Margin Based Active Learning

Maria-Florina Balcan[1], Andrei Broder[2], and Tong Zhang[3]

[1] Computer Science Department, Carnegie Mellon University, Pittsburgh, PA.
ninamf@cs.cmu.edu
[2] Yahoo! Research, Sunnyvale, CA 94089, USA.
broder@yahoo-inc.com
[3] Yahoo! Research, New York, 100111, USA.
tzhang@yahoo-inc.com

Abstract. We present a framework for margin based active learning of linear separators. We instantiate it for a few important cases, some of which have been previously considered in the literature. We analyze the effectiveness of our framework both in the realizable case and in a specific noisy setting related to the Tsybakov small noise condition.

1 Introduction

There has recently been substantial interest in using unlabeled data together with labeled data for machine learning. The motivation is that unlabeled data can often be much cheaper and more plentiful than labeled data, and so if useful information can be extracted from it that reduces dependence on labeled examples, this can be a significant benefit.

There are currently two settings that have been considered to incorporate unlabeled data in the learning process. The first one is the so-called *Semi-supervised Learning* [3, 5], where, in addition to a set of labeled examples drawn at random from the underlying data distribution, the learning algorithm can also use a (usually larger) set of unlabeled examples from the same distribution. In this setting, unlabeled data becomes informative under *additional* assumptions and beliefs about the learning problem. Examples of such assumptions are the one used by Transductive SVM (namely, that the target function should cut through low density regions of the space), or by Co-training (namely, that the target should be self-consistent in some way). Unlabeled data is then potentially useful in this setting because it allows one to reduce search space from the whole set of hypotheses, down to the set of *a-priori* reasonable with respect to the underlying distribution.

The second setting, an increasingly popular one for the past few years, is *Active Learning* [2, 6, 8]. Here, the learning algorithm has both the capability of drawing random unlabeled examples from the underlying distribution and that of asking for the labels of *any* of these examples, and the hope is that a good classifier can be learned with significantly fewer labels by *actively* directing the queries to *informative* examples. As opposed to the Semi-supervised learning setting, and similarly to the classical supervised learning settings (PAC and Statistical Learning Theory settings) the only prior

belief about the learning problem in the Active Learning setting is that the target function (or a good approximation of it) belongs to a given concept class. Luckily, it turns out that for simple concept classes such as linear separators on the line one can achieve an *exponential* improvement (over the usual supervised learning setting) in the labeled data sample complexity, under no additional assumptions about the learning problem [2, 6].[4] In general, however, for more complicated concept classes, the speed-ups achievable in the active learning setting depend on the match between the distribution over example-label pairs and the hypothesis class, and therefore on the target hypothesis in the class. Furthermore, there are simple examples where active learning does not help at all, even if there in the realizable case (see, for example, [8]). Recent interesting work of Dasgupta [8] gives a nice generic characterization of the sample complexity aspect of active learning in the realizable case.

A few variants and restrictions of the general active learning setting have also been considered lately. For instance the Query by Committee analysis [10] assumes realizability (i.e., there exists a perfect classifier in a known set) and a correct Bayesian prior on the set of hypotheses [10]. The analysis of the active Perceptron algorithm described in [9] relies on an even stronger assumption, of known and fixed distribution.

In the general active learning setting, for the realizable case, Cohen, Atlas and Ladner have introduced in [6] a *generic* active learning algorithm. This algorithm is a sequential algorithm that keeps track of two spaces — the current *version space* H_i, defined as the set of hypotheses in H consistent with all labels revealed so far, and the current *region of uncertainty* R_i, defined as the set of all x in the instance space X, for which there exists a pair of hypotheses in H_i that disagrees on x. In round i, the algorithm picks a random unlabeled example from R_i and queries it, eliminating all hypotheses in H_i inconsistent with the received label. The algorithm then eliminates those $x \in R_i$ on which all surviving hypotheses agree, and recurses.This algorithm was later analyzed and generalized to the non-realizable case in [2], and it was shown that in certain cases it does provide a significant improvement in the sample complexity.

In this paper we analyze a generic margin based active learning algorithm for learning linear separators and instantiate it for a few important cases, some of which have been previously considered in the literature. Specifically, the generic procedure we analyze is presented in Figure 1. To simplify calculation, we will present and analyze a few modifications of the algorithm as well.

Our Contributions: We present and analyze a framework for margin based active learning and also instantiate it for a few important cases. Specifically:

– We point out that in order to obtain a *significant* improvement in the labeled data sample complexity we have to use a strategy which is more *aggressive* than the one proposed by Cohen, Atlas and Ladner in [6] and later analyzed in [2]. We point out that this is true even in the special case when the data instances are drawn uniformly from the the unit ball in R^d, and when the labels are consistent with a linear separator going through the origin. Indeed, in order to obtain a truly exponential improvement, and to be able to learn with only $\tilde{O}\left(d \log\left(\frac{1}{\epsilon}\right)\right)$ labeled examples, we need, in each iteration, to sample our examples from a subregion carefully chosen,

[4] For this simple concept class one can achieve a pure exponential improvement [6] in the realizable case, while in the agnostic case the improvement depends upon the noise rate [2].

and not from the entire region of uncertainty, which would imply a labeled data sample complexity of $\tilde{O}\left(d^{\frac{3}{2}} \log\left(\frac{1}{\epsilon}\right)\right)$.

- We show that our algorithm and argument extend to the non-realizable case. A specific case we analyze here is again the setting where the data instances are drawn uniformly from the the unit ball in R^d, and a linear classifier w^* is the Bayes classifier. We additionally assume that our data satisfies the popular Tsybakov small noise condition along the decision boundary [14]. We consider both a simple version which leads to *exponential* improvement similar to the item 1 above, and a setting where we get only a polynomial improvement in the sample complexity, and where this is provably the best we can do [4].
- We analyze a "large margin" setting and show how active learning can dramatically improve (the supervised learning) sample complexity; the bounds we obtain here *do not depend* on the dimensionality d.
- We provide a general and unified analysis of our main algorithm – Algorithm 1.

Structure of this paper: For clarity, we start by analyzing in Section 3 the special case where the data instances are drawn uniformly from the the unit ball in R^d, and when the labels are consistent with a linear separator w^* going through the origin. We then analyze the noisy setting in Section 4, and give dimension independent bounds in a large margin setting in Section 5. We present our generic Margin Based learning algorithm and analysis in Section 6 and finish with a discussion and in Section 7.

2 Definitions and Notation

Consider the problem of predicting a binary label y based on its corresponding input vector x. As in the standard machine learning formulation, we assume that the data points (x, y) are drawn from an unknown underlying distribution P over $X \times Y$; X is called the *instance space* and Y is the *label space*. In this paper we assume that $Y = \{\pm 1\}$.

Our goal is to find a classifier f with the property that its expected true loss of $\mathrm{err}(f)$ is as small as possible. Here we assume $\mathrm{err}(f) = E_{(x,y)\sim P}\left[\ell(f(x), y)\right]$, where we use $E_{(x,y)\sim P}$ to denote the expectation with respect to the true (but unknown) underlying distribution P. Throughout the paper, without loss of generality, we assume that $f(x)$ is a real-valued function, which induces a classification rule $2I(f(x) \geq 0) - 1$, where $I(\cdot)$ is the set indicator function. The decision at $f(x) = 0$ is not important in our analysis. We consider in the following the classification error loss, defined as $\ell(f(x), y) = 1$ if $f(x)y \leq 0$ and $\ell(f(x), y) = 0$ otherwise. We denote by $\mathrm{d}(f, g)$ the probability that the two classifiers f and g predict differently on an example coming at random from P. Furthermore, for $\alpha \in [0, 1]$ we denote by $\mathrm{B}(f, \alpha)$ the set $\{g \mid \mathrm{d}(f, g) \leq \alpha\}$.

In this paper, we are interested in linear classifiers of the form $f(x) = w \cdot x$, where w is the weight vector which we need to learn from training data. We are interested in using active learning (selective sampling) algorithms to improve the performance of linear classification methods under various assumptions. In particular, we are interested in margin based selective sampling algorithms which have been widely used in practical applications (see e.g. [13]). A general version of the type of algorithm we analyze here

Input: unlabeled data set $\mathcal{U} = \{x_1, x_2, \dots, \}$
a learning algorithm $\mathcal{A}$ that learns a weight vector from labeled data
a sequence of sample sizes $0 < \tilde{m}_1 < \tilde{m}_2 < \ldots < \tilde{m}_s = \tilde{m}_{s+1}$
a sequence of cut-off values $b_k > 0$ ($k = 1, \dots, s$)
Output: classifier $\hat{w}_s$.
Label data points $x_1, \dots, x_{\tilde{m}_1}$ by a human expert
iterate $k = 1, \dots, s$
use $\mathcal{A}$ to learn weight vector $\hat{w}_k$ from the first $\tilde{m}_k$ labeled samples.
for $j = \tilde{m}_k + 1, \dots, \tilde{m}_{k+1}$
if $|\hat{w}_k \cdot x_j| > b_k$ **then** let $y_j = \text{sign}(\hat{w}_k \cdot x_j)$
else label data point x_j by a human expert
end for
end iterate

Fig. 1. Margin-based Active Learning

is described in Figure 1. Specific choices for the learning algorithm $\mathcal{A}$, sample sizes m_k, and cut-off values b_k depends on various assumptions we will make about the data, which we will investigate in details in the following sections.

