Noise Tolerant Learnability via the Dual Learning Problem

Shai Fine Ran Gilad-Bachrach Shahar Mendelson
Naftali Tishby

Institute of Computer Science
The Hebrew University
Jerusalem 91904, Israel
Email: {fshai,ranb,marsmen,tishby}@cs.huji.ac.il

Abstract. Much of the research in machine learning and neural computation assumes that the teacher gives the correct answers to the learning algorithm. However, in many cases this assumption is doubtful, since different sorts of noise may disable the learner from getting the correct answers all the time. The noise could be caused by noisy communication, human errors, measuring equipment and many other causes. In some cases, problems which are efficiently learnable without noise, become hard to learn when noise is introduced [3]. Partial answers to the noisy learning problem are known for specific classes, ([11]), but in general, no simple parameters are known which distinguish between classes that are learnable in the presence of noise and those which become hard to learn. The goal of this work is to introduce such parameters.
We use the *Membership Query* model to show that if the VC-dimension of the *dual* learning problem is moderate and if the dual learning problem is *dense* in some sense, then the class is learnable in the presence of *persistent random classification noise*. We show that our assumptions hold for a variety of learning problems, such as smooth functions, general geometrical concepts, and monotone monomials. We are particularly interested in the analysis of smooth function classes. We show uniform upper bound on the *fat-shattering dimension* of both the primal and dual learning problem which is derived by a geometric property of the class called *type*.

1 Introduction

In the PAC learning model it is assumed the learner has access to an *oracle* (a teacher) which randomly picks an instance from the sample space and returns this instance and its correct label according to the rule (concept) one wishes to learn. In real life problems, the existence of such oracle is doubtful: human mistakes, communication errors and various other problems prevent us from generating such an oracle. In these cases, a realistic assumption would be that the learner has access to some sort of a *noisy oracle*. This oracle may have internal mistakes which prevent it from generating the correct labels constantly,

or even influence its ability to randomly select points from the sample space. A variety of noise models appear in literature, ranging from random classification noise (in which only the labels of the instances are affected in such a way that the probability of receiving a correct answer from the oracle is better then random guessing [2]), to the malicious noise model, in which the oracle reacts to the learner's actions by changing attributes and labels in an adversary manner [12].

While the *Vapnik-Chervonenecis dimension* ([19],[4]) completely characterizes PAC learnability of classes of $\{0,1\}$-functions, the case of learning in a noisy environment is unresolved. Nevertheless, in a variety of cases one can show that even in the presence of noise learning is possible. Kearns [11] presented the model of *Statistical Queries* (SQ) and proved that if a class is learnable with SQs then it is also learnable in the presence of random labeling noise and presented a variety of problems in which this approach may be applied. Since Kearns' pioneering paper, much work was done on the subject of learning in a noisy environment [6, 10], however, little is known regarding the properties which make a concept class robust to noise. In this study we try to characterize those conditions.

We suggest two requirements: learnability and a density feature both of which, with respect to the dual learning problem. These requirements enable the learner to find many instances to which the target concept assigns the same value as to the given instance. We show that if the two assumptions hold, then it is possible to learn in a noisy environment. To this end, we introduce a general scheme which assumes the presence of a *Membership Query* (MQ) oracle.

A Membership Query oracle is a mechanism which enables the learner to sample the values of the target functions on specific instances which he chooses. The existence of such a mechanism provides substantial power to the learner, hence, this assumption may be regarded as questionable. However, in some cases, the MQ oracle can be efficiently simulated to meet our needs [8].

Throughout this paper we denote by Ω the sample space and by $\mathcal{H}$ the hypotheses class. We endow $\mathcal{H}$ with a known probability measure p, and assume that the target concept $c_t \in \mathcal{H}$ was selected according to p (*Bayesian Learning*). Set ϵ and δ to be the accuracy and confidence parameters respectively. Let d be the VC dimension of the primal learning problem and put d^* the VC dimension of the dual learning problem. Denote by η is the noise rate[1] , hence $0 \leq \eta < \frac{1}{2}$.

Our main result connects the ability to learn in the presence of persistent noise[2] with the complexity of the dual learning problem and its density features (cf. section 3).

