case only m and n need to depend on ε and δ while d is left as a variable whose value may be further optimized.

Definition 6 states the existence of an information operator $\hat{N}$ based on which $\mathscr{F}$ can be (d, ε)-approximable. There may be many such operators. Is there a notion of an optimal partial information operator? It may be defined as follows.

Definition 7 (Optimal partition-loss). For a fixed target class $\mathscr{F}$, fixed positive integers n and d, define the minimal partition-loss to be

$$I_{n, d, q}(\mathscr{F}) = \inf_{N_n} L_q^d(\Pi_{N_n}(\mathscr{F})),$$

where N_n runs over all linear information operators $N_n : \mathscr{F} \to \mathbb{R}^n$ of cardinality n.

We will refer to $I_{n, d, q}(\mathscr{F})$ as the *minimal information error* of the problem of PAC-learning $\mathscr{F}$.

3. THE MINIMAL INFORMATION ERROR

For a target class $\mathscr{F}$, the quantity $I_{n, d, q}(\mathscr{F})$ measures the minimal approximation error of the worst-case element in the target class given optimal partial information about it expressed, as n linear operations and given that the approximating class is of pseudo-dimension d.

The importance of $I_{n, d, q}$ arises from the following: First, it permits finding or estimating the most efficient form of partial information for learning, namely, that whose loss equals or almost equals the information error $I_{n, d, q}$. For this, one needs to obtain both upper and lower bounds on $I_{n, d, q}$. The tighter the estimation the better the information operator whose loss achieves the upper bound.

Second, $I_{n, d, q}$ permits a quantitative measure for the value of partial information for learning from examples. If we denote by N_n^* the operator whose associated partition has a loss $L_q^d(\Pi_{\substack{* \\ n}}(\mathscr{F})) = I_{n, d, q}$ then by Definition 5 for any $g \in \mathscr{F}$ there exists an hypothesis class $\mathscr{H}_y^{*d}$ and an hypothesis $h_y^* \in H_y^{*d}$, both depending on $y = N_n^*(g)$, such that $L_q(h_y^*) \leqslant I_{n, d, q}$. Let $\Delta \equiv \Delta(m, d, \delta) = c_1 \sqrt{(d \log^2 d \ln m + \ln(1/\delta))/m}$, the right side being an upper bound on the deviation between the empirical loss $\hat{L}_q(h)$ and the loss $L_q(h)$ uniformly over all $h \in \mathscr{H}_y^{*d}$, which holds for any fixed $q \geqslant 1$ and follows from a uniform SLLN result of Vapnik and Chervonenkis [32] (see Theorem 4 in [25]) using the fact that $\dim_p(\mathscr{H}_y^{*d}) = d < \infty$. Then using N^* and $\mathscr{H}_y^{*d}$ for the choice of $\hat{N}$ and $\hat{\mathscr{H}}^d$, respectively in Theorem 1 implies that for fixed n, $d \geqslant 1$, for any target $g \in \mathscr{F}$, an algorithm which minimizes the empirical loss obtains an hypothesis $\hat{h}$ with a loss,

$$\begin{aligned} L_q(\hat{h}) &\leqslant \hat{L}_q(\hat{h}) + \frac{\Delta}{2} \leqslant \hat{L}_q(h_y^*) + \frac{\Delta}{2} \leqslant L_q(h_y^*) + \Delta \\ &\leqslant I_{n, d, q} + \Delta(m, d, \delta). \end{aligned} \tag{1}$$

For any fixed d, fixing $m + n$ at some constant value and minimizing the upper bound with respect to m and n yields the optimal m^* and n^*. The value of partial information in terms of the number of training examples is reflected in the rate in which m^* grows with respect to n^*.

Third, $I_{n, d, q}$ must decrease with respect to d as for any $y \in \mathbb{R}^n$ the loss $L_q(\mathscr{H}^d) = \operatorname{dist}(\mathscr{F}_y, \mathscr{H}^d, L_q)$ decreases while $\Delta(m, d, \delta)$ increases with respect to increasing d. Hence, in case the learner has an available hypothesis class structure $\{\mathscr{H}_y^d\}_{y \in \mathbb{R}^n, d \in \mathbb{Z}_+}$ then there exists an optimal value d^* which minimizes the upper bound on $L_q(\hat{h})$. This in turns implies that, based on a finite amount of partial information and a finite sample size, in choosing the hypothesis $\hat{h}$, one should select a hypothesis class (or model) of optimal complexity d^*.