3 The Realizable Case under the Uniform Distribution

We consider here a commonly studied setting in the active learning literature [7–9]. Specifically, we assume that the data instances are drawn uniformly from the the unit ball in R^d, and that the labels are consistent with a linear separator w^* going through the origin (that is $P(w^* \cdot xy \leq 0) = 0$). We assume that $\|w^*\|_2 = 1$. It is worth noting that even in this seemingly simple looking scenario, there exists an $\Omega\left(\frac{1}{\epsilon}\left(d + \log\frac{1}{\delta}\right)\right)$ lower bound on the PAC learning sample complexity [12].

We start by informally presenting why active learning is in principle possible, at least when d is constant. We show it is not difficult to improve the labeled data sample complexity from $\tilde{O}\left(\frac{d}{\epsilon}\right)$ to $\tilde{O}\left(d^{\frac{3}{2}}\log\left(\frac{1}{\epsilon}\right)\right)$. Specifically, let us consider Procedure 1, where $\mathcal{A}$ is a learning algorithm for finding a linear classifier consistent with the training data. Assume that in each iteration k, $\mathcal{A}$ finds a linear separator $\hat{w}_k$, $\|\hat{w}_k\|_2 = 1$ which is consistent with the first $\tilde{m}_k$ labeled examples. We want to ensure that $\text{err}(\hat{w}_k) \leq \frac{1}{2^k}$ (with large probability), which (by standard VC bounds) requires a sample of size $\tilde{m}_k = \tilde{O}\left(2^k d\right)$; note that this implies we need to add in each iteration about $m_k = \tilde{m}_{k+1} - \tilde{m}_k = \tilde{O}\left(2^k d\right)$ new labeled examples. The desired result will follow if we can show that by choosing appropriate b_k, we only need to ask the human expert to label $\tilde{O}(d^{3/2})$ out of the $m_k = \tilde{O}\left(2^k d\right)$ data points and ensure that all m_k data points are correctly labeled (i.e. the examples labeled automatically are in fact correctly labeled).

Note that given our assumption about the data distribution the error rate of any given separator w is $\text{err}(w) = \frac{\theta(w,w^*)}{\pi}$, where $\theta(w, w^*) = \arccos(w \cdot w^*)$. Therefore $\text{err}(\hat{w}_k) \leq 2^{-k}$ implies that $\|\hat{w}_k - w^*\|_2 \leq 2^{-k}\pi$. This implies we can *safely* label all the points with $|\hat{w}_k \cdot x| \geq 2^{-k}\pi$ because w^* and $\hat{w}_k$ predict the same on those

examples. The probability of x such that $|\hat{w}_k \cdot x| \leq 2^{-k}\pi$ is $\tilde{O}(2^{-k}\sqrt{d})$ because in high dimensions, the 1-dimensional projection of uniform random variables in the unit ball is approximately a Gaussian variable with variance $1/d$. Therefore if we let $b_k = 2^{-k}\pi$ in the k-th iteration, and draw $m_{k+1} - m_k = \tilde{O}\big(2^k d\big)$ new examples to achieve an error rate of $2^{-(k+1)}$ for $\hat{w}_{k+1}$, the expected number of human labels needed is at most $\tilde{O}(d^{\frac{3}{2}})$. This essentially implies the desired result. For a high probability statement, we can use Procedure 2, which is a modification of Procedure 1.

Input: allowed error rate ϵ, probab. of failure δ, a sampling oracle for P_X, a labeling oracle
a sequence of sample sizes $m_k > 0$, $k \in Z^+$; a sequence of cut-off values $b_k > 0$, $k \in Z^+$
Output: weight vector $\hat{w}_s$ of error at most ϵ with probability $1 - \delta$
Draw m_1 examples from P_X, label them and put into a working set $W(1)$.
iterate $k = 1, \ldots, s$
find a hypothesis $\hat{w}_k$ ($\|\hat{w}_k\|_2 = 1$) consistent with all labeled examples in $W(k)$.
let $W(k+1) = W(k)$.
until m_{k+1} additional data points are labeled, draw sample x from P_X
if $|\hat{w}_k \cdot x| \geq b_k$, reject x
otherwise, ask for label of x, and put into $W(k+1)$
end iterate

Fig. 2. Margin-based Active Learning (separable case)

Note that we can apply our favorite algorithm for finding a consistent linear separator (e.g., SVM for the realizable case, linear programming, etc.) at each iteration of Procedure 2, and the overall procedure is *computationally efficient*.

Theorem 1. *There exists a constant C, s. t. for any $\epsilon, \delta > 0$, using Procedure 2 with $b_k = \frac{\pi}{2^{k-1}}$ and $m_k = Cd^{\frac{1}{2}}\left(d\ln d + \ln\frac{k}{\delta}\right)$, after $s = \lceil \log_2 \frac{1}{\epsilon} \rceil$ iterations, we find a separator of error at most ϵ with probability $1 - \delta$.*

Proof. The proof is a rigorous version of the informal one given earlier. We prove by induction on k that at the k'th iteration, with probability $1 - \delta(1 - 1/(k+1))$, we have $\mathrm{err}(\hat{w}) \leq 2^{-k}$ for all separators $\hat{w}$ consistent with data in the set $W(k)$; in particular, $\mathrm{err}(\hat{w}_k) \leq 2^{-k}$.

For $k = 1$, according to Theorem 7 in Appendix A, we only need $m_1 = O(d + \ln(1/\delta))$ examples to obtain the desired result. In particular, we have $\mathrm{err}(\hat{w}_1) \leq 1/2$ with probability $1 - \delta/2$. Assume now the claim is true for $k-1$. Then at the k-th iteration, we can let $S_1 = \{x : |\hat{w}_{k-1} \cdot x| \leq b_{k-1}\}$ and $S_2 = \{x : |\hat{w}_{k-1} \cdot x| > b_{k-1}\}$. Using the notation $\mathrm{err}(w|S) = \Pr_x((w \cdot x)(w^* \cdot x) < 0 | x \in S)$, for all $\hat{w}$ we have:

$$\mathrm{err}(\hat{w}) = \mathrm{err}(\hat{w}|S_1)\Pr(S_1) + \mathrm{err}(\hat{w}|S_2)\Pr(S_2).$$

Consider an arbitrary $\hat{w}$ consistent with the data in $W(k-1)$. By induction hypothesis, we know that with probability at least $1 - \delta(1 - 1/k)$, both $\hat{w}_{k-1}$ and $\hat{w}$ have errors at most 2^{1-k} (because both are consistent with $W(k-1)$). As discussed earlier, this implies that $\|\hat{w}_{k-1} - w^*\|_2 \leq 2^{1-k}\pi$ and $\|\hat{w} - w^*\|_2 \leq 2^{1-k}\pi$. So, $\forall x \in S_2$, we have

$(\hat{w}_{k-1} \cdot x)(\hat{w} \cdot x) > 0$ and $(\hat{w}_{k-1} \cdot x)(w^* \cdot x) > 0$. This implies that $\mathrm{err}(\hat{w}|S_2) = 0$. Now using the estimate provided in Lemma 4 with $\gamma_1 = b_{k-1}$ and $\gamma_2 = 0$, we obtain $\Pr_x(S_1) \leq b_{k-1}\sqrt{4d/\pi}$. Therefore $\mathrm{err}(\hat{w}) \leq 2^{2-k}\sqrt{4\pi d} \cdot \mathrm{err}(\hat{w}|S_1)$, for all $\hat{w}$ consistent with $W(k-1)$. Now, since we are labeling m_k data points in S_1 at iteration $k-1$, it follows from Theorem 7 that we can find C s. t. with probability $1-\delta/(k^2+k)$, for all $\hat{w}$ consistent with the data in $W(k)$, $\mathrm{err}(\hat{w}|S_1)$, the error of $\hat{w}$ on S_1, is no more than $1/(4\sqrt{4\pi d})$. That is we have $\mathrm{err}(\hat{w}) \leq 2^{-k}$ with probability $1 - \delta((1 - 1/k) + 1/(k^2 + k)) = 1 - \delta(1 - 1/(k + 1))$ for all $\hat{w}$ consistent with $W(k)$, and in particular $\mathrm{err}(\hat{w}_k) \leq 2^{-k}$, as desired. □

