NOISE TOLERANT LEARNING USING EARLY PREDICTORS

Shai Fine Ran Gilad-Bachrach Eli Shamir Naftali Tishby
Institute of Computer Science
The Hebrew University
Jerusalem 91904, Israel
Email: {fshai,ranb,shamir,tishby}@cs.huji.ac.il

Abstract

Generalization in most PAC learning analysis starts around $O(d)$ examples, where $d = VC_{dim}$ of the class. Nevertheless, analysis of learning curves using statistical mechanics shows much earlier generalization [7]. Here we introduce a gadget called *Early Predictor*, which exists if somewhat better than random prediction of the label of an arbitrary instance can be obtained from labels of $O(\log d)$ random examples. We were able to show that by taking a majority vote over a committee of Early Predictors, strong and efficient learning is obtained. Moreover, this learning procedure is robust to persistent classification noise. The margin analysis of the vote is used to explain this result. We also compare the suggested method to *Bagging* [11] and *Boosting* [5] and connect it to the *SQ* model [10]. A concrete example of Early Predictor is constructed for learning linear separators under uniform distribution. In this context we should mention the hardness result by Bartlett and

1 Introduction

Traditionally, analysis of the learning curve starts from $O(d)$ examples, where $d = VC_d$. For example, Helmbold and Warmuth [8] showed that $2d - \Omega\left(\sqrt{d \log d}\right)$ examples suffices for weak learning[1], while $d - O(\sqrt{d})$ are essential for distribution free learning. Haussler et. al. [7] showed, based on analysis taken from the statistical mechanics of learning, that for many classes of distributions generalization is possible even with very few examples. In this paper we continue this line of study and focus our attention on the very beginning of the learning curve. In particular we claim that if we can identify any non trivial behavior of the learning curve in the first $O(\log d)$ examples, then this information may be exploited and plugged into a general efficient learning scheme which is tolerant to random persistent classification noise.

The noise model which we refer to is the persistent random classification noise. In this model the learner has access to an oracle $\mathcal{O}$ which returns a pair (x, y) such that x is drawn independently from the instance space according to a fix distribution and y is the correct label of x with probability $1 - \eta$. Persistence means that once a label of an instance have been determined by the environment (the oracle $\mathcal{O}$, sampling through Membership Query, etc.) it will never be changed by the environment, regardless of the fact that it was correct or incorrect label. This extension to the PAC learning model was first examined by Angluin and Laird [1]. A most significant step towards PAC learning in this noise model have been accomplished by Kearns who presented the *Statistical Query* (SQ) model [10] in which the learning algorithm is allowed to ask for an estimate of the expected values of functions defined over the labelled examples and use these values to learn. This model may be viewed as a restriction on the way an algorithm uses the PAC example oracle. Kearns showed that approximating these values can be done by sampling efficiently as long as it is allowed to give not too small additive mistake. Since this simulation averages on many samples, it is less vulnerable to persistent random classification noise. Therefor, if a class is learnable via *statistical queries*, it is also learnable in the presence of noise. It turns out that many concept classes known to be efficiently learnable in the PAC model are also learnable in the SQ model and thus are noise tolerant. (cf. Decatur's thesis [3] for further reading).

One deficiency of the SQ model that we would like to address is of practical nature: The construction and usage of the statistical queries are problem dependant and hence there is no general scheme to covert a PAC learning algorithm to

[1] not necessarily polynomial weak learning

a SQ learning algorithm. Sometimes the conversion is quite intricate (cf. Jackson, Shamir and Shwartzman [9]). We therefore suggest a different technique to overcome the noise - voting of a committee.

Advanced voting algorithm such as Freund and Shapire's AdaBoost [5], are vulnerable to noise since they tend to over-fit the corrupted examples [4]: AdaBoost generate distributions which focus on mislabeled instances and by doing so generate hypotheses which are heavily influenced from these instances; More moderated algorithms, such as Bagging [11] do not change distributions but even so might be heavily affected by noise since the basic learning algorithm might not be able to find an hypothesis consistent with the sample it gets, if some of the instances are mislabeled. Gofer and Shamir [6] showed that it can be overcome even for malicious noise, but they used a sample exponentially big in the VC dimension.