In the next section we apply the learning framework to the problem of learning a Sobolev target class of the form

$$W_p^{r, l} = \{f : \|D^k f\|_{L_p([0, 1]^l)} \leqslant 1, k : k_1 + \cdots k_l \leqslant r\}, \quad p \geqslant 1,$$

where $D^k f = (\partial^{k_1 + \cdots + k_l})/(\partial x_1^{k_1} \cdots \partial x_l^{k_l}) f$ and $\|f\|_{L_p[0, 1]^l} = (\int_{[0, 1]^l} |f|^p)^{1/p})$. This family $\{W_p^{r, l}\}_{p \geqslant 1}$ corresponds to rich infinite dimensional classes with smoothness parameter r. We estimate $I_{n, d, q}(W_p^{r, l})$ for general q satisfying $1 \leqslant q \leqslant p \leqslant \infty$.

4. PAC-LEARNING A SOBOLEV TARGET CLASS

The general Sobolev class $W_p^{r,l}$ defined in the previous section plays a principal role in the field of approximation theory. The degree (or rate) of approximation of $W_p^{r,l}$ by classes of finite pseudo-dimension has been recently studied in Maiorov and Ratsaby [19]; see also [20]. The degree of approximation of many other target classes can be derived from these results using the well-known embedding characteristics of Sobolev spaces (see, for instance, Adams [4]).

In this section we prove that $W_p^{r,l}$ which has an infinite pseudo-dimension, is PAC-learnable in the sense of Theorem 1. The proof is based on constructing a linear operator $\hat{N}_n$ and a family of hypothesis classes $\{\hat{\mathscr{H}}_y^d\}_{y \in \mathbb{R}^n}$ such that for any $\varepsilon > 0$, the class $W_p^{r,l}$ is $(d, \varepsilon/2)$-approximable based on a finite information cardinality $n(\varepsilon/2)$. This is equivalent to saying that the loss $L_q^d(\Pi_{\hat{N}_n}(W_p^{r,l})) \leqslant \varepsilon/2$, for any $\varepsilon > 0$ which in turn implies that for any target function $g \in W_p^{r,l}$, there is an hypothesis class $\hat{\mathscr{H}}_y^d$, depending on $y = \hat{N}_n(g)$, which contains an optimal hypothesis h_y^* with an optimal loss $L_{g,d,q}^* \leqslant \varepsilon/2$. As $\hat{\mathscr{H}}_y^d$ is of pseudo-dimension $d < \infty$, it is PAC-learnable; hence by (1), for arbitrary $\varepsilon > 0$ and $0 < \delta < 1$, $L_q(\hat{h}) - L_{g,d,q}^* \leqslant 2 \sup_{h \in \hat{\mathscr{H}}_y^d} |L_q(h) - \hat{L}_q(h)| = \Delta(m, d, \delta)$ and is bounded from above by $\varepsilon/2$, with probability $1 - \delta$, provided that the randomly drawn i.i.d. sample is of a large enough size $m(\varepsilon, \delta, d)$. It then follows that for any $g \in W_p^{r,l}$ there exists an hypothesis $\hat{h}$ satisfying $L_q(\hat{h}) \leqslant \varepsilon$ which may be determined, based on finite sample size m and information cardinality n, thereby proving the PAC-learnability of $W_p^{r,l}$.

The dependence of ε on m is already known from the SLLN upper bound. We now estimate its dependence on n by obtaining an upper bound on $I_{n,d,q}(W_p^{r,l})$. We also obtain a lower bound on $I_{n,d,q}(W_p^{r,l})$ which reflects on the goodness of the particular information operator used for the upper bound.

Theorem 2 (Minimal information error of $W_p^{r,l}$). *For any $1 \leqslant q \leqslant p \leqslant \infty$ and for $n \geqslant c_2 > 1$, $d \geqslant 1$, the minimal information error for a Sobolev target class $W_p^{r,l}$ is bounded as*

$$\frac{c_3}{(d + n \ln n)^{r/l}} \leqslant I_{n,d,q}(W_p^{r,l}) \leqslant \frac{c_4}{(n+d)^{r/l}}$$

for some constants $c_2, c_3, c_4 > 0$ independent of n and d.