Theorem 1. *For any efficiently learnable problem, if the dual problem is efficiently learnable and dense then there exists a learning algorithm L which uses the MQ oracle, such that for every $\epsilon, \delta > 0$, and a noise corrupted sample, L learns the primal problem using*

$$Poly(n, \frac{1}{\epsilon}, \frac{1}{\delta}, \frac{1}{1-2\eta}, d) \tag{1}$$

[1] We shall often use the term *learning parameters* to address all the parameters above.

[2] once an instance's label is set, repeated sampling of the same instance does not change it.

samples, and in time

$$Poly(n, \frac{1}{\epsilon}, \frac{1}{\delta}, \frac{1}{1-2\eta}, d, d^*) \tag{2}$$

where n is the size of an instance (input dimension).

The proof we give is constructive: we present an algorithm which satisfies (1) and (2). Note that if d^* is bounded by some polynomial in the other learning parameters, this algorithm learns efficiently.

The rest of the paper is organized as follows: In section 2 we define the dual learning problem and discuss its implications briefly. Section 3 is devoted to the notion of a dense learning problem. In section 4 we present the construction of our scheme – which provides noise robustness to any learning problem which meets the learnability and density criteria of the dual learning problem (A few simple applications for binary values functions are presented in appendix A). In section 5 we generalize our discussion to real valued functions. We focus our attention to classes of continuous functions, discuss the concept of *type* and its applications to learning theory, and demonstrate the applicability of the suggested paradigm to classes of sufficiently smooth functions. We conclude with closing remarks and several leads for further study (section 6).

2 The Dual Learning Problem

A learning problem may be characterized by a tuple $\langle \Omega, \mathcal{H} \rangle$ where Ω is the sample space and $\mathcal{H}$ is the hypotheses class. Learning can be viewed as a two player game: one of them, the teacher, picks a target concept, while his counterpart, the learner, tries to identify this concept. Different learning models differ in the way the learner interacts with the teacher (PAC's Sampling Oracle, Statistical Queries, Membership Queries, Equivalence Queries, etc.) and in the method in which the learning performance are evaluated.

Every learning problem has a dual learning problem [17] which may be characterized by the tuple $\langle \mathcal{H}, \Omega \rangle$. In this representation the learning game reverses: first the teacher chooses an instance $\omega_0 \in \Omega$ and then the learner tries to approximate this instance by querying on the value ω_o assigns to different functions.

Note that each instance ω can by viewed as a function δ_ω on $\mathcal{H}$ such that $\delta_\omega(h) = h(\omega)$. We denote this collection of functions by Ω^*. Given a prior probability measure on the hypotheses space, then

$$\rho_\Omega \left(\delta_w, \delta_{\widehat{w}}\right) = \Pr\left(\{h \in \mathcal{H} : h\left(w\right) \neq h\left(\widehat{w}\right)\}\right) \tag{3}$$

is a pseudo-metric on Ω. Thus, the distance between two functions $\delta_w, \delta_{\widehat{w}} \in \Omega^*$ is the probability that an hypothesis h, drawn from $\mathcal{H}$ according to p, does not assign the same label to both w and $\widehat{w}$.

To clarify this notion we present two dual learning problems:

- Let Ω be the interval $[0,1]$ and set $\mathcal{H}$ to be the class of all intervals $[0,a]$ where $a \in [0,1]$. If ω is an instance and h_a is the hypothesis $[0,a]$ then

$h_a(\omega) = 1$ if and only if $a \geq \omega$. Turning to the dual learning problem, note that $\delta_\omega(h_a) = 1$ if and only if $\omega \leq a$. Hence, the dual learning problem is equivalent to learning intervals of the form $[\omega, 1]$ where $\omega \in [0, 1]$.

- Let $\Omega = \mathbb{R}^n$, and put $\mathcal{H}$ to be the class of linear separators (Perceptrons), i.e., $h_x \in \mathcal{H}$ is the hypothesis which assign to each $\omega \in \mathbb{R}^n$ the label $\text{sign}\langle x, w\rangle$. Therefore, the dual learning problem is the problem of learning linear separators, implying that this problem is dual to itself.