1.1 Summary of Results

The problem of learning from noisy sample become apparent even for the task of weak learning. Most algorithms need at least $O(d)$ examples in order to produce a weak hypothesis, but in the presence of noise the probability of getting uncorrupted sample is exponentially small. On the other hand a noisy sample, i.e. a sample which contains one or more mislabeled instances, might not allow the algorithm to learn at all. In order to overcome this problem we would like to gain information from as little as $O(\log d)$ examples, in this case we have significant probability of getting a sample without any mislabeled instances.

In this paper we show that if, for a fixed and known set of distributions, one is capable of a non trivial prediction from as few as $O(\log d)$ samples, it is possible to construct an efficient, noise-tolerant learning algorithm based on these early predictors using a simple voting scheme. We use the margin theorem of Schapire et al. [12] to prove the efficiency of this learning technique. Specifically for the class of linear separators, we show that under the assumption of *uniform distribution*, this learning technique is applicable. In contrast, it follows from Bartlett and Be-David [2] that learning a linear separator in the presence of noise is hard when the distribution is not known.

We would like to formulate the kind of information needed from a very small sample in order to allow efficient and robust learning. We define an *Early Predictor* as a function that upon receiving a (very small) random sample of labeled instances (examples) returns a prediction for the label of a given (unlabeled) instance. The accuracy of the prediction should be only slightly better then a random guess. In parallel with the *Weak Learning* paradigm, we assume the existence of a polynomial q in the learning parameters such that $\frac{1}{q}$ quantifies the advantage of the prediction over random guessing. More precisely, let $\mathcal{C}$ be a concept class defined over instance space X and $d = VC_d(\mathcal{C})$, let $c_t \in \mathcal{C}$ be the target concept and let $S^{(k)} = \{(x_1, y_1), \ldots, (x_k, y_k)\}$ be a sample of size k, where $x_i \in X$ is an instance and $y_i \in Y$ is the corresponding label.

Definition 1 *An* Early Predictor *is a function* $\hat{P}:(X \times Y)^k \times X \to Y$, *where* $k = O(\log d)$ *such that there exists a polynomial* $q(1/\delta, n, size(c_t), d) > 0$ *that for every* $\delta > 0$ *the following inequality holds*

$$Pr_x[Pr_{S^{(k)}}[\hat{P}(S^{(k)}, x) = c(x)] - \max_{y \neq c(x)} Pr_{S^{(k)}}[\hat{P}(S^{(k)}, x) = y] \leq \alpha_\delta] < \delta \tag{1}$$

where $\alpha_\delta = \frac{1}{q}$ *(thus emphasizing the dependence of the advantage on the confidence parameter* δ*).*

For almost every instance, we require that $\hat{P}$ will give a correct label *on average* over a given sample. Note the difference between early predictor and *weak prediction*[2]. *Weak prediction* guarantees that a correct label may be obtained for more then half of the sample space, while *early predictors* guarantees that for almost all the sample space it will give a correct label with probability more then half - the probability is on the predictors rather then on the instance space. Our first result states that if such a predictor exists then the class $\mathcal{C}$ is learnable: Using early predictors we can generate an hypothesis by forming a committee of predictors and voting on their predictions.

Definition 2 *Let* Q *be a set of samples such that for each* $i \in \{1 \ldots N\}$, Q_i *is a sample of size* k *(a total of* Nk *examples). Assuming binary concepts, i.e.* $Y = \{-1, +1\}$, *we define the* majority function

$$f_Q(x) = \frac{1}{n} \sum_i \hat{P}_{Q_i}(x)$$

and the corresponding majority hypothesis

$$h_Q(x) = sign(f_Q(x))$$

[2] *weak prediction* is equivalent to *weak learning* [8]

Hence, the hypothesis is the sign of the majority vote of a committee of early predictors $\hat{P}(Q_1, x), \ldots, \hat{P}(Q_N, x)$ while the absolute value $|f_Q(x)|$ of the vote is the margin or the confidence of the vote.

