A Universal Noise Cleaning Procedure via the Dual Learning Problem

S. Fine
IBM Research Center
Yorktown Heights
NY 10598, USA
fshai@us.ibm.com

R. Gilad-Bachrach S. Mendelson N. Tishby
Institute of Computer Science
The Hebrew University
Jerusalem 91904, Israel
{ranb,marsmen,tishby}@cs.huji.ac.il

Abstract

We present a new scheme for cleaning a sample corrupted by labeling noise. Our scheme is universal in the sense that we only make general assumptions on the *dual* learning problem and therefore it is completely detached from the specifics of the *primal* problem itself. In a nutshell, we turn to the dual learning problem to exploit valuable information about the underlying structure of the primal one, which in turn provides the means to device a simple "noise cleaning" mechanism, using *Membership Queries*. We demonstrate the strength and applicability of the suggested method with a few learning problems of different nature. Of particular interest is the problem of learning in the restricted class of parity functions, where only k out of n bits are active. We show that in the MQ model we can outperforme the recent result by Blum et al. [3] and handle $k = O\left(n - c\log\left(n\right)\log\log\left(n\right)\right)$. This also provides a sharp separation between our method and the SQ model. The suggested procedure works not only for classification problems but for regression problems as well. To this end, we present a uniform upper bound on the *fat-shattering* dimension of both the primal and dual problems, which is derived from a geometric property of classes of real-valued functions - called *type*.

1 Introduction

In the PAC learning model it is assumed that the learner has access to an *oracle* which randomly picks an instance from the sample space and returns it with its correct label. In most real-life problems the existence of such an oracle is doubtful due to human errors, measurement noise, faulty communication, etc. A more realistic approach will be to assume that information provided by the oracle could be distorted in some way. Several noise models have been considered in the literature. These models range from random classification noise, where the labels can be randomly flipped with certain fixed probability [2], to a malicious noise model, in which the oracle may change both the sampling distribution and classifications in an adverse manner [6]. Choosing an hypothesis with minimal training error provides

good approximation to the target concept [2] but can be NP-hard. In contrast, a learning algorithm in the *Statistical Query* (SQ) Model [7] uses statistics, as opposed to labeled instances, in the course of learning. SQ algorithms can in turn be converted to noise tolerant PAC algorithms. However, it was recently shown that there are classes which are not SQ learnable but they are still PAC learnable in the presence of labeling noise [3].

Our main result forms a link between the ability to learn in the presence of persistent noise[1] and the complexity (both sample and computational) and density (cf. section 3) of the dual learning problem: We show that these two properties provide the means to apply noise robustness to the primal problem. The dual learning problem is constructed by switching the roles of the teacher and the learner, i.e. a "target" instance is being learned based on a sample of values it is assigned by different hypotheses. The density feature guarantees that any target instance can be approximated by many other instances, to which most of the hypotheses assign the same value. This is, intuitively, a continuity assumption in both instances and hypotheses. The learnability of the dual problem enables us to find such approximating instances. Assuming the presence of a *Membership Query* (MQ) oracle[2], it is easy to construct a noise cleaning procedure by simply performing a majority vote (cf. section 4).

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 player, 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 and in the method by which the learning performance is evaluated.

Every learning problem has a dual learning problem [12] which may be characterized by the tuple $\langle \mathcal{H}, \Omega \rangle$. In this representation the learning game is reversed: first the teacher chooses an instance $\omega_0 \in \Omega$ and then the learner tries to approximate this instance by querying on the values 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 p on the hypotheses space, then

$$\rho_\Omega\left(\delta_w, \delta_{\widehat{w}}\right) = \Pr\left(\{h \in \mathcal{H} : h\left(w\right) \neq h\left(\widehat{w}\right)\}\right) \tag{1}$$

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 $a \in [0,1]$. The dual problem is learning intervals of the form $[\omega, 1]$ where $\omega \in [0,1]$.
- Let $\Omega = \mathbb{R}^n$ and set $\mathcal{H}$ to be the class of Perceptrons, i.e. $h_x(\omega) = sign\langle x, \omega \rangle$, where $x, \omega \in \mathbb{R}^n$. The symmetric role of the parameters in the $sign$ function implies that this problem is dual to itself.