Proof. From the previous series of definitions we have

$$I_{n,d,q}(W_p^{r,l}) = \inf_{N_n} \sup_{y \in \mathbb{R}^n} \inf_{\mathscr{H}^d} \sup_{f \in W_p^{r,l} \cap N_n^{-1}(y)} \inf_{h \in \mathscr{H}^d} \|f - h\|_{L_q}.$$

To obtain an upper bound, it suffices to consider a particular information operator $\hat{N}_n$ and a particular hypothesis class $\hat{\mathscr{H}}_y^d$ which may depend on y. The proof is the same as that of Lemma 6 in [25] which treats only the case of $W_\infty^{r,l}$. For any $y \in \mathbb{R}^n$, we define $\hat{\mathscr{H}}_y^d = \{\sum_{i=1}^n y_i \phi_i(x) + \sum_{i=n+1}^{n+d} a_i \phi_i(x) : a_i \in \mathbb{R}\}$, where $\{\phi_i\}_{i=1}^{n+d}$ are piecewise polynomials of degree at most $r-1$ over a uniform cubical partition of the domain $[0,1]^l$. We let $\hat{N}_n(f) = [b_1(f), ..., b_n(f)]$, where b_i, $1 \leqslant i \leqslant n$, are the coefficients of the projection of f onto the linear subspace spanned by the basis $\{\phi_i\}_{i=1}^n$. Thus, for any target $g \in W_p^{r,l}$, if we let $y = \hat{N}_n(g)$ then the optimal hypothesis $h_y^* \in \hat{\mathscr{H}}_y^d$ is $\sum_{i=1}^{n+d} b_i(g)\, \phi_i(x)$. Then using a result of Birman and Solomjak [7, Theorem 3.3], concerning spline approximation which states that $\sup_{f \in W_p^{r,l}} \|f - \sum_{i=1}^{n+d} b_i(g)\, \phi_i(x)\|_{L_q} \leqslant c_4/(n+d)^{r/l}$ yields the upper bound on $I_{n,d,q}(W_p^{r,l})$.

The proof of the lower bound follows next. Since $W_\infty^{r,l} \subset W_p^{r,l}$ for $p \geqslant 1$ and since $\|f\|_{L_1} \leqslant \|f\|_{L_q}$, $q \geqslant 1$, then $I_{n,d,q}(W_p^{r,l}) \geqslant I_{n,d,1}(W_\infty^{r,l})$. It then suffices to consider a subset of $W_\infty^{r,l}$ and compute the minimal information error for it. We construct this subset next. For $y \in \mathbb{R}$, let $\phi(y)$ be any function in $W_\infty^{r,1}$ which satisfies $|\phi(y)| \leqslant 1$, $\phi(y) = 0$ for $y \notin [0,1]$, $\phi(0) = \phi(1) = 0$, $\phi(\frac{1}{2}) = 1$. Let m be a fixed positive integer and $\tilde{m} = m^{1/l}$. Let $D = \{0, 1, ..., \tilde{m} - 1\}^l$. For $x \in \mathbb{R}^l$, $\bar{\imath} = [i_1, i_2, ..., i_l] \in D$ define the function $\phi_{\bar{\imath}}(x)$: $\prod_{j=1}^l \phi_{i_j}(x_j)$, where for $y \in \mathbb{R}$, $\phi_{i_j}(y) = \phi(\tilde{m}y - i_j)$, $0 \leqslant i_j \leqslant \tilde{m} - 1$, $1 \leqslant j \leqslant d$. Define the set of uniformly spaced m points $\{x_{\bar{\imath}}\}_{\bar{\imath} \in D}$ in $[0,1]^l$ as $x_{\bar{\imath}} = (1/\tilde{m})[i_1 + \frac{1}{2}, i_2 + \frac{1}{2}, ..., i_d + \frac{1}{2}]$.

