Convex Optimization for Active Learning with Large Margins

Eric J. Friedman[*]
School of Operations Research and Information Engineering
Cornell University
Ithaca, NY 14850
ejf27@cornell.edu

Abstract

In this paper we show how large margin assumptions make it possible to use ideas and algorithms from convex optimization for active learning. This provides an alternative and complementary approach to standard algorithms for active learning. These algorithms appear to be robust and provide approximately correct hypotheses with probability one, as opposed to the standard PAC learning results.

In particular we consider the problem of finding global convergence bounds for active learning with halfspaces. We show that a large margin assumption allows the reduction of active learning problem to that of convex optimization from which one can construct efficient algorithms. This work generalizes and clarifies previous results in this area and provides new insights.

1 Introduction

Large margin assumptions are fundamental in learning theory and practice, in particular in our understanding of Support Vector Machines [Vap00]. In this paper we show how large margin analysis can be applied to active learning. Our main insight is that large margin assumptions allow one to apply tools from convex optimization to develop active learning algorithms which exponentially improve on unsupervised learning algorithms. Our analysis is global and not asymptotic in contrast with recent results [BHW08].

Active learning is important because in many situations one has access to many examples which are costly to label. A well known example is that of spam detection. Unlabeled emails are widely available, but labeled emails require human intervention, so are costly. The goal of active learning algorithms in this setting is to learn to identify spam with a minimal number of labeling requests. Additional applications of active learning arise in web searching and also in biology, where labeling may involve *in vitro* experiments which are expensive and time consuming [DZW$^+$07].

The initial results for active learning were encouraging. For example, when instances are uniformly distributed on the unit sphere, simple algorithms can actively learn separating hyperplanes through the origin with error less than ϵ with only $O(d\log(d/\epsilon))$ label queries, where d is the dimensionality of the space. This is an exponential improvement over the number of labels requested by unsupervised learning which are $\Omega(d/\epsilon)$. In addition, in [BBZ07] Balcan et. al. show that under a large margin assumption (and other mild restrictions) one can remove the dependence on dimension and present an active learning algorithm that requires only $O(log(1/\epsilon))$ labels.

However, when the separating hyperplanes are not required to pass through the origin, the (non-asymptotic) complexity of active learning is $\Omega(d/\epsilon)$ which is the same as that for unsupervised learning (see, e.g., [Das05]). However, in a recent paper [BHW08] Balcan et. al. have shown that asymptotically active learning (under mild assumptions) improves on unsupervised learning and except for some unrealistic models the improvement is exponential. Formally, they show that the number of labels needed is $C(h)d\log(1/\epsilon)$ where $C(h)$ depends on the target hypothesis h.

In this paper we will show that, under an additional assumption, one can learn the generalized separating hyperplane with $Cd\log(1/(A\epsilon))$ labels which does not depend directly on the target hypothesis but only on its margin A. Thus, for large margins one gets non-asymptotic results and in addition, these results can be obtained for many interesting generalizations.

Our analysis relies on an important fact that is of independent interest. Large margins create a tight relationship between the distance of hypotheses in terms of errors and the geometric distances between them in the hypothesis space. In particular, if two hypotheses are geometrically close then they must be close in terms of errors. This allows one to transform the analysis from the standard combinatorial probabilistic approach to that of convex optimization, a topic which has been well studied. Thus, one can rigorously apply the ideas from convex optimization to active learning. This provides a complementary approach which may prove fruitful.

One somewhat surprising outcome of this approach is that our results are truly approximately correct not just probably approximately correct, which is the standard for analyses in this area.[1]

[*] http://www.people.cornell.edu/pages/ejf27/

[1] Although this may not seem possible for randomized samples the key is that the total number of required samples, which are not labeled, can be arbitrarily large. This raises interesting definitional questions which we will not pursue in this paper.

The paper is organized as follows: in the following section we present the abstract version of our analysis. Then, in Section 3 we apply our theory to some well known examples and conclude in Section 4.

2 General Theory

Let $X = B^d$, where B^d is the d dimensional unit ball in $\Re^d$, be the instance space where instances can have 2 labels $\{-,+\}$ which we will refer to as negative or positive instances. Assume there is some distribution D on X. The hypotheses will be halfspaces in $\Re^d$ which we will represent by $w \in \Re^d$, using a nonstandard, but useful, representation. Given some $w \in \Re^d$ we define the halfspace as $\hat{h}(w) = \{x \in X \mid w \cdot x \leq 1\}$. Note that this assumes that $x = 0$ is a positive instance. We will assume this throughout without loss of generality. (If this isn't true just swap the definition of the labels.) Let W be the set of allowed halfspace parameters and $\hat{H}$ the set of feasible halfspaces.