The choice of rejection region in Theorem 1 essentially follows the idea in [6]. It was suggested there that one should not sample from a region (S_2 in the proof) in which all classifiers in the current version space (in our case, classifiers consistent with the labeled examples in $W(k)$) predict the same label. A more general version, with theoretical analysis, was considered in [2]. Here we have used a more a refined VC-bound for the realizable case, e.g., Theorem 7, to get a better bound. However, the strategy of choosing b_k in Theorem 1 (thus the idea of [6]) is not optimal. This can be seen from the proof, in which we showed $\mathrm{err}(\hat{w}_s|S_2) = 0$. If we enlarge S_2 (using a smaller b_k), we can still ensure that $\mathrm{err}(\hat{w}_s|S_2)$ is small; furthermore, $\Pr(S_1)$ becomes smaller, which allows us to use fewer labeled examples to achieve the same reduction in error. Therefore in order to show that we can achieve an improvement from $\tilde{O}\big(\frac{d}{\epsilon}\big)$ to $\tilde{O}\big(d \log\big(\frac{1}{\epsilon}\big)\big)$ as in [9], we need a more *aggressive* strategy. Specifically, at round k we set as margin parameter $b_k = \tilde{O}\Big(\frac{\log(k)}{2^k\sqrt{d}}\Big)$, and in consequence use fewer examples to transition between rounds. In order to prove correctness we need to refine the analysis as follows:

Theorem 2. *There exists a constant C s. t. for $d \geq 4$, and for any $\epsilon, \delta > 0$, $\epsilon < 1/4$, using Procedure 2 with $m_k = C\sqrt{\ln(1+k)}\big(d\ln(1+\ln k) + \ln\frac{k}{\delta}\big)$ and $b_k = 2^{1-k}\pi d^{-1/2}\sqrt{5+\ln(1+k)}$, after $s = \lceil \log_2 \frac{1}{\epsilon} \rceil - 2$ iterations, we find a separator of error $\leq \epsilon$ with probability $1 - \delta$.*

Proof. As in Theorem 1, we prove by induction on k that at the k's iteration, for $k \leq s$, with probability at least $1 - \delta(1 - 1/(k + 1))$, we $\mathrm{err}(\hat{w}) \leq 2^{-k-2}$ for all choices of $\hat{w}$ consistent with data in the working set $W(k)$; in particular $\mathrm{err}(\hat{w}_k) \leq 2^{-k-2}$.

For $k = 1$, according to Theorem 7, we only need $m_k = O(d + \ln(1/\delta))$ examples to obtain the desired result; in particular, we have $\mathrm{err}(\hat{w}_1) \leq 2^{-k-2}$ with probability $1-\delta/(k+1)$. Assume now the claim is true for $k-1$ ($k > 1$). Then at the k-th iteration, we can let $S_1 = \{x : |\hat{w}_{k-1} \cdot x| \leq b_{k-1}\}$ and $S_2 = \{x : |\hat{w}_{k-1} \cdot x| > b_{k-1}\}$. Consider an arbitrary $\hat{w}$ consistent with the data in $W(k-1)$. By induction hypothesis, we know that with probability $1 - \delta(1 - 1/k)$, both $\hat{w}_{k-1}$ and $\hat{w}$ have errors at most 2^{-k-1}, implying $\theta(\hat{w}_{k-1}, w^*) \leq 2^{-k-1}\pi$ and $\theta(\hat{w}, w^*) \leq 2^{-k-1}\pi$. Therefore $\theta(\hat{w}, \hat{w}_{k-1}) \leq 2^{-k}\pi$. Let $\tilde{\beta} = 2^{-k}\pi$ and using $\cos\tilde{\beta}/\sin\tilde{\beta} \leq 1/\tilde{\beta}$ and $\sin\tilde{\beta} \leq \tilde{\beta}$ it is easy to verify that $b_{k-1} \geq 2\sin\tilde{\beta}d^{-1/2}\sqrt{5 + \ln\left(1 + \sqrt{\ln\max(1, \cos\tilde{\beta}/\sin\tilde{\beta})}\right)}$. By Lemma 7, we have both

$\Pr_x[(\hat{w}_{k-1} \cdot x)(\hat{w} \cdot x) < 0, x \in S_2] \leq \frac{\sin\tilde{\beta}}{e^5 \cos\tilde{\beta}} \leq \frac{\sqrt{2}\tilde{\beta}}{e^5}$ and
$\Pr_x[(\hat{w}_{k-1} \cdot x)(w^* \cdot x) < 0, x \in S_2] \leq \frac{\sin\tilde{\beta}}{e^5 \cos\tilde{\beta}} \leq \frac{\sqrt{2}\tilde{\beta}}{e^5}$.

Taking the sum, we obtain $\Pr_x[(\hat{w} \cdot x)(w^* \cdot x) < 0, x \in S_2] \leq \frac{2\sqrt{2}\tilde{\beta}}{e^5} \leq 2^{-(k+3)}$. Using now Lemma 4 we get that for all $\hat{w}$ consistent with the data in $W(k-1)$ we have:

$$\begin{aligned}\mathrm{err}(\hat{w}) \leq & \mathrm{err}(\hat{w}|S_1)\Pr(S_1) + 2^{-(k+3)} \leq \mathrm{err}(\hat{w}_k|S_1) b_{k-1}\sqrt{4d/\pi} + 2^{-(k+3)} \\ \leq & 2^{-(k+2)}\left(\mathrm{err}(\hat{w}|S_1) 16\sqrt{4\pi}\sqrt{5+\ln(1+k)} + 1/2\right).\end{aligned}$$

Since we are labelling m_k points in S_1 at iteration $k-1$, we know from Theorem 7 that $\exists C$ s. t. with probability $1-\delta/(k+k^2)$ we have $\mathrm{err}(\hat{w}_k|S_1)16\sqrt{4\pi}\sqrt{5+\ln(1+k)} \leq 0.5$ for all $\hat{w}$ consistent with $W(k)$; so, with probability $1-\delta((1-1/k)+1/(k+k^2)) = 1-\delta(1-1/(k+1))$, we have $\mathrm{err}(\hat{w}) \leq 2^{-k-2}$ for all $\hat{w}$ consistent with $W(k)$. □

The bound in Theorem 2 is generally better than the one in Theorem 1 due to the improved dependency on d in m_k. However, m_k depends on $\sqrt{\ln k}\ln\ln k$, for $k \leq \lceil \log_2 \frac{1}{\epsilon} \rceil - 2$. Therefore when $d \ll \ln k(\ln\ln k)^2$, Theorem 1 offers a better bound. Note that the strategy used in Theorem 2 is more aggressive than the strategy used in the selective sampling algorithm of [2, 6]. Indeed, we do not sample from the entire region of uncertainty – but we sample just from a subregion carefully chosen. This helps us to get rid of the undesired $d^{1/2}$. Clearly, our analysis also holds with very small modifications when the input distribution comes from a high dimensional Gaussian.

4 The Non-realizable Case under the Uniform Distribution

We show that a result similar to Theorem 2 can be obtained even for non-separable problems. The non-realizable (noisy) case for active learning in the context of classification was recently explored in [2, 4]. We consider here a model which is related to the simple one-dimensional problem in [4], which assumes that the data satisfy the increasingly popular Tsybakov small noise condition along the decision boundary[14]. We first consider a simple version which still leads to exponential convergence similar to Theorem 2. Specifically, we still assume that the data instances are drawn uniformly from the the unit ball in R^d, and a linear classifier w^* is the Bayes classifier. However, we do not assume that the Bayes error is zero. We consider the following low noise condition: there exists a known parameter $\beta > 0$ such that:

$$P_X(|P(Y=1|X) - P(Y=-1|X)| \geq 4\beta) = 1.$$

In supervised learning, such a condition can lead to fast convergence rates. As we will show in this section, the condition can also be used to quantify the effectiveness of active-learning. The key point is that this assumption implies the stability condition required for active learning:

$$\beta \min\left(1, \frac{4\theta(w,w^*)}{\pi}\right)^{1/(1-\alpha)} \leq \mathrm{err}(w) - \mathrm{err}(w^*) \qquad (1)$$

with $\alpha = 0$. We analyze here a more general setting with $\alpha \in [0, 1)$. As mentioned already, the one dimensional setting was examined in [4]. We call $\mathrm{err}(w) - \mathrm{err}(w^*)$ the *excess error* of w. In this setting, Procedure 2 needs to be slightly modified, as in Figure 3.