Consider the function subclass

$$F_m = \left\{ f_a(x) = \frac{1}{m^{r/l}} \sum_{\bar{\imath} \in D} a_{\bar{\imath}} \phi_{\bar{\imath}}(x) : \|a\|_{l_\infty^m} \leqslant 1 \right\},$$

where $\|a\|_{l_\infty^m} = \max_{\bar{\imath} \in D} a_{\bar{\imath}}$. We now prove that $F_m \subset W_\infty^{r,l}$. For a multiinteger $\alpha \in \mathbb{Z}_+^l$, satisfying $|\alpha| = \sum_{i=1}^l \alpha_i \leqslant r$, denote by $f^{(\alpha)}$ the partial derivative of order α. Denote by $x_{\bar{\imath},j}$ the jth component of $x_{\bar{\imath}}$ and let $\Delta_{\bar{\imath}} = \{x \in [0,1]^l : x_{\bar{\imath},j} - \frac{1}{2} \leqslant x_j \leqslant x_{\bar{\imath},j} + \frac{1}{2},\ 1 \leqslant j \leqslant l\}$. We have

$$\begin{aligned}
&\sup_{x \in [0,1]^l} |f_a^{(\alpha)}(x)| \\
&\quad = \frac{1}{m^{r/l}} \sup_{x \in [0,1]^l} \left| \sum_{\bar{\imath} \in D} a_{\bar{\imath}} \phi_{\bar{\imath}}^{(\alpha)}(x) \right| \\
&\quad = \frac{1}{m^{r/l}} \max_{\bar{\jmath} \in D} \sup_{x \in \Delta_{\bar{\jmath}}} \left| \sum_{\bar{\imath} \in D} a_{\bar{\imath}} \phi_{\bar{\imath}}^{(\alpha)}(x) \right| \\
&\quad = \frac{1}{m^{r/l}} \max_{\bar{\jmath} \in D} \sup_{x \in \Delta_{\bar{\jmath}}} |a_{\bar{\jmath}} \phi_{\bar{\jmath}}^{(\alpha)}(x)| \\
&\quad = \frac{1}{m^{r/l}} \max_{\bar{\jmath} \in D} |a_{\bar{\jmath}}| \sup_{x \in \Delta_{\bar{\jmath}}} |\phi_{\bar{\jmath}}^{(\alpha)}(x)| \\
&\quad = \frac{1}{m^{r/l}} \max_{\bar{\jmath} \in D} |a_{\bar{\jmath}}| \sup_{x \in \Delta_{\bar{\jmath}}} |\phi^{(\alpha_1)}(\tilde{m}x_1 - j_1) \\
&\qquad \times \phi^{(\alpha_2)}(\tilde{m}x_2 - j_2) \cdots \phi^{(\alpha_l)}(\tilde{m}x_l - j_l)|
\end{aligned}$$

$$= \frac{1}{m^{r/l}} \max_{\bar{j}} |a_{\bar{j}}| \, \tilde{m}^r \sup_{x \in [0,1]^l} |\phi^{(\alpha)}(x)|$$

$$\leqslant \sup_{x \in [0,1]^l} |\phi^{(\alpha)}(x)| \leqslant 1.$$

The last line follows, since by assumption, $\phi \in W_\infty^{r,1}$. This proves that $f_a \in W_\infty^{r,l}$, for any $f_a \in F_m$.

We proceed with bounding $I_{n,d,1}(F_m)$ from below. First we state and prove the following lemma. We will henceforth denote by $E^m = \{-1, +1\}^m$.

Lemma 1. *Let $m \geqslant 1$, $n \geqslant 160$ be integers satisfying $m \geqslant 2100 n \ln n$. For any subspace $Q^n \subset \mathbb{R}^m$ of co-dimension n there exists a subset $V^n \subset E^m$, of cardinality $|V^n| \geqslant 2^{\gamma m}$ such that for every $v \in V^n$, $\operatorname{dist}(v, Q^n, l_1^m) \leqslant \alpha m$, where $\frac{1}{2} < \gamma < 1$, $0 < \alpha < 1$ are absolute constants.*