The VC-dimension of the dual learning problems (also called *co-VC*) obeys the inequalities [15]:

$$\log(d) \leq d^* \leq 2^d$$

where d is the VC-dimension of the primal problem and d^* is the co-VC. This gives in a worst case scenario an exponential gap in the complexity of the two problems. However, in both examples presented here and in most realistic cases this gap is only polynomial. This observation makes Theorem 1 interesting for many noise corrupted learning problems (cf. Troyansky's thesis [18] for a comprehensive survey of dual representation). In section 5 we broaden the discussion to handle concepts which are not $\{0, 1\}$ -functions. Thus, we replace the notion of VC-dimension with *fat-shattering* dimension. We show that for classes which consist of sufficiently smooth functions, both the fat-shattering and the co-fat-shattering have an upper bound which is polynomial in the learning parameters and enables efficient learning of the dual problem.

3 Dense Learning Problems

A learning problem is dense [7] if every hypothesis has many hypotheses which are close but not equal to it.

Definition 1. *A learning problem $\langle \Omega, \mathcal{H}\rangle$ is* dense *with respect to the probability measure $\mathcal{D}$ if for every $h \in \mathcal{H}$ and every $\epsilon > 0$ there exists $h' \in \mathcal{H}$ such that $0 < Pr_{x\in\mathcal{D}}[h(x) \neq h'(x)] < \epsilon$.*

The density property is distribution dependent: for every learning problem there exists a distribution in which this learning problem is *not* dense. Indeed, if the distribution is singular, i.e., supported on a finite set , the problem cannot be dense. If a learning problem is indeed dense, any hypothesis can be approximated by infinite number of hypotheses [3].

We would like to extend the definition of a dense learning problem to finite hypothesis classes as well. We replace the demand of approximating h for every ϵ, by the same demand for every polynomial ϵ. In the above definition it is required that each hypothesis h can be approximated by infinite number of hypotheses. In the finite case we replace the infinity assumption by a super-polynomial number of approximating hypotheses.

[3] An alternative definition would be that for any h there exists infinite number of h' such that $\Pr_{x\in\mathcal{D}}[h(x) \neq h'(x)] < \epsilon$. Note that the two definitions are not identical.

Definition 2. *Let l be the index of a finite class $\mathcal{H}_l$. A sequence of finite learning problems $\langle \Omega_l, \mathcal{H}_l \rangle$, is* dense *with respect to the probability measure $\mathcal{D}_l$ if for every polynomial $p(l)$ and every $h \in \mathcal{H}_l$ there exists a set $H_l \subset \mathcal{H}_l$ of size super polynomial in l such that for every $h' \in H_l$, $0 < Pr_{\omega \in \mathcal{D}_l}[h(\omega) \neq h'(\omega)] < 1/p(l)$.*

In both definitions the governing idea is that each hypothesis can be approximated using large (infinite or super-polynomial) number of hypotheses.

4 Applying Noise Robustness to a Learning Problem

We now present our conversion scheme. We apply noise robustness to any learning problem which satisfies the requirements of learnability and density w.r.t its dual learning problem. Our scheme is based on the following simple argument.

Our object is to "clean" a noise corrupted sample, i.e., to restore the correct label of every instance in the sample. Since the noise is persistent, constant sampling by MQ wont do the trick. However, if there are enough instances in the sample space to which the target concept assigns the same label, it is enough to sample such instances and use majority vote. The problem is to identify those instances. To this end, we use the dual learning problem. The requirements of learnability and density ensure that with high probability, for any instance ω, the dual learning algorithm will find many instances ω' such that almost all concepts assign them with the same label. Since the learner of the primal learning problem knows the dual target ω and the probability measure on $\mathcal{H}$ (the Bayesian assumption), he can provide a clean sample to the dual learning problem. Hence, the dual learning problem is noise free and therefore far more easy.