First we note that it makes sense to restrict to the case where $||w|| \geq 1$, since if $||w|| < 1$ then the separating hyperplane generated by w does not intersect x so $h(w)$ can be removed from $\hat{H}$. This is because the halfspace generated by $w/||w||$ (or any other w with $||w|| = 1$) will classify all points in the same way as the halfspace generated by w so there is no loss of generality by removing all w's with norm less than 1.

Now we define the complexity of actively learning halfspaces. Note that this definition assumes that the algorithm learns an approximately optimal halfspace with probability 1.

Definition 1 *Given the problem of actively learning a halfspace given instance space X and distribution D, the problem has sure sample complexity $SSC(\epsilon)$ if there exists an active learning algorithm which can find a halfspace with error at most ϵ with labeling at most $SSC(\epsilon)$ instances.*

Now we define the margin of a halfspace as follows.

Definition 2 *A halfspace $\hat{h}(w)$ satisfies the margin condition with margin constant $\hat{A}$ if*

$$Pr[|w \cdot x - 1| \leq \gamma] \leq \gamma / \hat{A}$$

for all $\gamma > 0$. The margin constant of a halfspace is the infimum of $\hat{A}$ over all satisfying margins.

Note that this definition of margin is somewhat nonstandard and is not the same as the geometric margin. In Section 3 we will discuss the relationship between the two.

2.1 Large Margins

Define the error of a hypothesis $h(w')$ from a true hypothesis $h(w)$ to be

$$err(h(w'); h(w)) = Pr_X[h(w) \Delta h(w')]$$

where Δ represents the symmetric difference, i.e.,

$$h(w) \Delta h(w') = h(w) \setminus h(w') \bigcup h(w') \setminus h(w).$$

Note that $h(w) \Delta h(w')$ is the disagreement region, $DIS(\{h(w), h(w')\})$.

In general, for some $\hat{W} \subset W$ define the disagreement region for $\hat{W}$ to be the union of all the disagreement regions between pairs of w's. Thus,

$$DIS(\hat{W}) = \bigcup_{w, w' \in \hat{W}} h(w) \Delta h(w')$$

which is the set of instances which are uncertain given a set of hypotheses $\hat{W}$. Now we show how margins connect errors in parameter space with classification errors.

Theorem 3 *If $w \in W$ has margin constant $\hat{A}$ then $err(h(w'); h(w)) \leq ||w - w'|| / \hat{A}$.*

Proof: Note that the errors can only occur in the disagreement region $DIS(\{h(w), h(w')\})$, thus we need to bound $P_X[DIS(\{h(w), h(w')\})]$. Now, consider the problem of maximizing the margin over the disagreement region. It is easy to see that the maximum must be attained at the boundary which implies that the maximum margin is $|w \cdot x - 1|$ where $w' \cdot x = 1$. Thus

$$|w \cdot x - 1| = |(w - w') \cdot x + w' \cdot x - 1|$$

and by applying the constraint we see that

$$|w \cdot x - 1| = |(w - w') \cdot x| \leq ||w - w'|| \, ||x|| \leq ||w - w'||$$

since $||x|| \leq 1$ by assumption. Then, by the margin assumption, the total probability of instances with this margin is less than $||w - w'||/A$ completing the proof. QED

The importance of this result is that it reduces a probabilistic problem into a geometric one. If we can find a hypothesis which is guaranteed to be geometrically close to the true hypothesis then, under a margin assumption, it will have small error. Note that this statement is true with certainty, not just with high probability.

2.2 Applying Convex Optimization

Our (somewhat nonstandard) choice of representation for the halfspaces turns out to be important as it can be easily embedded into a convex space. Let W_t be the set of feasible halfspace parameters after t instances have been labeled, e.g. if $w \in W_t$ then $h(w)$ classifies the first t labeled points correctly. Now suppose that the next point x_{t+1} is labeled positively. (To reduce notation we will index the points by the time at which they are labeled.) If a hypothesis $w \in W$ is consistent with this point it must satisfy $w \cdot x_{t+1} \leq 1$ which is a halfspace in the *parameter space*. Thus $W_{t+1} = W_t \bigcup f_+(x)$ where $f_+(x) = \{w \in W \mid w \cdot x \leq 1\}$. Note that if this point were a negative instance then we would just use the complement of $f_+(x)$, denoted $f_-(x)$, in the update.[2] Thus, a newly labeled point generates a separating hyperplane for the feasible region.