Input: allowed error rate ϵ, probab. of failure δ, a sampling oracle for P_X, and a labeling oracle
a sequence of sample sizes $m_k > 0$, $k \in Z^+$; a sequence of cut-off values $b_k > 0$, $k \in Z^+$
a sequence of hypothesis space radii $r_k > 0$, $k \in Z^+$;
a sequence of precision values $\epsilon_k > 0$, $k \in Z^+$
Output: weight vector $\hat{w}_s$ of excess error at most ϵ with probability $1 - \delta$
Pick random $\hat{w}_0$: $\|\hat{w}_0\|_2 = 1$.
Draw m_1 examples from P_X, label them and put into a working set W.
iterate $k = 1, \ldots, s$
find $\hat{w}_k \in B(\hat{w}_{k-1}, r_k)$ ($\|\hat{w}_k\|_2 = 1$) to approximately minimize training error:
$\sum_{(x,y)\in W} I(\hat{w}_k \cdot xy) \leq \min_{w \in B(\hat{w}_{k-1}, r_k)} \sum_{(x,y)\in W} I(w \cdot xy) + m_k \epsilon_k$.
clear the working set W
until m_{k+1} additional data points are labeled, draw sample x from P_X
if $|\hat{w}_k \cdot x| \geq b_k$, reject x
otherwise, ask for label of x, and put into W
end iterate

Fig. 3. Margin-based Active Learning (non-separable case)

Theorem 3. *Let $d \geq 4$. Assume there exists a weight vector w^* s. t. the stability condition (1) holds. Then there exists a constant C, s. t. for any $\epsilon, \delta > 0$, $\epsilon < \beta/8$, using Procedure 3 with $b_k = 2^{-(1-\alpha)k} \pi d^{-1/2} \sqrt{5 + \alpha k \ln 2 - \ln \beta + \ln(2+k)}$, $r_k = 2^{-(1-\alpha)k-2}\pi$ for $k > 1$, $r_1 = \pi$, $\epsilon_k = 2^{-\alpha(k-1)-4}\beta / \sqrt{5 + \alpha k \ln 2 - \ln \beta + \ln(1+k)}$, and $m_k = C\epsilon_k^{-2}\left(d + \ln \frac{k}{\delta}\right)$, after $s = \lceil \log_2(\beta/\epsilon) \rceil$ iterations, we find a separator with excess error $\leq \epsilon$ with probability $1 - \delta$.*

Proof. The proof is similar to that of Theorem 2. We prove by induction on k that after $k \leq s$ iterations, $\mathrm{err}(\hat{w}_k) - \mathrm{err}(w^*) \leq 2^{-k}\beta$ with probability $1 - \delta(1 - 1/(k+1))$.

For $k = 1$, according to Theorem 8, we only need $m_k = \beta^{-2} O(d + \ln(k/\delta))$ examples to obtain $\hat{w}_1$ with excess error $2^{-k}\beta$ with probability $1 - \delta/(k+1)$. Assume now the claim is true for $k-1$ ($k \geq 2$). Then at the k-th iteration, we can let $S_1 = \{x : |\hat{w}_{k-1} \cdot x| \leq b_{k-1}\}$ and $S_2 = \{x : |\hat{w}_{k-1} \cdot x| > b_{k-1}\}$. By induction hypothesis, we know that with probability at least $1 - \delta(1 - 1/k)$, $\hat{w}_{k-1}$ has excess errors at most $2^{-k+1}\beta$, implying $\theta(\hat{w}_{k-1}, w^*) \leq 2^{-(1-\alpha)(k-1)}\pi/4$. By assumption, $\theta(\hat{w}_{k-1}, \hat{w}_k) \leq 2^{-(1-\alpha)k-2}\pi$. Let $\tilde{\beta} = 2^{-(1-\alpha)k-2}\pi$ and using $\cos\tilde{\beta}/\sin\tilde{\beta} \leq 1/\tilde{\beta}$ and $\sin\tilde{\beta} \leq \tilde{\beta}$, it is easy to verify that $b_{k-1} \geq 2\sin\tilde{\beta} d^{-1/2} \sqrt{5 + \alpha k \ln 2 - \ln\beta + \ln\left(1 + \sqrt{\ln(\cos\tilde{\beta}/\sin\tilde{\beta})}\right)}$. From Lemma 7, we have both

$$\Pr_x\left[(\hat{w}_{k-1} \cdot x)(\hat{w}_k \cdot x) < 0, x \in S_2\right] \leq \frac{\sin\tilde{\beta}}{e^5 \beta^{-1} 2^{\alpha k} \cos\tilde{\beta}} \leq \frac{\sqrt{2}\tilde{\beta}\beta}{2^{\alpha k} e^5} \quad \text{and}$$

$\Pr_x\left[(\hat{w}_{k-1}\cdot x)(w^*\cdot x)<0, x\in S_2\right]\leq \frac{\sin\tilde{\beta}}{e^5\beta^{-1}2^{\alpha k}\cos\tilde{\beta}}\leq \frac{\sqrt{2}\tilde{\beta}\beta}{2^{\alpha k}e^5}$.

Taking the sum, we obtain $\Pr_x\left[(\hat{w}_k\cdot x)(w^*\cdot x)<0, x\in S_2\right]\leq \frac{2\sqrt{2}\tilde{\beta}\beta}{2^{\alpha k}e^5}\leq 2^{-(k+1)}\beta$. Therefore we have (using Lemma 4):

$$\begin{aligned}\mathrm{err}(\hat{w}_k)-\mathrm{err}(w^*)\leq&(\mathrm{err}(\hat{w}_k|S_1)-\mathrm{err}(w^*|S_1))\Pr(S_1)+2^{-(k+1)}\beta\\ \leq&(\mathrm{err}(\hat{w}_k|S_1)-\mathrm{err}(w^*|S_1))b_{k-1}\sqrt{4d/\pi}+2^{-(k+1)}\beta\\ \leq&2^{-k}\beta\left((\mathrm{err}(\hat{w}_k|S_1)-\mathrm{err}(w^*|S_1))\sqrt{\pi}/(4\epsilon_k)+1/2\right).\end{aligned}$$

By Theorem 7, we can choose C s. t. with m_k samples, we obtain $\mathrm{err}(\hat{w}_k|S_1)-\mathrm{err}(w^*|S_1)\leq 2\epsilon_k/\sqrt{\pi}$ with probability $1-\delta/(k+k^2)$. Therefore $\mathrm{err}(\hat{w}_k)-\mathrm{err}(w^*)\leq 2^{-k}\beta$ with probability $1-\delta((1-1/k)+1/(k+k^2))=1-\delta(1-1/(k+1))$. □

If $\alpha=0$, then we can achieve exponential convergence similar to Theorem 2, even for *noisy* problems. However, for $\alpha\in(0,1)$, we must label $\sum_k m_k=O(\epsilon^{-2\alpha}\ln(1/\epsilon)(d+\ln(s/\delta))$ examples[5] to achieve an error rate of ϵ That is, we only get a polynomial improvement compared to the batch learning case (with sample complexity between $O(\epsilon^{-2})$ and $O(\epsilon^{-1})$). In general, one *cannot* improve such polynomial behavior – see [4] for some simple one-dimensional examples.

Note: Instead of rejecting x when $|\hat{w}_k\cdot x|\geq b_k$, we can add them to W using the automatic labels from $\hat{w}_k$. We can then remove the requirement $\hat{w}_k\in B(\hat{w}_{k-1},r_k)$ (thus removing the parameters r_k). The resulting procedure will have the same convergence behavior as Theorem 3 because the probability of making error by $\hat{w}_k$ when $|\hat{w}_k\cdot x|\geq b_k$ is no more than $2^{-(k+2)}\beta$.