The detailed algorithm is as follows:

1. Pick a random sample $\mathcal{S} = \{\omega_1, \ldots, \omega_m\}$, where $m = \text{Poly}(n, \frac{1}{\epsilon}, \frac{1}{\delta}, d)$.
2. For each instance $\omega_i \in \mathcal{S}$,
   - By simulating the dual learning problem $k = \text{Poly}\left(\frac{1}{1-2\eta}, \log(\frac{m}{\delta})\right)$ times, generate an ensemble $\mathcal{S}_i = \left\{\omega_i^1, \ldots, \omega_i^k\right\}$ such that every ω_i^j is δ/m close to ω_i in the dual representation.[4]
   - Use MQ to get a label for every $\omega_i^j \in \mathcal{S}_i$.
   - Assign a label to x_i based on majority vote over the labels of $\mathcal{S}_i$
3. Learn the primal learning problem using the newly assigned labels of $\mathcal{S}$.

As stated it Theorem 1, the sample complexity of the algorithm is $m \times k$ which is polynomial in the learning parameters. The time complexity depends polynomially on the same parameters as well as the VC-dimension of the dual problem. The proof that the algorithm is correct is straightforward, and follows using the union bound and the Chernoff bound.

[4] The density requirement guarantees that it is possible to find many such different approximating instances.

As shown in appendix A, it is possible to use this paradigm for a large variety of problems. Geometric oriented concepts such as *neural networks*, and *general geometric concepts*, boolean problems such as *monotone monomials* and many other problems, all have a dense and moderate dual problem. Therefore, this paradigm may be used to solve such problems. Note that the class of *parity functions* does not satisfy our assumptions. Indeed, although the learning problem is dual to itself, the density property is violated, since every two different instances disagree on exactly half of the hypotheses.

5 Dealing with classes of continuous functions

Next, we present an example in which the concepts are continuous functions: For every given ω_0 we attempt to find "many" instances ω, such that for most of the concepts f, $f(\omega)$ is "almost" $f(\omega_0)$. When the concepts are continuous functions, it is natural to search for the desired ω near ω_0. However, if there is no a–priori bound on the modulus of continuity of the concepts, it is not obvious when ω is "close enough" to ω_0. Moreover, in certain examples the desired instances are not necessarily close to ω_0, but may be found "far away" from it (cf. App. A.3).

In general, the question of learnability of the dual problem may be divided into two parts. The first is to construct a learning algorithm L which assigns to each sample $S_n = \{f_1, ..., f_n\}$ a point ω such that for every f_i, $|f_i(\omega) - f_i(\omega_0)| < \varepsilon$. The second part is to show that the class of functions $\Omega^* = \{\delta_\omega | \omega \in \Omega\}$ on $\mathcal{H}$ satisfies some compactness condition (e.g., a finite fat-shattering dimensions $VC_\varepsilon(\mathcal{H}, \Omega^*)$).

5.1 Estimating $VC_\varepsilon(\mathcal{H}, \Omega^*)$

As a starting point, recall the following standard definitions and notation: given a normed space $(X, \| \|)$, let X^* be the space of the continuous linear functionals on X. The norm on X^* is given by $\|x^*\| = \sup_{\|x\|=1} |x^*(x)|$. Each Banach space X is isometrically embedded in the space of continuous functionals on X^*, denoted by X^{**}, via the duality map $x \to x^{**}$ given by $x^{**}(x^*) = x^*(x)$. Put $B(X)$ the unit ball of X and $B(X^*)$ the dual unit ball.

Definition 3. *Let $(X, \| \|)$ be a Banach space consisting of functions on $\Omega \subset \mathbb{R}^d$. We say that X has a reproducing kernel if for every $\omega \in \Omega$, the evaluation functional δ_ω is norm continuous.*

Note that if the class $\mathcal{H}$ is a subset of such a space X, then the question of dual learning is always a linear learning problem in the following sense: the "base" space is now $\mathcal{H}$, and we wish to learn the continuous linear functional δ_ω. In other words, $\Omega^* \subset X^*$.

We will show that if $\mathcal{H}$ is a subset of a Banach space with a reproducing kernel and if Ω^* is a bounded subset of X^*, then the fat-shattering dimension $VC_\varepsilon(\mathcal{H}, \Omega^*)$ is finite for every $\varepsilon > 0$ – provided that X satisfies a geometric property called *type*.

Recall that a Rademacher random variable on $[0,1]$ is of the form $r(t) = sign(sin(2^n \pi t))$ for some n. Hence, it is a symmetric $\{-1,1\}$–valued random variable.