[2] Note that $f_+(x)$ contains the separating hyperplane while f_- does not. This leads to bookkeeping issues which don't significantly affect our analysis.

This connection can be made precise by noting that we can think of an active learning algorithm as approximately choosing points $x \in X$ with approximate knowledge of the distribution D.

Definition 4 *A point $x \in X$ is in the smooth support of D if any neighborhood of $x \in X$ has nonzero probability.*

Note that if a point has smooth support then if the algorithm waits long enough it can always find a point arbitrarily close to x to label. In the following, we will assume that we can find the exact point. As will be clear, our algorithms are robust to small errors in the chosen point so this will not affect our results.

Combining these ideas with the large margin result we can reduce the problem of actively learning a halfspace to that of sequentially choosing a sequence of points $x_1, \dots, x_t$ in the smooth support of D to label, such that the feasible region after labeling these points W_t has small geometric diameter. Since each of these points generates a separating hyperplane we can use this to reduce the size of the feasible region until it is sufficiently small that we can estimate the true w with sufficient accuracy.

The precise method we use depends on our assumptions about the set of smooth support of D. We begin with the simplest version which is when all of X is contained in the smooth support of D. In this case, the following simple bisection algorithm suffices.

Active Bisection Algorithm

```
1. Set $W_0 = W$ and $t = 0$.
2. For t=1 to T
   (a) Let $i = t \pmod{d}$
   (b) Compute
       $\bar{b}_i = (\max\{w_i |\ w \in W_t\} + \min\{w_i |\ w \in W_t\})/2.$
   (c) Label $x = \bar{b}_i e_i$, where $e_i$ is the $i$'th unit vector and let $s \in \{-,+\}$ be the label.
   (d) Set $W_t = W_{t-1} \bigcap f_s(x)$.
3. Return $w^*$ such that $w^*_i = (\max\{w_i \mid w \in W_t\} + \min\{w_i \mid w \in W_t\})/2$.
```

Since this is essentially a bisection algorithm it converges rapidly and allows us to actively learn efficiently.

Theorem 5 *Assume that W is convex, has diameter ϕ, and all of X in $DIS(W)$ is contained in the smooth support of D. Then its sure sample complexity is $O(d \log(d\phi/(\hat{A}\epsilon)))$.*

Proof: Apply the Active Bisection Algorithm. As we show below, after $O(d \log(d\phi/(\hat{A}\epsilon)))$ iterations it will find a w with error at most ϵ.

Consider the set of t such that $i = t \pmod{i}$. At each such iteration of the algorithm the size of the feasible region in the i'th direction is cut in half. Since the diameter of W_0 is less than ϕ it is guaranteed to be less than $2^{-T/d}\phi$ by the end of the algorithm. Thus the diameter of W_T is less than $2^{-T/d}\sqrt{d}\phi$. Applying Theorem 1 completes the analysis. QED

Note that in the case where W has significantly different extent in different directions one could strengthen this result. For example if W is large in k directions but small, $O(\epsilon)$, in the others, then the number of labels depends linearly on k not d.

As we will see in the following section, this method will allow us, with a mild assumption to solve problems where D has smooth support on a convex subset of X; however, without such full support, as in the well studied example where the smooth support is on the surface of x, one needs a more sophisticated approach.

Our analysis will apply some ideas from convex optimization, known as cutting plane methods. These are generalized bisection algorithms. The best known of which is the Ellipsoid method which was used to show the polynomiality of linear programming; however, the Ellipsoid method is theoretically suboptimal, so we will develop an algorithm modeled on the center of gravity method.[3]

Given a set $M \in \Re^d$ let $Vol(M)$ be the Euclidean volume of H. It us useful to define the center of gravity of M as

$$CG(M) = Vol(M)^{-1} \int_M x dx.$$

Now we recall Grunbaum's theorem [Grü60] in our notation.

Theorem 6 (Grunbaum) *Let $M \in \Re^d$ be a convex set and $h(w)$ a halfspace such that $w \cdot CG(M) = 1$. Then*

$$Vol(M \bigcap h(w)) \leq e^{-1} Vol(M).$$

Thus, by a judicious choice of w we can always geometrically reduce the volume of the feasible set. Applying this technique we can reduce the assumptions required to obtain a good active learning algorithm. However, there are two main issues to resolve. The first is that we need to be able to find at least one separating hyperplane and the second is that reducing the volume of W_t is not sufficient as we must also guarantee that it has small diameter. In order to achieve both of these goals we need the following definition.