Theorem 1 *Fix $c_t \in \mathcal{C}$ and let $\epsilon, \delta > 0$ be the accuracy and confidence parameters respectively. Let $\mathcal{H} = \{\hat{P}_{S^{(k)}} | S^{(k)} \in (X \times Y)^k\}$ and $d = VC_d(\mathcal{H})$. For any $N \geq N^*$ where*

$$N^* = O\left(\frac{1}{(\alpha_{\frac{\epsilon}{4}})^2} \log \frac{d}{\epsilon\delta(\alpha_{\frac{\epsilon}{4}})^2}\right) \tag{2}$$

if $Q \in \left((X \times Y)^k\right)^N$, then

$$Pr_Q[Pr_x[h_Q(x) \neq c_t(x)] > \epsilon] < \delta \tag{3}$$

Notice that if the sample feeding an early predictor is small enough, i.e. $O(\log d)$, then it's prediction is tolerant to noise. This observation and the fact that the voting method is tolerant to noise (cf. section 3) we conclude that

Corollary 1 *Learning by voting using early predictors is tolerant to random persistent classification noise.*

Notice that here one could use the SQ model in the following scheme:

1. Randomly select $x_1, \ldots, x_m$.
2. For each i, pool early predictors to construct a statistical query which predicts the label of x_i.
3. Assuming that all the labels are correct, use any (not necessarily noise tolerant) learning technique to generate an hypothesis based on the resulting training set.

Note that if all the x_i's are "good", i.e. non of them fall in the set of instances where α_δ bound doesn't holds, then with high probability all the resulting labels will be correct. Since SQ is noise tolerant, this method can handle noise as well. Analyzing the SQ method can be done using standard techniques.

Learning in the SQ model using early predictors might seem strange at first glance since we use a large sample to construct a smaller but uncorrupted sample (training set) which in turn is used for learning. But if one is capable of correcting a corrupted sample doesn't it mean he already learned? This observation is the heart of the proof of Theorem 1 which employs an argument based on the margin of the hypothesis [12].

We conclude by demonstrating the construction of early predictors for the class of linear separators using only one labeled instance (example). The sample space is a uniformly distributed n dimensional unit ball. We use homogeneous linear separators, i.e. each classifier is a vector v and the classification rule is $c_v(x) = \text{sign}(x \cdot v)$.

Let $z \in X$ and $l \in Y$. The following function is an early predictor for the class of homogeneous linear separators

$$\hat{P}(\langle z, l\rangle, x) = l \cdot \text{sign}(z \cdot x) \tag{4}$$

This function predicts that the label of x is the same as the label of z if the angle between them is less then $\frac{\pi}{2}$.

Theorem 2 *$\hat{P}$ is an Early Predictor:*

For every linear separator c_v and $\alpha_\delta = \frac{\gamma\delta}{\sqrt{n}}$ (γ is a constant independent of n and δ) the following holds

$$Pr_x[Pr_z[\hat{P}(\langle z, c_v(z)\rangle, x) = c_v(x)] > \frac{1}{2} + \alpha_\delta] > 1 - \delta \tag{5}$$

Plugging $\alpha_\delta = \frac{\gamma\delta}{\sqrt{n}}$ in Theorem 1 implies that the voting method is an efficient learning algorithm for this class. Moreover, since the constructed early predictor $\hat{P}$ posses certain qualities (which will be specified later), the learning algorithm is also noise tolerant.

2 Early Predictors

In Definition 1 we first introduced the notion of early predictors. We've presented two methods for exploiting the information provided by such functions. The first method uses voting on the predictions of a committee of early predictors, while the second method uses statistical queries to generate an uncorrupted training set. In this section we analyze the performance of the voting method.

The following notions governs the proof of Theorem 1. We start by drawing at random a set T that may be considered as a validation set. For each instance in T, the committee of early predictors vote on it's label. For almost all the instances, with high probability the majority of the early predictors will vote for the correct label. Moreover, the majority will be significant, i.e. with a large margin. Hence, the large margin theorem of Schapire et. al. [12] can be used to upper bound the generalization error. Finally we note, that the set T is only used for the sake of the analysis and does not have any practical meaning. For the sake of completeness, we quote the large margin theorem:

Theorem 3 [Schapire, Freund, Bartlett and Lee] *Let T be a sample of m examples chosen independently at random according to a probability distribution $\mathcal{D}$ over the sample space $\{(x, y)\,|x \in X, y \in \{-1, 1\}\}$. Suppose the hypothesis space $\mathcal{H}$ has VC-dimension d, and let $\delta > 0$. Assume that $m \geq d \geq 1$. Then with probability at least $1-\delta$ over the random choice of the set T, every weighted average function $f(x) = \sum w_i h_i(x)$ (where $w_i \geq 0 \quad \sum w_i = 1$) satisfies the following bound for every $\theta > 0$:*

$$Pr_{\mathcal{D}}[yf(x) \leq 0] \leq Pr_T[yf(x) \leq \theta] + O\left(\sqrt{\frac{1}{m}\left(\frac{d\log^2(m/d)}{\theta^2} + \log(1/\delta)\right)}\right) \tag{6}$$

Recall our definition of the majority hypothesis h_Q (Definition 2) then substituting $\Pr_{\mathcal{D}}[h_Q(x) \neq y] = \Pr_{\mathcal{D}}[yf(x) \leq 0]$ in the statement of Theorem 3 provides an upper bound to the generalization error of h_Q. The proof of theorem 1 follows.

Proof: of theorem 1

Let T be a random sample of m examples which we term the *validation set* and let $Q \in \left((X \times Y)^k\right)^N$ be an independently drawn set of samples which is used to construct a committee of N early predictors, each of them is based on k random examples (cf. Definition 2). Q will be termed the *committee set*. In order to apply Theorem 3 to bound the generalization error of the hypothesis h_Q, we shall choose N, m and θ such that the left side of (6) will be bounded by ϵ with probability $1-\delta$ over the choice of T and Q.

Let $\theta = \frac{1}{2}\alpha_{\frac{\epsilon}{4}}$ and let m be such that $m \geq O(\frac{16}{\epsilon^2}(\frac{4d\log^2(\frac{\epsilon^2}{64}\alpha^2_{\frac{\epsilon}{4}})}{\alpha^2_{\frac{\epsilon}{4}}} + \log\frac{3}{\delta}))$ then the second term in (6) is bounded by $\epsilon/2$ with probability $1 - \frac{\delta}{3}$.

In order to bound the first term let us fix the target concept and consider the following two faulty events: The first occurs when a bad validation set T is chosen and the second occurs when a bad committee set Q is chosen. At the first event there are too many instances in T for which most committees of early predictors can not guaranty reasonable predictions[3], while at the second event we fail to select a set Q upon which a committee of early predictors will be constructed, such that with high probability it will succeed to correctly predict the labeling of most of the instances in any good T. We shall choose m and N such that the probability for each one of these events is bounded by $\frac{\delta}{3}$.

Let $A(x)$ be the expected advantage of an early predictor $\hat{P}$ for the instance x over a random guess, i.e. $A(x) = \Pr_S[\hat{P}(S) = c(x)] - \frac{1}{2}$. We can restate definition 1 to be

$$\forall\delta : \Pr_x[A(x) < \alpha_\delta/2] < \delta$$

T is a bad validation set if more then $\frac{m\epsilon}{2}$ of its instances have expected advantage less then $\frac{1}{2}\alpha_{\frac{\epsilon}{4}}$. Using Chernoff bound we may conclude that if $m \geq \frac{12}{\alpha_{\frac{\epsilon}{4}}}\log\frac{3}{\delta}$, the probability that T would be a bad validation set is bounded by $\frac{\delta}{3}$.

One can look at f_Q as an estimation of $A(x)$. For our discussion we shall argue that $f_Q(x)$ is a good estimate if for every $(x, y) \in T$ such that $A(x) \geq \frac{1}{2}\alpha_{\frac{\epsilon}{4}}$ then $yf_Q(x) \geq \frac{1}{4}\alpha_{\frac{\epsilon}{4}}$ (Note that the notation $yf_Q(x)$ means the proportion of the correct prediction minus the proportion of the wrong prediction - which is exactly the margin as defined by

[3]It should be emphasized that a reasonable prediction by the committee is a correct majority vote with a large enough margin. The margin corresponds to the level of confidence in the decision of the committee.

[12]). Chernoff bound implies that if $N \geq \frac{8}{(\alpha_{\frac{\epsilon}{4}})^2} \log \frac{3m}{\delta}$ and (x, y) is such that $A(x) \geq \frac{1}{2}\alpha_{\frac{\epsilon}{4}}$ then the probability that $yf_Q(x) < \frac{1}{4}\alpha_{\frac{\epsilon}{4}}$ is less then $\frac{\delta}{3m}$. Using the union bound we conclude that with probability greater then $\frac{\delta}{3}$, the majority function f_Q is a good estimate of A for every instance in T.