Definition 4. *A Banach space X has type p, if there is some constant C such that for every $x_1, ..., x_n \in X$,*

$$\int_0^1 \left\| \sum_1^n r_i(t) x_i \right\| dt \leq C \left(\|x_i\|^p \right)^{1/p} \tag{4}$$

where $r_i(t)$ are i.i.d. Rademacher random variables on $[0,1]$.

The basic facts concerning the concept of type may be found, for example, in [13] or in [16]. Clearly, for every Banach space (4) holds in the case $p = 1$. If we set

$$T(X) = \sup\{p | X \text{ has type } p\} \tag{5}$$

then it follows that $T(X) \in [1,2]$. For example, Hilbert spaces and L_p spaces for $p \geq 2$ have type 2, thus, if X is a Hilbert space or if $X = L_p$ for $p \geq 2$ then $T(X) = 2$ and the supermum is attained.

The case which interests us the most, is that of Banach spaces with a non–trivial type (i.e. $T(X) > 1$) in which the evaluation functionals are uniformly bounded. The classical example for such spaces is that of the Sobolev spaces $W^{k,p}$, (cf. [9] for the basic facts regarding Sobolev spaces, or [1] for a comprehensive survey). An example of such a space is $W^{1,2}(0,1)$, which is the space of continuous functions $f : [0,1] \to \mathbb{R}$, for which the derivative f' belongs to L_2 with respect to the Lebesgue measure. The inner product on this space is given by $\langle f, g \rangle = \int_0^1 (fg + f'g')dx$.

The concept of *type* may be regarded as a learning theory parameter, because of the striking fact that $T(X) = 1$ if and only if for every integer n, the n–dimensional simplex can be linearly embedded in $B(X)$ by an operator T_n, such that $\|T_n\| \left\|T_n^{-1}\right\| \leq C$, where C is a universal constant (see [16]). The ability to embed a simplex $\{0, Tx_1, ..., Tx_n\}$ such that $\|Tx_j\| = \varepsilon$, is measured by $VC_\varepsilon(B(X), B(X^*))$. Indeed, given any classification of the vertices of the embedded simplex, there is a linear functional which separates the vertices such that the size of the margin is ε. The norm of this functional in X^* is determined by the ability to embed the simplex in $B(X)$ (see [14]).

It turn out that $VC_\varepsilon(B(X), B(X*))$ is determined by the type of X:

Theorem 2. *([14]) For every infinite dimensional Banach space X the fat-shattering dimension $VC_\varepsilon(B(X), B(X^*))$ is finite if and only if $T(X) > 1$. If $T(X) = p$ and $p' < p$ then there are constants C and c such that*

$$c \left(\frac{1}{\varepsilon}\right)^{\frac{p}{p-1}} \leq VC_\varepsilon\left(B(X), B(X^*)\right) \leq C \left(\frac{1}{\varepsilon}\right)^{\frac{p'}{p'-1}}$$

If X has type $p = T(X)$ then $VC_\varepsilon\left(B(X), B(X^*)\right) = \Omega\left(\frac{1}{\varepsilon}\right)^{\frac{p}{p-1}}$

Corollary 1. *Assume that X is a Banach space which has a non trivial type and that $\mathcal{H}$ is a bounded subset of X. Assume further that the evaluation functionals are uniformly bounded in X^*. Then $VC_\varepsilon(\mathcal{H}, \Omega^*) < \infty$ for every $\varepsilon > 0$.*

Proof: Clearly, we may assume that $\mathcal{H}$ is a subset of the unit ball of X, and that Ω^* is a subset of the dual unit Ball. Hence, by our assumptions, for every $\varepsilon > 0$

$$VC_\varepsilon\left(\mathcal{H}, \Omega^*\right) \leq VC_\varepsilon\left(B(X), B(X^*)\right) < \infty \tag{6}$$