5 Dimension Independent Bounds

Although we showed that active learning can improve sample complexity, the bounds depend on the dimensionality d. In many practical problems, such dependency can be removed if the classifier can separate the data with large margin. We consider the following simple case, with x drawn from a d-dimensional Gaussian with bounded total variance: $x\sim N(0,\Sigma)$, $\Sigma=\mathrm{diag}(\sigma_1^2,\ldots,\sigma_d^2)$ and $\sigma_1\geq\cdots\geq\sigma_d>0$. Note that $\mathbf{E}_x\|x\|_2^2=\sum_j\sigma_j^2$. The Gaussian assumption can also be replaced by other similar assumptions such as uniform distribution in an ellipsoid. We employ the Gaussian assumption for computational simplicity. We assume further that the label is consistent with a weight vector w^* with $\|w^*\|_2=1$. However, if we do not impose any restrictions on w^*, then it is not possible to learn w^* without the d-dependence. A standard assumption that becomes popular in recent years is to assume that w^* achieves a good margin distribution. In particular, we may impose the following margin distribution condition $\forall\gamma>0$:

$$P_x(|w^*\cdot x|\leq\gamma)\leq\frac{2\gamma}{\sqrt{2\pi}\sigma}\qquad(2)$$

Condition (2) says that the probability of small margin is small. Since the projection $w^*\cdot x$ is normal with variance $\sigma^2=\sum_j\sigma_j^2(w_j^*)^2$, the margin condition (2) can be replaced by

[5] We are ignoring dependence on β here.

$$\|w^*\|_\Sigma \geq \sigma \qquad (3)$$

where $\|\xi\|_\Sigma = \sqrt{\sum_j \xi_j^2 \sigma_j^2}$, which says that the variance of x projected to w^* is at least σ. This condition restricts the hypothesis space containing w^* so that we may develop a learning bound that is independent of d. Although one can explicitly impose a margin constraint based on (3), for simplicity, we shall consider a different method here that approximates w^* with a vector in a small dimensional space. Lemma 1 shows that it is possible. For $w, w' \in R^d$, we define $\theta_\Sigma(w, w') = \arccos \frac{\sum_j \sigma_j^2 w_j w'_j}{\|w\|_\Sigma \|w'\|_\Sigma}$.

Lemma 1. *If w^* with $\|w^*\|_2 = 1$ satisfies (3) and let $w^*[k] = [w_1^*, \ldots, w_k^*, 0, \ldots, 0]$, then $\sin \theta_\Sigma(w^*, w^*[k]) \leq \sigma_{k+1}/\sigma$.*

Proof. By assumption, we have:
$\sin(\theta_\Sigma(w^*, w^*[k]))^2 = \frac{\sum_{j=k+1}^d \sigma_j (w_j^*)^2}{\sum_{j=1}^d \sigma_j^2 (w_j^*)^2} \leq \sigma_{k+1}^2 \frac{\sum_{j=k+1}^d (w_j^*)^2}{\sum_{j=1}^d \sigma_j^2 (w_j^*)^2} \leq \sigma_{k+1}^2 \frac{\sum_j (w_j^*)^2}{\sum_j \sigma_j^2 (w_j^*)^2} = (\sigma_{k+1}/\sigma)^2$, as desired. □

Note that the error of classifier w is given by $\mathrm{err}(w) = \frac{\theta_\Sigma(w, w^*)}{\pi}$. Therefore Lemma 1 shows that under the margin distribution condition (2), it is possible to approximate w^* using a low dimensional $w^*[k]$ with small error. We can now prove that:

Theorem 4. *Assume that the true separator w^* with $\|w^*\|_2 = 1$ satisfies (3). There exists C s. t. $\forall \epsilon, \delta > 0$, $\epsilon < 1/8$, using Procedure 4 with $b_k = 2^{1-k}\pi\sqrt{5 + \ln(1+k)}$, $b_0 = 0$, $d_k = \inf\{\ell : \sin(2^{-(k+4)} e^{-b_{k-1}^2/2}\pi \geq \sigma_{\ell+1}/\sigma\}$, $r_k = 2^{-k}\pi$ for $k > 1$, $r_1 = \pi$, $\epsilon_k = 2^{-5}/\sqrt{5 + \ln(1+k)}$, and $m_k = C\epsilon_k^{-2}\left(d_k + \ln \frac{k}{\delta}\right)$, after $s = \lceil \log_2 \left(\frac{1}{\epsilon}\right) \rceil - 2$ iterations, we find a separator with excess error $\leq \epsilon$ with probability $1 - \delta$.*

Proof. We prove by induction on k that after $k \leq s$ iterations, $\mathrm{err}(\hat{w}_k) - \mathrm{err}(w^*) \leq 2^{-(k+2)}$ with probability $1 - \delta(1 - 1/(k+1))$. Note that by Lemma 1, the choice of d_k ensures that $\theta_\Sigma(w^*, w^*[d_k]) \leq 2^{-(k+3)}\pi$, and thus $\mathrm{err}(w^*[d_k]) \leq 2^{-(k+3)}$.

For $k = 1$, according to Theorem 7, we only need $m_k = O(d_k + \ln(k/\delta))$ examples to obtain $\hat{w}_1 \in \mathcal{H}[d_k]$ with excess error $2^{-(k+2)}$ with probability $1 - \delta/(k+1)$. Assume now the claim is true for $k-1$ ($k \geq 2$). Then at the k-th iteration, we can let $S_1 = \{x : |\hat{w}_{k-1} \cdot x| \leq b_{k-1}\}$ and $S_2 = \{x : |\hat{w}_{k-1} \cdot x| > b_{k-1}\}$. By induction hypothesis, we know that with probability at least $1 - \delta(1 - 1/k)$, $\hat{w}_{k-1}$ has excess errors at most $2^{-(k+1)}$, implying $\theta(\hat{w}_{k-1}, w^*) \leq 2^{-(k+1)}\pi$. By assumption, $\theta(\hat{w}_{k-1}, \hat{w}_k) \leq 2^{-k}\pi$. Let $\tilde{\beta} = 2^{-k}\pi$ and use $\cos\tilde{\beta}/\sin\tilde{\beta} \leq 1/\tilde{\beta}$ and $\sin\tilde{\beta} \leq \tilde{\beta}$, it is easy to verify that the following inequality holds $b_{k-1} \geq \sqrt{2}\sin\tilde{\beta}\sqrt{5 + \ln\left(1 + \sqrt{\ln(\cos\tilde{\beta}/\sin\tilde{\beta})}\right)}$.

Let $P = \Pr_x[(\hat{w}_{k-1} \cdot x)(\hat{w}_k \cdot x) < 0, x \in S_2]$, and let $(\xi_1, \xi_2) \sim N(0, I_{2\times 2})$ and $\theta = \theta_\Sigma(\hat{w}_k, \hat{w}_{k-1})$. By Lemma 3, we have

$$\begin{aligned}
P =& 2\Pr_x[\xi_1 \leq 0, \xi_1\cos(\theta) + \xi_2\sin(\theta) \geq b_{k-1}] \\
\leq& 2\Pr_x\left[\xi_1 \leq 0, \xi_1 + \xi_2\sin(\tilde{\beta})/\cos(\tilde{\beta}) \geq b_{k-1}/\cos(\tilde{\beta})\right] \\
\leq& \frac{\sin\tilde{\beta}}{\cos\tilde{\beta}}\left(1 + \sqrt{\ln(\cos(\tilde{\beta})/\sin(\tilde{\beta}))}\right)e^{-b_{k-1}^2/(2\sin(\tilde{\beta})^2)} \leq \frac{\sqrt{2}\tilde{\beta}}{e^5}.
\end{aligned}$$

Similarly, we also have $\Pr_x\left[(\hat{w}_{k-1}\cdot x)(w^*\cdot x)<0, x\in S_2\right]\le\frac{\sqrt{2}\tilde{\beta}}{e^5}$. This implies that $\Pr_x\left[(\hat{w}_k\cdot x)(w^*\cdot x)<0, x\in S_2\right]\le\frac{2\sqrt{2}\tilde{\beta}}{e^5}\le 2^{-(k+3)}$. Now using Lemma 2, we have

$$\begin{aligned}\mathrm{err}(\hat{w}_k)\le&\mathrm{err}(\hat{w}_k|S_1)\Pr(S_1)+2^{-(k+3)}\le\mathrm{err}(\hat{w}_k|S_1)b_{k-1}/\sqrt{2\pi}+2^{-(k+3)}\\ \le&2^{-(k+2)}\left(\mathrm{err}(\hat{w}_k|S_1)8\sqrt{5+\ln(1+k)}+1/2\right).\end{aligned}$$

Our choice of d_k ensures that $\mathrm{err}(w^*[d_k]|S_1)\le 2^{-6}/\sqrt{5+\ln k}$. From Theorem 8, we know it is possible to choose a constant C such that with m_k samples we have $\mathrm{err}(\hat{w}_k|S_1)8\sqrt{5+\ln(1+k)}\le 0.5$ with probability $1-\delta/(k+k^2)$. Hence $\mathrm{err}(\hat{w}_k)\le 2^{-k-2}$ with probability $1-\delta((1-1/k)+1/(k+k^2))=1-\delta(1-1/(k+1))$. □