If $N \geq \frac{8}{(\alpha_{\frac{\epsilon}{4}})^2} \log \frac{3m}{\delta}$ and $m \geq \frac{12}{\alpha_{\frac{\epsilon}{4}}} \log \frac{3}{\delta}$ then with probability greater then $1 - \frac{2}{3}\delta$ over the choices of Q and T

$$\Pr_{(x,y)\in T}[yf_Q(x) \leq \theta] < \frac{\epsilon}{2} \tag{7}$$

(note that $A(x) = E_Q[\frac{1}{2}yf_Q(x)]$) and so we get the bound on the first term in (6).

Let us review all the resulted conditions on N and m

1. $m \geq O\left(\frac{16}{\epsilon^2}\left(\frac{4d\log^2\left(\frac{\epsilon^2}{64}\alpha^2_{\frac{\epsilon}{4}}\right)}{\alpha^2_{\frac{\epsilon}{4}}} + \log\frac{3}{\delta}\right)\right)$
2. $m \geq \frac{12}{\alpha_{\frac{\epsilon}{4}}} \log \frac{3}{\delta}$
3. $N \geq \frac{8}{(\alpha_{\frac{\epsilon}{4}})^2} \log \frac{3m}{\delta}$

So, by setting $m \geq O(\frac{16}{\epsilon^2}(\frac{4d\log^2(\frac{\epsilon^2}{64}\alpha^2_{\frac{\epsilon}{4}})}{\alpha^2_{\frac{\epsilon}{4}}} + \log\frac{3}{\delta})) + \frac{12}{\alpha_{\frac{\epsilon}{4}}} \log \frac{3}{\delta}$ and $N^* = \frac{8}{(\alpha_{\frac{\epsilon}{4}})^2} \log \frac{3m}{\delta}$ the right side of (6) is upper bounded by $\epsilon/2 + \epsilon/2 = \epsilon$ with probability $1 - \delta$ over the choices of T and Q. Setting N^* as defined in (2) guarantees that for every $N \geq N^*$ all the above restrictions holds.

Finally note that T is used only for the sake of the analysis: f_Q does not use any validation set T to generate it's estimates for $A(x)$. It suffices to know that, with high probability, if we would have chosen T, f_Q should have big margins on almost all the instances in T. □

Note that instead of the use we made of margins, we could have used Vapnik and Chervonenkis's original bounds but this would cause the bounds to become worst as N becomes larger (as in the case of AdaBoost [12]).

3 Noise Tolerance

In the previous section we introduced the notion of early predictors: We were able to show that if such a predictor exists, the class is learnable and the sample complexity of learning was computed. In this section, we expand the analysis to the case of learning with noise. We could have proceed via SQ conversion as outlined in the introduction. However, SQ conversion is problem dependent, while direct analysis of early predictors will provide uniform conversion.

We shall assume that the noise is persistent random labeling noise. It is assumed that the learner has access to a noisy oracle $\mathcal{O}$, i.e. for each call to the oracle $\mathcal{O}$, it returns a pair (x, y) such that x is chosen according to some fixed distribution and $y = c_t(x)$ with probability $1 - \eta$, were c_t is the target concept. The learner's task is to approximate c_t, even when η is close to $\frac{1}{2}$. We shall analyze the behavior of $\hat{P}$ in this setting.

Recall that $\hat{P}$ is a function that gets a sample S of size k and an instance x and returns an estimate for the correct label of x. We will show that if the predictor has some symmetry property, then learning by using the voting method is noise tolerant. The symmetry property demands that there exists a function $f(\eta)$ such that for every x, y and $\eta < \frac{1}{2}$

$$\Pr_{S,e\sim B(k,\eta)}[\hat{P}(S \oplus e, x) = y] \geq f(\eta)\Pr_S[\hat{P}(S, x) = y] \tag{8}$$

where e is the error vector which operates on S by changing the label of S_i whenever $e_i = 1$, and $B(n, p)$ is the binomial distribution. To simplify notation and make our argument clear, we shall assume that $Y = \{-1, +1\}$ and $k = 1$. In this case $S = \{(z, y)\}$, we require that for every x and $y \in Y$

$$\Pr[\hat{P}((z, y), x) = y] = \Pr[\hat{P}((z, \bar{y}), x) = \bar{y}] \tag{9}$$

Note that the probability is over the random choices of $\hat{P}$.