The proof of the lemma follows: The square of the l_2^m-distance between any vertex $v \in E^m$ and the subspace Q^n is the squared norm of the projection of v onto the n-dimensional subspace Q_n orthogonal to Q^n. The latter is the sum squared of the norm of the dot products of v with n orthogonal vectors u_i, $1 \leqslant i \leqslant n$, which span Q_n. We estimate the number of vertices in E^m whose distance from Q^n is no more than $\alpha\sqrt{m}$ in the l_2^m. Draw uniformly a vertex $v \in E^m$ by picking its ith component, $1 \leqslant i \leqslant m$, from $\{-1, +1\}$ with probability $\frac{1}{2}$. Then we obtain $\mathbf{P}(v \in E^m\colon \operatorname{dist}(v, Q^n, l_2^m) > \alpha\sqrt{m}) \leqslant \sum_{i=1}^n \mathbf{P}(|(v, u_i)|) > \alpha\sqrt{m}/\sqrt{n})$ which for $\alpha = 0.047$ is bounded from above by $2ne^{-m\alpha^2/4n}$ by using a standard application of Chebychev's inequality applied to a weighted sum of i.i.d. Bernoulli random variables. Using the inequality $\|a\|_{l_2^m} \leqslant \sqrt{m}\,\|a\|_{l_2^m}$ for any $a \in \mathbb{R}^m$, it then follows that, there are at least $2^m(1 - 2(1/n)^{2100\alpha^2/4 - 1)}$ vertices in E^m whose distance from Q^n in the l_1^m-metric is less than αm, given that $m \geqslant 2100 n \ln n$. For $n \geqslant 160$, $2^m(1 - 2(1/n)^{2100\alpha^2/4-1}) \geqslant 2^{\gamma m}$ for $\gamma = 0.986$, which proves the lemma. ▮

Consider the subspace $F_m \cap N_n^{-1}(0)$ of co-dimension n and denote by Q^n its corresponding subspace of co-dimension n in $\mathbb{R}^m$. Define the set of functions $F_m(V^n) = \{f_v \in F_m : v \in V^n\}$, where V^n is as defined in Lemma 1. We claim the following:

$$\sup_{f_v \in F_m(V^n)} \operatorname{dist}(f_v, \mathscr{H}^d, L_1)$$

$$\leqslant \sup_{f_a \in F_m \cap N_n^{-1}(0)} \operatorname{dist}(f_a, \mathscr{H}^d, L_1) + \frac{c}{m^{r/l}}$$

for an absolute constant $c > 0$. The proof follows next. Let $\hat{v} \in V^n$ be such that $\operatorname{dist}(f_{\hat{v}}, \mathscr{H}^d, L_1) = \sup_{f_v \in F_m(V^n)} \operatorname{dist}(f_v, \mathscr{H}^d, L_1)$. Let $B_\infty^m = \{a \in \mathbb{R}^m\colon \|a\|_{l_\infty^m} \leqslant 1\}$. Let $\hat{a} \in Q^n \cap B_\infty^m$ be such that $\operatorname{dist}(\hat{v}, \hat{a}, L_1) = \operatorname{dist}(\hat{v}, Q^n \cap B_\infty^m, L_1)$, namely, $\hat{a}$ is the closest vector in $Q^n \cap B_\infty^m$ to $\hat{v}$. We define the augmented approximating class $\mathscr{H}_0^d$ to be $\mathscr{H}^d \cup \{0\}$ which amounts to adding the zero element. It is straightforward to show that the pseudo-dimension of $\mathscr{H}_0^d$ is d. Then we have

$$\sup_{f_v \in F_m(V^n)} \operatorname{dist}(f_v, \mathscr{H}_0^d, L_1)$$

$$= \operatorname{dist}(f_{\hat{v}}, \mathscr{H}_0^d, L_1)$$

$$= \operatorname{dist}(f_{\hat{a}} + (f_{\hat{v}} - f_{\hat{a}}), \mathscr{H}_0^d, L_1)$$

$$\leqslant \operatorname{dist}(f_{\hat{a}}, \mathscr{H}_0^d) + \operatorname{dist}(f_{\hat{v}} - f_{\hat{a}}, \mathscr{H}_0^d, L_1)$$

$$\leqslant \operatorname{dist}(f_{\hat{a}}, \mathscr{H}_0^d) + \|f_{\hat{v}} - f_{\hat{a}}\|_{L_1}$$

$$\leqslant \sup_{f_a \in F_m \cap N_n^{-1}(0)} \operatorname{dist}(f_a, \mathscr{H}^d, L_1) + \|f_{\hat{v}} - f_{\hat{a}}\|_{L_1}$$

where for the second inequality we used $\operatorname{dist}(f_{\hat{v}} - f_{\hat{a}}, \mathscr{H}_0^d, L_1) \leqslant \operatorname{dist}(f_{\hat{v}} - f_{\hat{a}}, 0, L_1) = \|f_{\hat{v}} - f_{\hat{a}}\|_{L_1}$. We have