[1] Repeated sampling of the same instance does not change its label.

[2] The MQ oracle [1] is a mechanism which enables the learner to obtain the value of the target function on specific instances. The existence of such a tool provides the learner with substantial power, though in some important cases it is inevitable [8].

The VC-dimension of the dual learning problems (also called the co-VC) obeys the inequalities [10]: $\log(d) \leq d^* \leq 2^d$, where d is the VC-dimension of the primal problem and d^* is the co-VC. Although this inequality guarantees only exponential gap, in many learning problems the gap is smaller (as demonstrated in the previous examples). In section 6 we broaden the discussion to handle continuous functions and obviously we replace VC with *fat-shattering* dimension. We show that for classes of sufficiently smooth functions, both the fat-shattering and the co-fat-shattering have a polynomial upper bound (in the learning parameters) which enables efficient learning of the dual problem.

3 Dense Learning Problems

A learning problem is dense if every hypothesis has many hypotheses which are close but not equal to it [4]:

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 supported on a finite set, the problem cannot be dense. In order to deal with finite hypothesis classes, we modified the above definition in two aspects: we restrict ϵ to be polynomial; and we require that any hypothesis is approximated by super-polynomial number of hypotheses.

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 $h' \in H_l$, $\Pr_{\omega \in \mathcal{D}_l}[h(\omega) \neq h'(\omega)] < 1/p(l)$.*

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

4 Cleaning a Noisy Sample

We present a procedure for cleaning a random labeling noise in a sample of 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:

Continuous re-sampling a given instance by the MQ oracle will not restore its true label since the noise is persistent. However, if the target concept assigns to many other instances the same label it assigned to the given one, then its possible to query the labels of sufficient amount of them and use a majority vote. It remains to show how do we 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 the concepts assign ω and ω' with the same label. Since the learner of the primal problem knows the dual target ω and the probability measure on $\mathcal{H}$ (a Bayesian assumption), he can provide a clean sample to the dual learning algorithm. Hence, the dual learning problem is noise free and therefore easier to solve.

Theorem 1 *For any efficiently PAC learnable problem, if the dual problem is efficiently PAC learnable and dense with respect to a known probability measure,*

then there exists a learning algorithm L which uses a MQ oracle such that for any $0 < \epsilon, \delta < 1$ and a sample corrupted by random classification noise with rate $\eta < \frac{1}{2}$, L learns the primal problem using:

$$Poly(n, \frac{1}{\epsilon}, \frac{1}{\delta}, \frac{1}{1-2\eta}, d) \text{ examples, and } Poly(n, \frac{1}{\epsilon}, \frac{1}{\delta}, \frac{1}{1-2\eta}, d, d^*) \text{ time.}$$

where n is the size of an instance (input dimension) and d, d^* are the VC dimensions of the primal and the dual learning problems, respectively.

We provide a constructive proof by presenting an algorithm which satisfies the computational and sample complexity requirements. Note that if d^* is bounded by some polynomial in the other learning parameters, this algorithm learns efficiently. The detailed algorithm follows:

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

The sample complexity of the algorithm is $m \times k$, i.e. polynomial in the learning parameters. The computational complexity depends polynomially on the same parameters as well as on the VC-dimension of the dual problem. The proof of correctness follows immediately using standard techniques and hence will be omitted.

5 Examples of learning Binary-Valued functions

Consider the class of *monotone monomials*, in which concepts are conjunctions over $\{0,1\}^n$ formed solely by positive literals (e.g. $x_1 \wedge x_2 \wedge x_5$), thus $\mathcal{H} = \{c_I : I \subset \{0,1\}^n\}$ where $c_I(x) = \bigwedge_{i \in I} x_i$.