Input: allowed error rate ϵ, probab. of failure δ, a sampling oracle for P_X, and a labeling oracle
$\Sigma=\mathrm{diag}(\sigma_1^2,\ldots,\sigma_d^2)$, a sequence of sample sizes $m_k>0$, $k\in Z^+$
a sequence of cut-off values $b_k>0$, $k\in Z^+$ and one of hypothesis space radii $r_k>0$, $k\in Z^+$
a sequence of hypothesis space dimensions $d_k>0$, $k\in Z^+$
a sequence precision values $\epsilon_k>0$, $k\in Z^+$.
Output: weight vector $\hat{w}_s$ of excess error at most ϵ with probability $1-\delta$
Pick random $\hat{w}_0$: $\|\hat{w}_0\|_\Sigma=1$.
Draw m_1 examples from P_X, label them and put into a working set W.
iterate $k=1,\ldots,s$
find $\hat{w}_k\in\mathcal{H}[d_k]$ ($\|\hat{w}_k\|_\Sigma=1, \|\hat{w}_k-\hat{w}_{k-1}\|_\Sigma\le 2(1-\cos(r_k))$) such that
$\sum_{(x,y)\in W} I(\hat{w}_k\cdot xy)\le m_k\epsilon_k$,
where $\mathcal{H}[d_k]=\{w\in R^d : w_{d_k+1}=\cdots=w_d=0\}$
clear the working set W
until m_{k+1} additional data points are labeled, draw sample x from P_X
if $|\hat{w}_k\cdot x|\ge b_k$, reject x
otherwise, ask for label of x, and put into W
end iterate

Fig. 4. Margin-based Active Learning (with low-dimensional approximation)

Using a more refined ratio VC-bound, one can easily improve the choice of $m_k=C\epsilon_k^{-2}(d_k+\ln(k/\delta))$ to $m_k=C\epsilon_k^{-1}(d_k\ln\epsilon^{-1}+\ln(k/\delta)$ in Theorem 4. In Algorithm 4, instead of putting constraint of $\hat{w}_k$ using r_k, one can also use $\hat{w}_{k-1}$ to label data x and put them into the working set W such that $|\hat{w}_{k-1}\cdot x|\ge b_{k-1}$, which introduces error at most $2^{-(k+3)}$. One may then train a $\hat{w}_k$ using labeled data in W without the constraint $\|\hat{w}_k-\hat{w}_{k-1}\|_\Sigma\le 2(1-\cos(r_k))$; the results will be similar.
The sample complexity of Procedure 4 depends on d_k which is determined by the decay of σ_k instead of d. In particular we can consider a few possible decays with $d=\infty$:

– Assume $\sigma_k\le O(2^{-\beta k})$ with constant $\beta>0$, which is the eigenvalue decaying behavior for exponential kernels. In this case d_k is $O(k/\beta)$. Therefore we only need $m_k=O(k^2\ln k)$ examples at each iteration k.

– Assume $\sigma_k \leq O(k^{-\beta})$ with constant $\beta > 0$, which is the eigenvalue decaying behavior for spline kernels. In this case d_k is $O(2^{k/\beta})$. Therefore we need $m_k = \tilde{O}(2^{k/\beta})$ examples at each iteration k. The total samples needed to achieve accuracy ϵ is $\tilde{O}(\epsilon^{-1/\beta})$. Note that when $\beta > 1$, we achieve faster than $O(1/\epsilon)$.
– When the total variation is bounded: $\sum_j \sigma_j^2 \leq 1$, which means that $\|x\|_2$ is bounded on average (corresponding to standard large margin kernel methods with bounded $\|x\|_2$), then $\sigma_k \leq 1/\sqrt{k}$. Therefore we can take $d_k = O(2^{2k})$ and $m_k = \tilde{O}(2^{2k})$. The total sample size needed to achieve accuracy ϵ is $\tilde{O}(\epsilon^{-2})$. The constant will depend on the margin $\sigma/\sqrt{\sum_j \sigma_j^2}$ but independent of the dimensionality d which is infinity.

6 A General Analysis for Margin Based Active Learning

We show here a general bound for Algorithm 1 based on assumptions about the algorithm $\mathcal{A}$, the sample sizes m_k, and the thresholds b_k. This is a more abstract version of the same underlying idea used in proving the results presented earlier in the paper.

Theorem 5. *Consider Algorithm 1. Let $\mathcal{A}$ be empirical risk minimization algorithm with respect to the hypothesis space $\mathcal{H}$ and assume that given $\epsilon, \delta > 0$, with $m \geq M(\mathcal{H}, \epsilon, \delta)$ samples, we have distribution free uniform convergence bound. I.e.:*

$$P\left[\sup_{w\in\mathcal{H}} \left|\mathbf{E}I(w \cdot xy \leq 0) - \frac{1}{m}\sum_{i=1}^m I(w \cdot x_i y_i \leq 0)\right| \leq \epsilon\right] \geq 1 - \delta. \quad (4)$$

Let $\delta \in (0,1)$ be the probability of failure. Assume that we ensure that at each stage k:

– *Choose margin threshold b_{k-1} such that with probability $1 - 0.5\delta/(k+k^2)$, $\exists \hat{w}_*$: $P((\hat{w}_{k-1} \cdot x)(\hat{w}_* \cdot x) \leq 0, |\hat{w}_{k-1} \cdot x| > b_{k-1}) \leq 2^{-(k+2)}$ and $P(\hat{w}_* \cdot xy \leq 0) \leq \inf_{w\in\mathcal{H}} \mathrm{err}(w) + 2^{-(k+2)}$.*
– *Take $m_k = \tilde{m}_k - \tilde{m}_{k-1} = M(\mathcal{H}, 2^{-(k+3)}, 0.5\delta/(k+k^2))$.*

Then after s iterations, $\mathrm{err}(\hat{w}_s) \leq \inf_{w\in\mathcal{H}} \mathrm{err}(w) + 2^{-s}$ with probability at least $1 - \delta$.

Proof Sketch: By the assumption on m_k, with probability $1 - \delta/(k+k^2)$, we have: $\mathrm{err}(\hat{w}_k) \leq P(\hat{w}_k \cdot xy \leq 0, x \in S_1) + P((\hat{w}_k \cdot x)(\hat{w}_* \cdot x) \leq 0, x \in S_2) + P(\hat{w}_* \cdot xy \leq 0, x \in S_2) \leq P(\hat{w}_k \cdot xy \leq 0, x \in S_1) + P((\hat{w}_k \cdot x)(\hat{w}_{k-1} \cdot x) \leq 0, x \in S_2) + P(\hat{w}_* \cdot xy \leq 0, x \in S_2) + 2^{-(k+2)} \leq P(\hat{w}_* \cdot xy \leq 0, x \in S_1) + P((\hat{w}_* \cdot x)(\hat{w}_{k-1} \cdot x) \leq 0, x \in S_2) + P(\hat{w}_* \cdot xy \leq 0, x \in S_2) + 2 \cdot 2^{-(k+2)} \leq \mathrm{err}(\hat{w}_*) + 3 \cdot 2^{-(k+2)} \leq \inf_{w\in\mathcal{H}} \mathrm{err}(w) + 2^{-k}$. □

In order to obtain a robust active learning algorithm that does not depend on the underlying data generation assumptions, one can estimate $M(\mathcal{H}, \epsilon, \delta)$ using sample complexity bounds. For example, we have used standard bounds such as Theorem 8 in earlier sections. A similar approach is taken in [2]. One can also replace (4) with a ratio uniform convergence bound such similar to the realizable case VC bound in Theorem 7. For some problems, this may lead to improvements.