$$\|f_{\hat{v}} - f_{\hat{a}}\|_{L_1}$$

$$= \frac{1}{m^{r/l}} \int_{[0,1]^d} \left| \sum_{\bar{\imath} \in D} (\hat{a}_{\bar{\imath}} - \hat{v}_{\bar{\imath}})\, \phi_{\bar{\imath}}(x) \right| dx$$

$$= \left(\frac{1}{m^{r/l}} \int_{\varDelta} \left| \prod_{j=1}^d \phi(\tilde{m} x_j) \right| dx \right) \sum_{\bar{\imath} \in D} |\hat{a}_{\bar{\imath}} - \hat{v}_{\bar{\imath}}|$$

$$= \frac{1}{m^{r/l+1}} \left(\int_0^1 |\phi(y)|\, dy \right)^d \|\hat{a} - \hat{v}\|_{l_1^m}$$

$$= \frac{1}{m^{r/l+1}} \|\hat{a} - \hat{v}\|_{l_1^m}. \qquad (2)$$

Let a^* be the closest vector in Q^n to $\hat{v}$. Then by the triangle inequality we have $\|\hat{v} - \hat{a}\| \leqslant \|\hat{v} - a^*\| + \|a^* - \hat{a}\|$, the first term on the right being $\operatorname{dist}(\hat{v}, Q^n, l_1^m)$ which by Lemma 1 is bounded from above by αm, provided that we henceforth choose $m \geqslant 2100 n \ln n$. The second term on the right can be shown by a geometric argument to be bounded from above by $\alpha' m$ for some absolute constant $\alpha' > 0$. We may now bound (2) from above by $c_5/m^{r/l}$ for some absolute constant $c_5 > 0$, which proves the above claim, We may now continue bounding $I_{n,d,1}(F_m)$ from below as follows:

$$I_{n,d,1}(F_m) = \inf_{N_n} \sup_{y \in \mathbb{R}^n} \inf_{\mathscr{H}^d} \sup_{f \in F_m \cap N_n^{-1}(y)} \inf_{h \in \mathscr{H}^d} \|f - h\|_{L_1}$$

$$\geqslant \inf_{N_n} \inf_{\mathscr{H}_0^d} \sup_{f \in F_m \cap N_n^{-1}(0)} \inf_{h \in \mathscr{H}_0^d} \|f - h\|_{L_1}$$

$$\geqslant \inf_{N_n} \inf_{\mathscr{H}_0^d} \sup_{f_v \in F_m(V^n)} \operatorname{dist}(f_v, \mathscr{H}_0^d, L_1) - \frac{c_5}{m^{r/l}}. \qquad (3)$$

We now make use of a proof of Theorem 2 in [20] which establishes that for any subset $K \subset E^m$ of cardinality $|K| \geqslant 2^{c_6 m}$ for some absolute constant $c_6 > 0$, there exists a subset $G \subset K$ such that for every $u, v \in G$, $u \neq v$, we have

$\|u-v\|_{l_1^m} > 2\beta m$, β a positive absolute constant, and $|G| = 2^{c_7 m}$, where $c_7 > 0$ a constant which depends only on γ and β. We apply this result by substituting the set V^n for K and denoting the resulting $2\beta m$-separated subset as G^n. Thus (3) is now bounded from below by

$$\inf_{N_n} \inf_{\mathscr{H}_0^d} \sup_{f_v \in F_m(G^n)} \operatorname{dist}(f_v, \mathscr{H}_0^d, L_1) - \frac{c_5}{m^{r/l}}. \tag{4}$$

In the proof of Theorem 1 in [19], it is shown that for a class $F_m(G^n)$, for any $\mathscr{H}^d$, $\sup_{f \in F_m(g^n)} \operatorname{dist}(f, \mathscr{H}^d, L_1) \geqslant c_8/m^{r/l}$, provided $m = c_9 d$ for some absolute constant $c_9 > 0$. We may therefore lower bound (4) by $c_{10}/m^{r/l}$ for some absolute constant $c_{10} > 0$. Satisfying the above two constraints on m we substitute for $m = c_{11}(n \ln n + d)$ which then proves the theorem. ▌