Definition 7 *A hyperplane $w \in W$ is fully covered by D if there exist d linearly independent instances $x \in X$ in the smooth support of X such that $w \cdot x = 1$.*

To construct the algorithm it will be useful to develop some notation. Given a set of vectors $V_i \in \Re^d$ for $i \in \{1, 2, \dots, k\}$ define the linear space

$$LIN(V) = \{v' \in \Re^d \mid \forall\, i,\ V_i \cdot v' = 0\}$$

and given an additional set of scalars $\psi_i \in \Re$ for $i \in \{1, 2, \dots, k\}$ define the affine space

$$AFF(V, \psi) = \{v' \in \Re^d \mid \forall\, i,\ V_i \cdot v' = \psi_i\}.$$

[3] These methods are all discussed in the lectures by Nemirovski [Nem94].

Given a vector $y \in \Re^d$ we say that $y \perp V$ if y is orthogonal to all the vectors in V and $y \not\perp V$ if that is not true. It is useful to note that if $y \not\perp V$ then the projection of y onto $LIN(V)$ is nonzero.

In addition, we need to consider the smallest "slab" containing a set. Recall that a slab is the region between 2 parallel hyperplanes. We denote a slab as

$$SL(\gamma, \psi, \omega) = \{v \in \Re^d \mid v \cdot \gamma \in [\psi - \omega, \psi + \omega]\}$$

where $v \in \Re^d$ is the defining unit vector, $\psi \in \Re$ is the center and ω is the width. Given a set $W \in \Re^d$ we will define the minimal slab $SL_m(W)$ to the slab with the smallest width ω containing W. The parameters of this slab will be denoted by $\gamma(W)$, $\psi(W)$, and $\omega(W)$.

Lastly, we require a projection operator to an affine space $\pi^{AFF(V,\gamma)}$ which takes a point v to the closest point in $AFF(V, \gamma)$.

Active Center of Gravity Algorithm

```
1. Set $W^d = W$, $V = \emptyset$ and $\psi = \emptyset$.

2. For $\hat{d} = d$ to 1

   (a) Compute $\hat{W}^{\hat{d}} = CGReduce(W^{\hat{d}}, V, \psi)$.
   (b) Set $\gamma^{\hat{d}} = \gamma(\hat{W}^{\hat{d}})$ and $\psi^{\hat{d}} = \psi(\hat{W}^{\hat{d}})$.
   (c) Set $W^{\hat{d}-1} = \pi^{AFF(V,\psi)}(\hat{W}^{\hat{d}})$.
   (d) Set $V = V \bigcup \gamma^{\hat{d}}$ and $\psi = \psi \bigcup \psi^{\hat{d}}$.

3. Return the $w^*$ which solves $\gamma^{\hat{d}} = \psi^{\hat{d}}$ for
   all $\hat{d} \in \{1, 2, \ldots, d\}$.
```

CGReduce

```
1. Input $W^{\hat{d}}, V$ and $\psi$.

2. While $\phi(\pi^{AFF(V,\psi)}(W^{\hat{d}})) \geq 2\epsilon/(\hat{A}d^{1/2})$ do

   (a) Let $w = CG(\pi^{AFF(V,\psi)}(W^{\hat{d}}))$.
   (b) Choose some $x$ in the smooth support
       of $X$ such that $w \cdot x = 1$ and $V \not\perp x$.
   (c) Set $s \in \{-.+\}$ to the label of $x$.
   (d) Set $W^{\hat{d}} = W^{\hat{d}} \bigcap f_s(x)$.

3. Return $W^{\hat{d}}$.
```

Now we show that this algorithm converges quickly and provides an upper bound for the sure sample complexity.

Theorem 8 *Assume that $X = B^d$, W is convex, has diameter $\phi(W)$, and all of $w \in W$ is fully covered. Then its sure sample complexity is $O(d \log(d\phi/(A\epsilon)))$.*

Proof: Apply the Active Center of Gravity Algorithm. As we show below, after $O(d \log(d\phi/(\hat{A}\epsilon)))$ iterations it will find a w with error at most ϵ.