Instead of showing that the dual class is dense, we will give a direct argument showing that the label of (almost) every instance can be approximated: Let $Y_x = \{y \in \{0,1\}^n | \forall i (x_i = 1) \Rightarrow (y_i = 1)\}$. Since the concept class is monotone, then $c_I(x) = 1$ implies $c_I(y) = 1$ for every $y \in Y_x$ On the other hand, if $c_I(x) = 0$ then there exists some $i \in I$ such that $x_i = 0$. Half of the instances $y \in Y_x$ have $y_i = 0$, implying that $c_I(y) = 0$ for each such y. Thus, $\Pr_{y \in Y_x}[c_I(y) = 0] \geq 1/2$ with respect to the uniform distribution on Y_x. Hence, if $c_I(x) = 1$ then $\Pr_{y \in Y_x}[c_I(y) = 0] = 0$, whereas if $c_I(x) = 0$ then $\Pr_{y \in Y_x}[c_I(y) = 0] \geq 1/2$. 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 need Y_x to be large 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 large. This condition holds for the uniform distribution, as well as for many other distributions. Note that in this case there is no need for a Bayesian assumption. However, we have used a slightly

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

relaxed definition of density in which for most of the instances (instead of all of them) there exists a sufficient number of approximating instances.

The class of parity functions is dual to itself (and hence $d^* = n$), but it is clearly not dense. Blum et al. [3] showed that if we restrict the parity functions to depend only on the first $k = c\log(n)\log\log(n)$ bits then the class is learnable in the presence of noise but not with SQ. This restricted class is indeed dense since any point has 2^{n-k} approximating instances which are easy to find (they have the same first k bits). The only restriction on k is that 2^{n-k} must be super-polynomial, hence k can be as big as $n - c\log(n)\log\log(n)$. This separates our model from the SQ model.

In contrast to the previous examples, when dealing with geometric concepts, the sample space and the concept class are both infinite. Consider the class of axis-aligned rectangles in $\mathbb{R}^2$. In this case, the VC-dimension of the primal problem is 4 and the co-VC is 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 simple example can easily be extended to handle general geometric concepts and neural networks.

6 Dealing with classes of continuous functions

Next, we generalize our discussion to the case of learning 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. Section 6.2).

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 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 dimension $VC_\varepsilon(\mathcal{H}, \Omega^*)$).

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

Given a Banach 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 $x^{**}(x^*) = x^*(x)$. Let B_X denote the unit ball of X and B_{X^*} denote the dual unit ball.

We will focus on Banach spaces of functions on Ω for which the point evaluation functionals are uniformly bounded, i.e. there is some m such that $\sup_\omega \|\delta_\omega\| \leq m$. This implies that Ω^* is a subset of mB_{X^*}. It is possible to show that if $\mathcal{H}$ is a bounded subset of such a Banach space X, then $VC_\varepsilon(\mathcal{H}, \Omega^*)$ is finite for every $\varepsilon > 0$, provided that X satisfies a geometric property called *type*[4] (cf. [11]).

Definition 3 *A Banach space X has type p, if there is some constant C such that*

[4]For further details regarding the connection of *type* and learning, see [9].

for every $x_1, ..., x_n \in X$,

$$\int_0^1 \left\| \sum_1^n r_i(t) x_i \right\| dt \leq C \left(\|x_i\|^p \right)^{1/p} \tag{2}$$

where $r_i(t)$ *are i.i.d. Rademacher random variables*[5] *on* $[0,1]$. *Denote* $T(X) = \sup\{p | X \text{ has type } p\}$

Theorem 2 *([9]) For every infinite dimensional Banach space* X *the fat-shattering dimension* $VC_\varepsilon(B_X, B_{X^*})$ *is finite iff* $T(X) > 1$. *If* $T(X) = p$ *and if* X *has type* p', *then there are constants* C *and* c *(which depend on* X*) such that*

$$(c/\varepsilon)^{\frac{p}{p-1}} \leq VC_\varepsilon(B_X, B_{X^*}) \leq (C/\varepsilon)^{\frac{p'}{p'-1}}$$

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

Note that $\mathcal{H} \subset m_1 B_X$ and $\Omega^* \subset m_2 B_{X^*}$ implies $VC_\varepsilon(\mathcal{H}, \Omega^*) \leq VC_{\frac{\varepsilon}{m_1 m_2}}(B_X, B_{X^*})$.

Corollary 3 *Assume that* X *is a Banach space of functions on* Ω *which has type* $p > 1$ *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* $\varepsilon > 0$.