Let us now apply Theorem 1 for PAC-learning the class $W_p^{r,l}$. As mentioned in Section 2, based on $I_{n,d,q}$, one may determine an upper bound on the loss of the hypothesis $\hat{h}$ obtained by PAC-learning the target class. Doing that yields

$$\begin{aligned} L_q(\hat{h}) &\leqslant I_{n,d,q} + \Delta(m, d, \delta) \\ &\leqslant \frac{c_4}{(n+d)^{r/l}} + c_1 \sqrt{\frac{d \log^2 d \ln m + \ln(1/\delta)}{m}} \end{aligned}$$

for some constants c_1, $c_4 > 0$ not depending on n, d, and m. It is clear from this expression that $W_p^{r,l}$ is PAC-learnable; i.e., for any $\varepsilon > 0$, $1 \leqslant d < \infty$, one can find finite m and n such that the loss $L(\hat{h}) \leqslant \varepsilon$. Interestingly enough for a fixed sample size m and fixed information cardinality n there is an optimal complexity

$$d^* \leqslant c_{12} \left(\left\{ \frac{rm}{l\sqrt{\ln m}} \right\}^{2l/(l+2r)} - n \right), \tag{5}$$

which minimizes the upper bound on the loss. Thus if a structure of hypothesis classes $\{\mathscr{H}^d\}_{d=1}^{\infty}$ is available to the learner then the best choice for a hypothesis class on which the learner should run empirical loss minimization is $\mathscr{H}^{d^*}$.

Let us compute how n and m trade-off. Fixing d and $m+n$ at some constant values and then minimizing the upper bound on $L_q(\hat{h})$ over m and n yields m^* and n^*. When $l < 2r$ we find that m^* grows polynomially in n^* at a rate no larger than $n^{*(1+r/l)}$; i.e. roughly speaking, partial information about the target g is worth 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.

The lower and upper bounds on $I_{n,d,q}(W_p^{r,l})$ stated in Theorem 2 are tight up to a logarithm factor in n; hence it follows that partial information based on the particular information operator $\hat{N}_n$ and family of hypothesis classes $\{\hat{\mathscr{H}}_h^d\}_{y \in \mathbb{R}^n}$ almost achieves the loss of the optimal information operator. Hence, for the classes $W_p^{r,l}$, $p \geqslant 1$, information based on a linear projection onto a linear hypothesis class $\hat{\mathscr{H}}_y^d$, as the one used in the proof of the upper bound on $I_{n,d,q}(W_p^{r,l})$, is close to being optimal.

5. CONCLUSION

We introduced a theoretical framework which extends the PAC model of learning to a scenario where a learner has general partial information about the target function, in addition to randomly drawn labeled examples. The framework extends PAC learnability to rich target classes with possibly infinite pseudo-dimension where now, in addition to a finite random sample, a finite amount of information is needed. For a family of Sobolev classes $W_p^{r,l}$, $p \geqslant 1$, it is found that the value of partial information, as compared to the sample size, depends on the ration of the smoothness parameter r and the dimensionality of the domain l. Moreover, information based on a linear projection operator onto a linear hypothesis class in determined to be almost optimal as it achieves (up to a logarithm factor in n) the lower bound on the error of the best linear information operator for $W_p^{r,l}$.

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. Y. S. Abu-Mostafa, Learning from hints in neural networks, *J. Complexity* **6** (1990), 192–198.
2. Y. S. Abu-Mostafa, Hints and the VC dimension, *Neural Comput.* **5** (1993), 278–288.
3. Y. S. Abu-Mostafa, Machines that learn from hints, *Sci. Am.* **272**, No. 4 (1995).
4. R. A. Adams, "Sobolev Spaces," Academic Press, New York, 1975.
5. N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, Scale-sensitive dimensions, uniform convergence, and learnability, *in* "Proceedings of the 34rd Annual Symposium on Foundations of Computer Science," pp. 292–301, IEEE Comput. Soc. Press, Los Alamitos, CA, 1993.
6. A. Barron, Approximation and estimation bounds for artificial neural networks, *Mach. Learning* **14** (1994), 115–133.
7. M. S. Birman and M. Z. Solomjak, Piecewise-polynomial approximations of functions of the classes W_p^α, *Math. USSR-Sb.* **2**, No. 3 (1967), 295–317.
8. A. Blumer, E. Ehrenfeucht, D. Haussler, and M. Warmuth, Learnability and the Vapnik–Chervonenkis dimension, *J. Am. Comput. Mach.* **36**, No. 4 (1989), 929–965.