□

From Theorem 2 it also follows that in certain cases the fat-shattering dimension of the primal learning problem is also polynomial in the learning parameters. Indeed, note that X has a non-trivial type if and only if X^* has a non–trivial type (see [16]). Also, $(\omega_1, ..., \omega_n)$ is ε–shattered by $f \in \mathcal{H}$ if and only if $(\delta_{\omega_1}, ..., \delta_{\omega_n})$ is ε–shattered by $f^{**} \in \mathcal{H}^{**}$, where f^{**} is the image of f using the duality map. Thus, $f^{**}(\delta_\omega) = f(\omega)$. Hence, if $\mathcal{H}$ is bounded in X, Ω^* is bounded in X^* and X^* has type p^*, then

$$VC_\varepsilon\left(\Omega, \mathcal{H}\right) = O\left(VC_\varepsilon\left(B(X^*), B(X^{**})\right)\right) = O\left(\frac{1}{\varepsilon}\right)^{\frac{p^*}{p^*-1}} \tag{7}$$

Also, it was shown in [14] that the primal learning problem is efficiently learnable for classes of functions which are bounded subsets of Hilbert spaces with reproducing kernels in which the evaluation functionals are uniformly bounded.

5.2 Approximating ω_0 on a sample of functions

Next, we turn to the problem of approximating ω_0 by some $\omega \neq \omega_0$ on any given sample $\mathcal{S}_k = \{f_1, ..., f_k\}$. Note that when $\mathcal{H}$ consists of continuous functions on Ω, then for every $\varepsilon > 0$ and every sample $f_1, ...f_k$, there are infinitely many instances ω, such that $|f_i(\omega) - f_i(\omega_0)| < \varepsilon$ for $1 \leq i \leq k$. Indeed, since $\{f_1, ..., f_k\}$ is a finite set of continuous functions, they have a uniform upper bound for their modulus of continuity. Thus, there is some $r > 0$, such that if $\omega \in B_{\omega_0}(r)$ and $1 \leq i \leq k$, then $|f_i(\omega) - f_i(\omega_0)| < \varepsilon$.

Corollary 1 *Let X be a Banach space consisting of continuous functions on Ω which has a non–trivial type. Then, the dual learning problem has infinitely many solutions. In particular, $\langle \mathcal{H}, \Omega^* \rangle$ is dense.*

Proof: The proof follows because the class of functions Ω^* on the space $\mathcal{H}$ is a uniform Glivenko–Cantelli class and since we are able to approximate δ_{ω_0} on any given sample of functions. □

6 Concluding Remarks and Further Research

The main idea which motivated this study was to formulate a set of conditions in which noise tolerant learnability can be achieved. We have proposed two requirements: learnability and density of the dual learning problem. In a nutshell, these requirements suggest that the key to overcome noise lies in the ability to exploit useful information from the underline structure of the concept class and to approximate the target function. Adding an access to a MQ oracle, we were able to device a conversion scheme which applies noise robustness to any learning problem which meets these demands.

In this presentation we handled only random classification noise. However, it seems that our method works in many other noise models (e.g. malicious noise).

This study suggests an hierarchy of the difficulty in noise tolerance learnability via the dual learning problem: Some problems are learnable in the presence of noise (monomials), while others are learnable only under the *Bayes* assumption (e.g. continuous functions). Note that some learning problems which are believed to be hard to handle in the presence of noise, indeed do not meet our set of demands (e.g. parity functions).

In section 5 we have generalized our scheme to handle real valued functions. We were able to show that when the space has a non–trivial type, the fat-shattering dimension of both the primal and the dual learning problem are polynomial in the accuracy parameter ε. One open problem we are interested in, is to generalize the techniques suggested in section 5 to handle $\{0,1\}$-functions as well. Note that when the concepts are not continuous functions, then the modulus of continuity will not be a useful parameter. We were able to obtain partial results (measure dependent) provided that for every function the decision boundary (the points in which the function is not continuous) is small in some sense.

Finally, our construction provides a set of sufficient conditions for noise tolerance learnability. However, we believe that the essence of these conditions reflects fundamental principles which may turn out to be a step towards a complete characterization of noise tolerant learnability.

References

1. R.A. Adams. *Sobolev Spaces*, volume 69 of *Pure and Applied Mathematics series.* Academic Press, 1975.
2. D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.
3. P. Bartlett and S. Ben-David. Hardness results for neural network approximation problems. In *In the proceedings of the 4'th European Conference on Computational Learning Theory*, 1999.
4. A. Blumer, A. Ehrenfeucht, D. Haussler, and Warmuth. M. K. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.
5. N. H. Bshouty, S. A. Goldman, H. D. Mathias, S. Suri, and H. Tamaki. Noise-tolerant distribution-free learning of general geomtric concepts. In *In the proceedings of the 28th Annual ACM Symposium of Theory of Computing*, 1996.