In principle, it is also possible to estimate b_k using theoretical analysis. We only need to find b_k such that when $\hat{w}_k \cdot x > b_k$, no weight vector w can disagree with

$\hat{w}_k$ with probability more than $2^{-(k+3)}$ if $\mathrm{err}(w)$ is within 2^{-k} of the optimal value. However, the computation is more complicated, and requires that we know the underlying distribution of x. Note that in the theorems proved in earlier sections, we were able to estimate b_k because specific distributions of x were considered. Without such knowledge, practitioners often pick b_k by heuristics. Picking the right b_k is necessary for achieving good performance in our analysis. One practical solution is to perturb $\hat{w}_k$ (e.g. using bootstrap samples) and find b_k such that the perturbed vectors agrees with $\hat{w}_k$ with large probability when $\hat{w}_k \cdot x > b_k$. Another possibility is to use a procedure that tests for the best b_k. This is relatively easy to do for realizable problems, as shown in Figure 5. We can then prove that:

Theorem 6. *Consider Algorithm 5. Let $\mathcal{A}$ be the empirical risk minimization algorithm with respect to the hypothesis space $\mathcal{H}$, and assume that $\forall \epsilon, \delta > 0$, with $m \geq M(\mathcal{H}, \epsilon, \delta)$ samples we have distribution free uniform convergence bound: i.e. with probability $1-\delta$, $\forall w \in \mathcal{H}$, we have both*

$$\mathbf{E}I(w \cdot xy \leq 0) \leq \tfrac{2}{m}\textstyle\sum_{i=1}^m I(w \cdot x_i y_i \leq 0) + \epsilon \quad \text{and}$$

$$\tfrac{1}{m}\textstyle\sum_{i=1}^m I(w \cdot x_i y_i \leq 0) \leq 2\mathbf{E}I(w \cdot xy \leq 0) + \epsilon.$$

Let $N(\epsilon, \delta)$ be a distribution free convergence bound for the binary random variables $\xi \in \{0,1\}$: i. e. for $m \geq N(\epsilon, \delta)$ with probability $1-\delta$ we have both

$$\mathbf{E}\xi \leq \tfrac{1.5}{m}\textstyle\sum_{i=1}^m \xi_i + \epsilon \text{ and } \tfrac{1}{m}\textstyle\sum_{i=1}^m \xi_i \leq 1.5\mathbf{E}\xi + \epsilon.$$

Let $m_k = M(\mathcal{H}, 2^{-(k+5)}, 0.5\delta/(k+k^2))$, $n_k = N(2^{-(k+3)}, 0.25\delta/(\ell_k(k+k^2)))$, and $\epsilon_k = 2^{-(k+1)}$. Assume also we take b_{k,ℓ_k} s.t. $P(\hat{w}_{k-1} \cdot x \geq b_{k,\ell_k}) \leq 2^{-(k+5)}$.
If $\inf_{w \in \mathcal{H}} I(w \cdot xy \leq 0) = 0$, then after s iterations, with probability $1-\delta$, we have:

- *At each iteration $k \leq s$, before the for loop over q stops: $\forall \hat{w}_* \in \mathcal{H}$ such that $P(\hat{w}_* \cdot xy \leq 0) > 2^{-(k+6)}$: $P((\hat{w}_{k-1} \cdot x)(\hat{w}_* \cdot x) \leq 0, |\hat{w}_{k-1} \cdot x| > b_{k,q}) > 2^{-(k+6)}$.*
- *The final error is $\mathrm{err}(\hat{w}_s) \leq 2^{-s}$.*

We omit the proof here due to lack of space. Note that Theorem 6 implies that we only need to label a portion of data, with margins b_{k,q_k}, where q_k is the smallest q such that $\exists \hat{w}_* \in \mathcal{H}$ with $P(\hat{w}_* \cdot xy \leq 0) \leq 2^{-(k+6)}$ and $P((\hat{w}_{k-1} \cdot x)(\hat{w}_* \cdot x) \leq 0, |\hat{w}_{k-1} \cdot x| > b_{k,q}) \leq 2^{-(k+6)}$. It does not require us to estimate b_k as in earlier theorems. However, it requires an extra n_k labeled data at each iteration to select the optimal margin $b_{k,q}$. This penalty is usually small because the testing sample size n_k is often significantly smaller than m_k. For example, for d dimensional linear classifiers consider earlier, m_k needs to depend on d but n_k can be d-independent. Therefore it is possible to achieve significant improvement with this testing procedure. Its advantage is that we can choose b_k based on data, and thus the procedure can be applied to distributions that are not uniform.

Input: a learning algorithm $\mathcal{A}$ that learns a weight vector from labeled data
a sequence of training sample sizes $m_1, \ldots, m_s$;
a sequence of validation sample sizes $n_1, \ldots, n_s$ and one of acceptance thresholds $\epsilon_1, \ldots, \epsilon_s$
a sequence of cut-off points $\{-1 = b_{k,0} < b_{k,1} < \cdots < b_{k,\ell_k}\}$ $(k = 1, \ldots, s)$
Output: classifier $\hat{w}_s$
label data points $x_1, \ldots, x_{m_1}$ by a human expert and use $\mathcal{A}$ to learn weight vector $\hat{w}_1$.
iterate $k = 2, \ldots, s$
generate and label n_k samples $(x'_1, y'_1), \ldots, (x'_{n_k}, y'_{n_k})$
generate m_k samples x_j with labels $y_j = \text{sign}(\hat{w}_{k-1} \cdot x_j)$ $(j = 1, \ldots, m_k)$
for $q = 1, \ldots, \ell_k$
label y_j by a human expert if $|\hat{w}_{k-1} \cdot x_j| \in (b_{k,q-1}, b_{k,q}]$ $(j = 1, \ldots, m_k)$
use $\mathcal{A}$ to learn weight vector $\hat{w}_k$ from examples (x_j, y_j) $(j = 1, \ldots, m_k)$
if (error of $\hat{w}_k$ on (x'_j, y'_j) $(j = 1, \ldots, n_k)$ is less than ϵ_k) **break**
end for
end iterate

Fig. 5. Margin-based Active Learning with Testing

7 Discussion and Open Problems

While our procedure is computationally efficient in the realizable case, it remains an open problem to make it efficient in the general case. It is conceivable that for some special cases (e.g. the marginal distribution over the instance space is uniform, as in section 4) one could use the recent results of Kalai et. al. for Agnostically Learning Halfspaces [11]. In fact, it would be interesting to derive precise bounds for the more general of class of log-concave distributions.

Acknowledgements. We thank Alina Beygelzimer, Sanjoy Dasgupta, Adam Kalai, and John Langford for a number of useful discussions. Part of this work was done while the first author was visiting Yahoo! Research.

References

1. M. Anthony and P. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.
2. M.-F. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. In *ICML*, 2006.
3. M.-F. Balcan and A. Blum. A PAC-style model for learning from labeled and unlabeled data. In *Proceedings of the Annual Conference on Computational Learning Theory*, 2005.
4. R. M. Castro and R. D. Nowak. Upper and lower error bounds for active learning. In *The 44th Annual Allerton Conference on Communication, Control and Computing*, 2006.
5. O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.
6. D. Cohen, L. Atlas, and R. Ladner. Improving generalzation with active learning. 15(2):201–221, 1994.
7. S. Dasgupta. Analysis of a greedy active learning strategy. In *Advances in Neural Information Processing Systems*, 2004.
8. S. Dasgupta. Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems*, 2005.

9. S. Dasgupta, A. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. In *Proceedings of the Annual Conference on Computational Learning Theory*, 2005.
10. Y. Freund, H. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168, 1997.
11. A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. In *Proceedings of the 46th Annual Symposium on the Foundations of Computer Science*, 2005.
12. P. M. Long. On the sample complexity of PAC learning halfspaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556–1559, 1995.
13. S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 4:45–66, 2001.
14. A. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 2004.

A Useful Facts

We state here two standard Sample Complexity bounds [1] and a few useful probability bounds for standard normal variable.

Theorem 7. *Let H be a set of functions from X to $\{-1,1\}$ with finite VC-dimension $V \geq 1$. Let P be an arbitrary, but fixed probability distribution over $X \times \{-1,1\}$. For any $\epsilon, \delta > 0$, if we draw a sample from P of size $N(\epsilon,\delta) = \frac{1}{\epsilon}\left(4V\log\left(\frac{1}{\epsilon}\right) + 2\log\left(\frac{2}{\delta}\right)\right)$, then with probability $1-\delta$, all hypotheses with error $\geq \epsilon$ are inconsistent with the data.*

Theorem 8. *Let H be a set of functions from X to $\{-1,1\}$ with finite VC-dimension $V \geq 1$. Let P be an arbitrary, but fixed probability distribution over $X \times \{-1,1\}$. There exists a universal constant C, such that for any $\epsilon, \delta > 0$, if we draw a sample $((x_i, y_i))_i$ from P of size $N = N(\epsilon,\delta) = \frac{C}{\epsilon^2}\left(V + \log\left(\frac{1}{\delta}\right)\right)$, then with probability $1-\delta$, for all $h \in H$, we have $\left|\frac{1}{N}\sum_{i=1}^{N} I(h(x_i) \neq y_i) - \mathbf{E}_{(X,Y)} I(h(X) \neq Y)\right| \leq \epsilon$.*

Lemma 2. *Assume $x = [x_1, x_2] \sim N(0, I_{2\times 2})$, then any given $\gamma_1, \gamma_2 \geq 0$, we have $\Pr_x((x_1,x_2) \in [0,\gamma_1] \times [\gamma_2, 1]) \leq \frac{\gamma_1}{2\sqrt{2\pi}} e^{-\gamma_2^2/2}$.*