Let us compute the probability that $\hat{P}$ will give correct prediction when it has access to the noisy oracle $\mathcal{O}$. Let x be such that when $\hat{P}$ has access to the "honest" oracle, then $\Pr_{(z,y)}[\hat{P}((z, y), x) = c_t(x)] = \gamma$ and hence

$$\Pr_{(z,y)\sim\mathcal{O}}[\hat{P}((z, y), x) = c_t(x)] \quad = \tag{10}$$

$$
\begin{aligned}
&= (1-\eta)\left(\text{Pr}_{(z,y)}[\hat{P}((z,y),x) = c_t(x) | y = c_t(z)]\right) \\
&\quad +\eta\left(\text{Pr}_{(z,y)}[\hat{P}((z,y),x) = c_t(x) | y = \overline{c_t(z)}]\right) &(11)\\
&= (1-\eta)\gamma + \eta(1-\gamma) &(12)
\end{aligned}
$$

Therefore, if $\eta < \frac{1}{2}$ and $\gamma \geq \frac{1}{2} + \alpha_\delta$ then

$$\text{Pr}_{(z,y)\sim\mathcal{O}}[\hat{P}((z,y),x) = c_t(x)] \geq (1-\eta)(\frac{1}{2}+\alpha_\delta) + \eta(\frac{1}{2}-\alpha_\delta) = \frac{1}{2} + \alpha_\delta(1-2\eta) \tag{13}$$

and so we can replace α_δ by the new value $\alpha_\delta(1-2\eta)$ and conclude that the voting method is noise tolerant. Note that if α_δ and $\frac{1}{2} - \eta$ are not worst then polynomial small in the learning parameters, then the learning algorithm will be efficient.

When $k > 1$ the symmetry property (9) works on the 2^k noisy combinations that $\hat{P}$ might get. Note however, that if $k = O(\log d)$ then the probability to get a sample S without any mistake is not worse than polynomial small in d. This means that if for a concept class we have any non-trivial bound on the behavior of the learning curve on the first $O(\log d)$ examples, we can turn it into a noise tolerant learning algorithm.

4 Applications for linear separators

In the previous section we developed a general approach of learning in the presence of noise. In this section we apply this method to the class of linear separators. The sample space is a uniformly distributed n dimensional unit ball and the concept class is the homogeneous linear separators. Let $z \in X$ and $l \in Y$. The following function is an early predictor for the class of homogeneous linear separators

$$\hat{P}(\langle z, l\rangle, x) = l \cdot \text{sign}(z \cdot x) \tag{14}$$

This function predicts that the label of x is the same as the label of z if the angle between them is less then $\frac{\pi}{2}$. If we can show that $\hat{P}$ is an early predictor, then since the symmetry property (9) holds for $\hat{P}$, we may conclude that learning by voting using $\hat{P}$ is noise tolerant.

Proof: theorem 2
Fix x and define θ to be the angle between x and v. The probability (over z) that $\hat{P}$ will give a wrong prediction is $\frac{2\theta}{2\pi} = \frac{\theta}{\pi}$ (see figure 1). Therefore, as long as θ is bounded from $\frac{\pi}{2}$ then

$$\text{Pr}_z[\hat{P}(\langle z, c(z)\rangle, x) = c(x)] > 1 - \frac{\theta}{\pi} \tag{15}$$

It will suffice to give a bound on the probability that θ is larger then some θ_0.

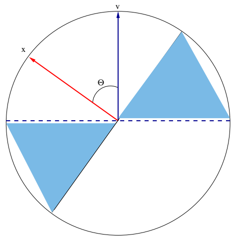

Figure 1: For z colored area, $\hat{P}$ will make a prediction error on x's label.