6. S. E. Decator. *Efficient Learning from Faulty Data.* PhD thesis, Harvard University, 1995.
7. B. Eisenberg and R. L. Rivest. On the sample complexity of pac-learning using random and chosen examples. In *In the Proceedings of the Third Annual Conference on Computational Learning Theory*, pages 154–162. Morgan-Kaufmann, 1990.
8. S. Fine, R. Gilad-Bachrach, E. Shamir, and N. Tishby. Noise tolerant learning using early predictors. subbmitted to NIPS 1999.
9. D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order.* Springer Verlag, 1998.
10. E. Jackson, E. Shamir, and C. Shwartzman. Learning with queries corrupted by classification noise. *Discrete Applied Mathematics*, 1999. to appear.
11. M. J. Kearns. Efficient noise-tolerant learning from statistical queries. *In the proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computation*, pages 392–401, 1993.
12. M. J. Kearns and M. Li. Learning in the presence of malicious errors. *SIAM J. Comput*, 22, N0. 4:807–837, 1993.
13. J. Lindenstrauss and L. Tzafriri. *Classical Banach Spaces*, volume 2. Springer Verlag, 1979.
14. S. Mendelson. Learnability in banach spaces with reproducing kernels. Preprint, 1999.
15. J. Pach and P. K Agarwal. *Combinatorial Geometry.* John Wiley & Sons, Inc., 1995.
16. G. Pisier. Probabilistic methods in the geometry of banach spaces. In *Probability and Analysis*, number 1206 in Lecture Notes in Mathematics, pages 167–241. Springer Verlag, 1986.
17. L. Pitt and M. K. Warmuth. Prediction-preseving reducibility. *Journal of Computer and System Sciences*, 41:430–467, 1990.
18. L. Troyansky. *Faithful Representations and Moments of Satisfaction: Probabilistic Methods in Learning and Logic.* PhD thesis, Hebrew University, Jerusalem, 1998.
19. V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its applications*, XVI(2):264–280, 1971.

A Some Examples of Use

Most of the paper is devoted to the introduction of the paradigm of learning with noise and the case of smooth functions. In this appendix we present a few other examples.

A.1 Monotone Monomials

The first problem we would like to present is a Boolean learning problem of discrete nature – the problem of learning *monotone monomials*. In this problem the sample space is $\{0,1\}^n$ and the concepts are conjunctions (e.g. $x_1 \wedge x_2 \wedge x_5$). The hypothesis class is $\mathcal{H} = \{c_I : I \subset \{0,1\}^n\}$ were $c_I(x) = \bigwedge_{i \in I} x_i$. To simplify notation we will assume that $x \subseteq [1..n]$, and $c \subseteq [1..n]$ such that $c(x) = 1 \iff c \subseteq x$. The dual learning problem for this class is learning monotone monomials with reverse order, i.e., $x(c) = 1 \iff x \supseteq c$. Both the primal and the dual learning problems have the same VC-dimension, n.

Instead of showing that the dual class is dense, we will give a direct argument showing that the label of each instance can be approximated. Let $Y_x = \{y \subseteq [1..n] : x \subseteq y\}$. Since the concept class is monotone, then if $c(x) = 1$ it follows that for every $y \in Y_x$, $c(y) = 1$. On the other hand, if $c(x) = 0$ then there exists some $i \in c \setminus x$. Half of the instances $y \in Y_x$ have $i \notin y$, implying that $c(y) = 0$ for each such y. Thus, $\Pr_{y \in Y_x}[c(y) = 0] \geq 1/2$ with respect to the uniform distribution on Y_x.

Hence, if $c(x) = 1$ then $\Pr_{y \in Y_x}[c(y) = 0] = 0$ whereas if $c(x) = 0$ then $\Pr_{y \in Y_x}[c(y) = 0] \geq 1/2$. This allows us to distinguish between the two cases. In order to be able to do the same thing in the presence of noise we have to demand that Y_x is big enough. From the definition of Y_x it follows that $||Y_x|| = 2^{n-||x||}$. It will suffice to require that $||x|| \leq np$ for $p < 1$ with high probability, since in this case $||Y_x||$ is exponentially big. This condition holds for uniform distribution and many other distributions.