Lemma 3. *Assume $x = [x_1, x_2] \sim N(0, I_{2\times 2})$. For any given $\gamma, \beta > 0$, the following holds: $\Pr_x(x_1 \leq 0, x_1 + \beta x_2 \geq \gamma) \leq \frac{\beta}{2}\left(1 + \sqrt{-\ln\left[\min(1,\beta)\right]}\right) e^{-\gamma^2/(2\beta^2)}$.*

B Probability estimation in high dimensional ball

Consider $x = [x_1, \ldots, x_d] \sim P_x$ uniformly distributed on unit ball in R^d. Let A be an arbitrary set in R^2; we are interested in estimating the probability $\Pr_x((x_1,x_2) \in A)$. Let V_d be the volume of d-dimensional ball; we know $V_d = \pi^{d/2}/\Gamma(1+d/2)$ where Γ is the Gamma-function. In particular $V_{d-2}/V_d = d/(2\pi)$. It follows:

$$\Pr_x((x_1,x_2) \in A) = \frac{V_{d-2}}{V_d} \int_{(x_1,x_2)\in A} (1 - x_1^2 - x_2^2)^{(d-2)/2} dx_1 dx_2 =$$

$$\frac{d}{2\pi}\int_{(x_1,x_2)\in A} (1 - x_1^2 - x_2^2)^{(d-2)/2} dx_1 dx_2 \leq \frac{d}{2\pi}\int_{(x_1,x_2)\in A} e^{-(d-2)(x_1^2+x_2^2)/2} dx_1 dx_2.$$

where we use the inequality $(1-z) \leq e^{-z}$.

Lemma 4. *Let $d \geq 2$ and let $x = [x_1, \ldots, x_d]$ be uniformly distributed in the d-dimensional unit ball. Given $\gamma_1 \in [0,1]$, $\gamma_2 \in [0,1]$, we have:*

$$\Pr_x((x_1, x_2) \in [0, \gamma_1] \times [\gamma_2, 1]) \leq \tfrac{\gamma_1\sqrt{d}}{2\sqrt{\pi}} e^{-(d-2)\gamma_2^2/2}.$$

Proof. Let $A = [0, \gamma_1] \times [\gamma_2, 1]$. We have
$\Pr_x((x_1, x_2) \in A) \leq \frac{d}{2\pi} \int\limits_{(x_1,x_2)\in A} e^{-(d-2)(x_1^2+x_2^2)/2} dx_1 dx_2 \leq \frac{\gamma_1 d}{2\pi} \int\limits_{x_2 \in [\gamma_2, 1]} e^{-(d-2)x_2^2/2} dx_2$
$\leq \frac{\gamma_1 d}{2\pi} e^{-(d-2)\gamma_2^2/2} \int\limits_{x \in [0, 1-\gamma_2)} e^{-(d-2)x^2/2} dx \leq \frac{\gamma_1 d}{2\pi} e^{-(d-2)\gamma_2^2/2} \min\left[1 - \gamma_2, \sqrt{\tfrac{\pi}{2(d-2)}}\right].$
Note that when $d \geq 2$, $\min(1, \sqrt{\pi/(2(d-2))}) \leq \sqrt{\pi/d}$. □

Lemma 5. *Assume $x = [x_1, \ldots, x_d]$ is uniformly distributed in the d-dimensional unit ball. Given $\gamma_1 \in [0,1]$, we have $\Pr_x(x_1 \geq \gamma_1) \leq \frac{1}{2} e^{-d\gamma_1^2/2}$.*

Proof. Let $A = [\gamma_1, 1] \times [-1, 1]$. Using a polar coordinate transform, we have:
$\Pr_x((x_1, x_2) \in A) = \frac{d}{2\pi} \int_{(x_1,x_2)\in A} (1 - x_1^2 - x_2^2)^{(d-2)/2} dx_1 dx_2 =$
$\frac{d}{2\pi} \int_{(r, r\cos\theta) \in [0,1] \times [\gamma_1, 1]} (1 - r^2)^{\frac{d-2}{2}} r dr d\theta = \frac{1}{2\pi} \int_{(r, r\cos\theta) \in [0,1] \times [\gamma_1, 1]} d\theta d(1 - r^2)^{\frac{d}{2}}$
$\leq \frac{1}{2\pi} \int_{(r,\theta) \in [\gamma_1, 1] \times [-\pi/2, \pi/2]} d\theta d(1 - r^2)^{d/2} = 0.5(1 - \gamma_1^2)^{d/2} \leq \frac{1}{2} e^{-d\gamma_1^2/2}.$ □

Lemma 6. *Let $d \geq 4$ and let $x = [x_1, \ldots, x_d]$ be uniformly distributed in the d-dimensional unit ball. Given $\gamma, \beta > 0$, we have:*

$$\Pr_x(x_1 \leq 0, x_1 + \beta x_2 \geq \gamma) \leq \tfrac{\beta}{2}(1 + \sqrt{-\ln \min(1, \beta)}) e^{-d\gamma^2/(4\beta^2)}.$$

Proof. Let $\alpha = \beta\sqrt{-2d^{-1} \ln \min(1, \beta)}$, we have:
$\Pr_x(x_1 \leq 0, x_1 + \beta x_2 \geq \gamma)$
$\leq \Pr_x(x_1 \leq -\alpha, x_1 + \beta x_2 \geq \gamma) + \Pr_x(x_1 \in [-\alpha, 0], x_1 + \beta x_2 \geq \gamma)$
$\leq \Pr_x(x_1 \leq -\alpha, x_2 \geq (\alpha + \gamma)/\beta) + \Pr_x(x_1 \in [-\alpha, 0], x_2 \geq \gamma/\beta)$
$\leq \frac{1}{2} \Pr_x(x_2 \geq (\alpha + \gamma)/\beta) + \Pr_x(x_1 \in [0, \alpha], x_2 \geq \gamma/\beta)$
$\leq \frac{1}{4} e^{-d(\alpha+\gamma)^2/(2\beta^2)} + \frac{\alpha\sqrt{d}}{2\sqrt{\pi}} e^{-d\gamma^2/(4\beta^2)}$
$\leq \left[\frac{1}{4} e^{-\frac{d\alpha^2}{2\beta^2}} + \frac{\alpha\sqrt{d}}{2\sqrt{\pi}}\right] e^{-\frac{d\gamma^2}{4\beta^2}} = \left[\frac{\min(1,\beta)}{4} + \frac{\beta\sqrt{-2\ln\min(1,\beta)}}{2\sqrt{\pi}}\right] e^{-\frac{d\gamma^2}{4\beta^2}}.$ □

Lemma 7. *Let u and w be two unit vectors in R^d, and assume that $\theta(u, w) \leq \tilde{\beta} < \pi/2$. Let $d \geq 4$ and let $x = [x_1, \ldots, x_d]$ be uniformly distributed in the d-dimensional unit ball. Consider $C > 0$, let $\gamma = \frac{2\sin\tilde{\beta}}{\sqrt{d}} \sqrt{\ln C + \ln\left(1 + \sqrt{\ln \max(1, \cos\tilde{\beta}/\sin\tilde{\beta})}\right)}$. Then $\Pr_x\left[(u \cdot x)(w \cdot x) < 0, |w \cdot x| \geq \gamma\right] \leq \frac{\sin\tilde{\beta}}{C\cos\tilde{\beta}}$.*

Proof. We rewrite the desired probability as $2\Pr_x\left[w \cdot x \geq \gamma, u \cdot x < 0\right]$. W.l.g., let $u = (1, 0, 0, ..., 0)$ and $w = (\cos(\theta), \sin(\theta), 0, 0, ..., 0)$. For $x = [x_1, x_2, ..., x_d]$ we have $u \cdot x = x_1$ and $w \cdot x = \cos(\theta)x_1 + \sin(\theta)x_2$. Using this representation and Lemma 6, we obtain $\Pr_x\left[w \cdot x \geq \gamma, u \cdot x < 0\right] = \Pr_x[\cos(\theta)x_1 + \sin(\theta)x_2 \geq \gamma, x_1 < 0] \leq \Pr_x\left[x_1 + \frac{\sin(\tilde{\beta})}{\cos(\tilde{\beta})} x_2 \geq \frac{\gamma}{\cos(\tilde{\beta})}, x_1 < 0\right] \leq \frac{\sin\tilde{\beta}}{2\cos\tilde{\beta}} \left(1 + \sqrt{\ln\max(1, \frac{\cos\tilde{\beta}}{\sin\tilde{\beta}})}\right) e^{-\frac{d\gamma^2}{4\sin^2\tilde{\beta}}} = \frac{\sin\tilde{\beta}}{2\cos\tilde{\beta}} C^{-1}$, as desired. □