9. L. Devroye, L. Gyorfi, and G. Lugosi, "A Probabilistic Theory of Pattern Recognition," Springer-Verlag, New York/Berlin, 1996.
10. R. O. Duda and P. E. Hart, "Pattern Classification and Scene Analysis," Wiley, New York, 1973.
11. K. Fukunaga, "Introduction to Statistical Pattern Recognition," Academic Press, New York, 1972.
12. U. Grenander, "Abstract Inference," Wiley, New York, 1981.
13. D. Haussler, Decision theoretic generalizations of the PAC model for neural net and other learning applications, *Inform. Comput.* **100**, No. 1 (1992), 78–150.
14. D. Haussler, Sphere packing numbers for subsets of the Boolean *n*-cube with bounded Vapnik–Chervonenkis dimension, *J. Combin. Theory Ser. A* **69** (1995), 217–232.
15. J. A. Jacobs, Methods for combining experts' probability assessments, *Neural Comput.* **7** (1995), 867–888.
16. G. G. Lorentz, M. v. Golitschek, and Y. Makovoz, "Constructive Approximation, Advanced Problems," Springer-Verlag, New York/Berlin, 1996.
17. G. Lugosi and A. Nobel, Adaptive model selection using empirical complexities, submitted.
18. G. Lugosi and K. Zeger, Concept learning using complexity regularization, *IEEE Trans. Inform. Theory* **42**, No. 1 (1995).
19. V. Maiorov and J. Ratsaby, Nonlinear function approximation using function classes of finite pseudo-dimension, *J. Constr. Approx.*, to appear.
20. V. Maiorov and J. Ratsaby, The degree of approximation of sets in Euclidean space using sets with bounded Vapnik–Chervonenkis dimension, *J. Discrete Appl. Math.* **86** (1998), 81–93.
21. A. Pinkus, "*n*-Widths in Approximation Theory," Springer-Verlag, New York, 1985.
22. D. Pollard, "Convergence of Stochastic Processes," Series in Statistics, Springer-Verlag, New York/Berlin, 1984.
23. D. Pollard, "Empirical Processes, Theory and Applications," NSF-CBMS Regional Conf. Ser., SIAM, Philadelphia, 1989.
24. J. Ratsaby, R. Meir, and V. Maiorov, Towards robust model selection using estimation and approximation error bounds, *in* "Proc. 9th Annual Conference on Computational Learning Theory," pp. 57, ACM, New York, 1996.
25. J. Ratsaby and V. Maiorov, Generalization of the PAC-model for learning with partial information (extended abstract), *in* "Proceedings of the Third European Conference on Computational Learning Theory, EuroCOLT 97" Springer-Verlag, New York/Berlin, 1997.
26. B. D. Ripley, "Pattern Recognition and Neural Networks," Cambridge Univ. Press, Cambridge, 1996.
27. M. Roscheisen, R. Hofmann, and V. Trespt, Incorporating prior knowledge into networks of locally-tuned units, *in* "Computational Learning Theory and Natural Learning Systems" (S. Hanson, T. Petsche, M. Kearns, and R. Rivest, Eds.), MIT Press, Cambridge, MA, 1994.
28. J. Shawe-Taylor, P. Bartlett, R. Williamson, and M. Anthony, *in* "A Framework for Structural Risk Minimisation," NeuroCOLT Technical Report, Series NC-TR-96-032, Royal Holloway, University of London, 1996.
29. G. G. Towell and J. W. Shavlik, Interpretation of artificial neural networks: Mapping knowledge-based neural networks into rules, *in* "Advances in Neural Information Processing Systems 4," pp. 977–984, Morgan Kaufmann, Denver, 1991.
30. J. F. Traub, G. W. Wasilkowski, and H. Wozniakowski, "Information-Based Complexity," Academic Press, San Diego, 1988.
31. L. G. Valiant, A theory of the learnable, *Comm. ACM* **27**, No. 11 (1984), 1134–1142.
32. V. N. Vapnik, "Estimation of Dependences Based on Empirical Data," Springer-Verlag, Berlin, 1982.
33. H. White, "Estimation, Inference and Specification Analysis," Cambridge Univ. Press, Cambridge, 1994.