From the geometry of the n dimensional unit ball it is clear that

$$\text{Pr}_x[\frac{\pi}{2} + \tau_0 > \text{angle}(x,v) > \frac{\pi}{2} - \tau_0] = \int_{\frac{\pi}{2}-\tau_0}^{\frac{\pi}{2}+\tau_0} \frac{n-1}{n} \frac{\pi^{\frac{n-1}{2}}}{\pi^{\frac{n}{2}}} \frac{\Gamma(\frac{n+2}{2})}{\Gamma(\frac{n+1}{2})} \sin^{n-1}\theta d\theta \tag{16}$$

Bounding $\sin\theta$ by 1[4], we obtain the following result:

$$\text{Pr}_x[\frac{\pi}{2} + \tau_0 > \text{angle}(x,v) > \frac{\pi}{2} - \tau_0] < 2\tau_0 \frac{1}{\sqrt{\pi}} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \leq \gamma\tau_0\sqrt{n} \tag{17}$$

[4] This bound is a reasonable approximation whenever θ is at a $\frac{1}{\sqrt{n}}$ neighborhood of $\frac{\pi}{2}$

The last inequation is due to $\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} < \frac{Const}{\sqrt{n}}$. This implies that if we choose $\tau_0 = \frac{\delta}{\gamma\sqrt{n}}$, the probability that the angle between x and c will be closer to $\frac{\pi}{2}$ then τ_0, is less then δ. Hence, if x is not closer then τ_0, then

$$\mathrm{Pr}_z[\hat{P}(\langle z, c_t(z)\rangle, x) = c(x)] > \frac{1}{2} + \frac{\tau_0}{\pi} \tag{18}$$

Plugging $\alpha_\delta = \frac{\gamma\delta}{\sqrt{n}}$ we conclude that $\hat{P}$ with the constants α_δ is an early prediction function for the class of linear separators. □

5 Conclusions and Further Research

In this paper we presented the use of some properties of the very beginning of the learning curve to construct an efficient, noise tolerant, learning algorithm. Note that we've assumed that a bound on the level of noise is known, but this can be overcome easily by similar arguments to the one used in the SQ model [10]. We made use of the large margin theorem [12] in the analysis of this algorithm. This analysis suggest another perspective for understanding of the SQ model in terms of margins.

The suggested learning technique is general in the sense that the only feature that is problem dependant is the construction of an early predictor. It is also of practical nature because it enable parallelism of the learning algorithm.

The voting method we used gives equal weights to all the early predictors. Our approach is comparable to the Bagging approach [11] while AdaBoost [5] chooses weights to minimize the empirical error. We suggest that the uniform weights methods are more robust to noise then adaptive methods.

We've also presented a week predictor for the class of linear separators under uniform distribution and so concluded that this class is learnable in the presence of noise. Note that in the SVM model [13] the problem of drawing the "best" linear separator is addressed in a different approach by charging each mislabeled instance by it's distance form the hyperplane and trying to minimize the empirical charge. This could be viewed as a metric approach as opposed to the probabilistic approach presented in this paper.

Finally note that using early predictor one can actually simulate *membership oracle*. Given a point x the oracle returns the label of x (the difference between membership and a PAC oracle: The later picks x randomly). The simulation can work for almost all queries to the membership oracle.

References

[1] D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

[2] 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.

[3] S. E. Decator. *Efficient Learning from Faulty Data*. PhD thesis, Harvard University, 1995.

[4] T. G. Dietterich. An experimental comparison of three methods for constructing ensemble of decision trees: Bagging, boosting, and randomization. *Machine Learning*, pages 1–22, 1998.

[5] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[6] E. Gofer and E. Shamir. 1998. Unpublished manuscript.

[7] D. Haussler, M. Kearns, H.S. Seung, and N. Tishby. Rigorous learning curve bounds from statistical mechanics. *Machine Learning*, 25:195–236, 1997.

[8] D. P. Helmbold and M. K. Warmuth. On weak learning. *Journal of Computer and System Sciences*, 50(3):551–573, 1995.

[9] E. Jackson, E. Shamir, and C. Shwartzman. Learning with queries corrupted by classification noise. *Discrete Applied Mathematics*, 1999. to appear.

[10] 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.

[11] Breiman L. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[12] R Schapire, Y. Freund, P Bartlett, and W. Lee. Boosting in the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

[13] V. Vapnik, S. Golowich, and A Smola. Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems*, 1996.