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 type $p > 1$ type iff X^* has a type $p^* > 1$ (see [11]). Also, $(\omega_1, ..., \omega_n)$ is ε–shattered by $f \in \mathcal{H}$ iff $(\delta_{\omega_1}, ..., \delta_{\omega_n})$ 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

$$VC_\varepsilon(\Omega, \mathcal{H}) = O\left(VC_\varepsilon(B(X^*), B(X^{**}))\right) = O(1/\varepsilon)^{\frac{p^*}{p^*-1}} \tag{3}$$

Moreover, it was shown in [9] that the primal problem is efficiently learnable for classes of functions which are bounded subsets of Hilbert spaces in which the evaluation functionals are uniformly bounded.

6.2 Approximating ω_0 on a sample of functions

When $\mathcal{H}$ consists of continuous functions on Ω, then for every $\varepsilon > 0$ and every sample $\mathcal{S}_k = \{f_1, ..., f_k\}$, there are infinitely many instances ω, such that $|f_i(\omega) - f_i(\omega_0)| < \varepsilon$ for $1 \leq i \leq k$. 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 4 *Let* X *be a Banach space consisting of continuous functions on* Ω *which has a non–trivial type. Then, the dual learning problem has an infinite number of solutions. In particular,* $\langle \mathcal{H}, \Omega^* \rangle$ *is dense.*

The approximating points need not be geometrically close, as demonstrated by the next example: 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, a probability measure on $\mathcal{H}$ is induced via a measure on $\mathbb{N}$: $\nu(\{n\}) = 1/n(n+1)$. Note that $\mathcal{H}$ 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 $m \in \mathbb{N}$, $f(\omega_0) = f(\omega_0 + mx)$.

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

7 Concluding Remarks and Further Research

We presented a universal pre-processing technique for cleaning a noisy sample. This result implies that any hypotheses class for which its dual problem is learnable and dense, can be learned in a noisy environment by first applying the cleaning procedure and then learning using the noise free sample. This result holds as long as the MQ oracle is accessible and the noise model is a *persistence random classification noise*. However, we already know that there are some cases in which MQ can be efficiently simulated [5], and we conjecture that the suggested procedure will also be useful for dealing with other models of classification noise.

A problem which is still open is to devise similar techniques to those used in section 6, to handle $\{0,1\}$-functions. Note that when the concepts are not continuous functions, the modulus of continuity is not a useful parameter. We were able to obtain partial results (measure dependent) provided that for every function the decision boundary is small in some sense.

Finally, this study suggests a new characterization of the difficulty associated with learning in a noisy environment based on general properties possessed by the dual learning problem. This may 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. Pursuing this direction is left for further study.

Acknowledgments

We would like to thank Ronitt Rubinfeld for many insightful comments.

References

[1] D. Angluin. Queries and concept learning. *Machine Learning*, 2:319–342, 1988.

[2] D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2:343–370, 1988.

[3] A. Blum, A. Kalai, and H. Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. In *Proc. of the 22'nd Ann. ACM Symp. on Theory of Computing*, 2000.

[4] B. Eisenberg and R. L. Rivest. On the sample complexity of pac-learning using random and chosen examples. In *Proc. of the 3'rd Ann. Conference on Computational Learning Theory*, pages 154–162, 1990.

[5] S. Fine, R. Gilad-Bachrach, E. Shamir, and N. Tishby. Noise tolerant learning using early predictors. Preprint, 1999.

[6] M. Kearns and M. Li. Learning in the presence of malicious errors. In *Proc. of the 20'th Ann. ACM Symp. on Theory of Computing*, pages 267–280, May 1988.

[7] M. J. Kearns. Efficient noise-tolerant learning from statistical queries. In *Proc. of the 25'th Ann. ACM Symp. on Theory of Computation*, pages 392–401, 1993.

[8] E. Kushilevitz and Y. Mansour. Learning decision trees using the fourier spectrum. *SIAM Journal on Computing*, 22(6):1331–1348, 1993.

[9] S. Mendelson. Learnability in banach spaces with reproducing kernels. Preprint, 1999.

[10] J. Pach and P. K Agarwal. *Combinatorial Geometry*. John Wiley & Sons, Inc., 1995.

[11] 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.

[12] L. Pitt and M. K. Warmuth. Prediction-preseving reducibility. *Journal of Computer and System Sciences*, 41:430–467, 1990.