Note that in this case there is no need for a Bayesian assumption, i.e., we do not assume the existence of a distribution on the sample space. However, we have used a slightly relaxed definition of density in which for most of the instances there exists a sufficient number of approximating instances.

A.2 Geometric Concepts

In contrast to the previous example, when dealing with geometric concepts, the sample space and the concept class are both infinite. For the sake of simplicity Let the sample space be $\mathbb{R}^2$ and the assume that the concept class consists of axis aligned rectangles. In this case, the VC-dimension of the primal problem is 4 and dual problem has VC-dimension 3. Moreover if a "smooth" probability measure is defined on the concept class, it is easily seen that each instance ω_0 is approximated by all the instances with distance r from it (when r depends on the defined measure and the learning parameters). Therefore, this class is dense.

This example can be expanded to large variety of problems, such as neural networks, general geometric concepts [5] and high dimensional rectangles. We would like to describe two methods of doing so.

Using the Bayesian Assumption: Each geometric concept divides the instance space into two sets. The edge of these sets is the *decision boundary* of the concept. Assume that for every instance there is a ball around it which does not intersect the decision boundary of "most" of the concepts. Denote by Pr a probability measure on $\mathcal{H}$, and assume for every $\delta > 0$ there exists $r > 0$ ($r = r(\delta, \omega_0)$) such that

$$\Pr_{h \in \mathcal{H}}[B(\omega_0, r) \cap \partial h \neq \phi] < \delta \tag{8}$$

($B(\omega_0, r)$ is the ball of radius r centered at ω_0 and ∂h is the decision boundary of h). If (8) holds then all the points in the $B(\omega_0, r)$ can be used to predict the label of ω_0, and therefore, for verifying the correct label of ω_0.

Geometric Concepts without Bayesian Assumption: A slightly different approach can be used when there is no Bayesian assumption. Given $\delta > 0$, then for every concept h there exists a distance $r^h > 0$ such that the measure of all points with distance smaller than r^h from the decision boundary of h does not exceed δ. If $0 < r = \inf_{h \in H} r^h$, then with high probability (on the instance space) a random point ω_0 has a distance which is greater than r to the decision boundary of the pre–selected target concept h. Hence, the ball of radius r around ω_0 can be used to select approximating instances.

A.3 Periodic Functions

In this example we present a case where the approximating instances are not in the near neighborhood of ω_0. Let $\Omega = \mathbb{R}$ and set $\mathcal{H} = \{\sin(2\pi t/p) | p \text{ is a prime } \}$. Since the number of primes is countable, the probability measure on $\mathcal{F}$ is induced via a measure on $\mathbb{N} : \nu(\{n\}) = 1/n(n+1)$. Note that the $\mathcal{F}$ consists of periodic functions, but for each function, the period is different. Given a point $\omega_0 \in \mathbb{R}$, and a confidence parameter δ, there is a finite set of functions A, such that $\nu(A) \geq 1 - \delta$. Hence, the elements of A have a common period. Therefore, there is some x, such that for every $f \in A$ and every $m \in \mathbb{N}$, $f(\omega_0) = f(\omega_0 + mx)$. It is reasonable to assume that the noise in the primal learning problem is not periodic (because the elements of the class do not have a common period), therefore, it is possible to find many points which agree with ω_0 with high probability, but are far away from the metric point of view. Moreover, using the same idea, given any sample $f_1, ..., f_k \in \mathcal{F}$, it is possible to construct an infinite number of points ω_i which agree with ω_0 on the given sample. It can be shown the learning rule L which assigns to each sample $s_k = \{f_1, ..., f_k\}$ one of the points ω_i, generalizes well in the sense that

$$\lim_{k \to \infty} \Pr\{S_k | sup_{\omega_o \in \mathbb{R}} \mathbb{E}_\nu (\delta_{\omega_0} - \delta_{L(S_k)})^2 \geq \epsilon\} = 0 \tag{9}$$