First, by Grunbaum's Theorem it is clear that the Center of Gravity Method will reduce the volume of $W^{\hat{d}}$ in CGReduce and clearly at some point the smallest slab will be sufficiently small. In addition, at the termination all slabs are less than $2\epsilon/(\hat{A}d^{1/2})$ thick so the solution is accurate to $\epsilon/(\hat{A}d^{1/2})$ in d orthonormal coordinates which by Theorem 3 guarantees the accuracy.

Next, since every $w \in W$ is fully covered it is always possible to find a valid x in step 2b of CGReduce. Since there are d linearly independent x's in the smooth support of D at w at least one of them must be non-orthogonal to any non-empty linear subspace of $\Re^d$.

Lastly, we need to check the number of iterations. Let $n_{\hat{d}}$ be the number of labeled point during the stage $\hat{d}$ call to CGReduce and let $\phi_{\hat{d}}$ be the volume of $W^{\hat{d}}$ in the main algorithm. Note that due to the projection step $\phi^{\hat{d}}$ is the volume of the projection which is less than α times the unprojected volume, where $\alpha = 2\epsilon/(\hat{A}d^{1/2})$. Thus by Grunbaum's theorem $\phi_{\hat{d}-1} \leq e^{-n_{\hat{d}}}/\alpha$. Iterating this bound yields $1 \leq e^{-n}/\alpha^d$, where $n = \sum_{\hat{d}=1}^{d} n_{\hat{d}}$, which completes the proof. QED

3 Applications

In this section we apply our results from the previous section to generalized versions of our motivating examples: learning halfspaces when the instances are uniformly distributed on surface of the unit ball. In this case we use the standard representation of halfspaces, $h(v, v_0) = \{x \in X \mid v \cdot x - v_0 \leq 0\}$, where $v \in \Re^d$ satisfies $||v|| = 1$, which we denote by $v \in V$, and the geometric margin defined as follows:

Definition 9 *A halfspace $h(v, v_0)$ satisfies the geometric margin condition with geometric margin A if*

$$Pr[|v \cdot x - v_0| \leq \gamma] \leq \gamma/A$$

for all $\gamma > 0$. The geometric margin constant of a halfspace is the infimum of A over all satisfying geometric margins.

First we note the simple relationship between the two margin constants.

Lemma 10 *Suppose that $\hat{h}(w)$ has margin constant $\hat{A}$. Then the geometric margin constant A of $\hat{h}(w)$ satisfies $A = \hat{A}||w||$.*

Proof: Note that $\hat{h}(w) = h(w/||w||, 1/||w||)$. Then

$$Pr[|w \cdot x - 1| \leq \gamma] = Pr[|(w/||w||) \cdot x - 1/(||w||)| \leq \gamma/||w||]$$

which implies that if

$$Pr[|w \cdot x - 1| \leq \gamma] \leq \gamma/\hat{A}$$

then

$$Pr[|w/||w|| \cdot x - 1/||w||| \leq \gamma] \leq \gamma/(||w||\hat{A})$$

which proves the lemma as $||w||\hat{A} = A$. QED

Thus, in order to get accuracy ϵ under the geometric margin we need to solve the problem to accuracy $\hat{\epsilon} = \epsilon/||w_m||$ in our representation where $w_m \in W$ maximizes $||w||$ over W.

Next we recall that the set of "relevant" halfspaces for instances contained in (or on) the unit ball is the set of $w \in$

$W \subset \Re^d$ such that $||w|| \geq 1$. In order to apply our techniques we need to restrict to a bounded set of parameters, so we define $W^\sigma = \{w \in \Re^d \mid 1 \leq ||w|| \leq \sigma\}$ which introduces small additional error under the large margin assumption.

Lemma 11 *Suppose that $w \in W$ has margin constant $\hat{A}$ and $w \notin W^\sigma$. Then there exists some $\hat{w} \in W^\sigma$ such that $err(\hat{w}; w) \leq 1/(A\sigma)$.*

Proof: Consider $\hat{w} = \sigma w/||w|| \in W^\sigma$ and consider the standard representation of the halfspaces, $h(w/||w||, 1/||w||)$ and $h(w/||w||, 1/\sigma)$. Then we use a similar argument to that of Theorem 3, except in this case the maximum margin between the two is simply $|1/||w|| - 1/\sigma| < 1/\sigma$ which proves the result. QED

Thus, bounding W^σ only increases the error by a small amount if σ is chosen to be large; however, the feasible region is not convex. To remedy this we make what seems to be a mild, yet powerful assumption.

Assumption 1 *The algorithm begins with a negative instance x_-. Recall that by assumption the origin is a positive instance, so this is equivalent to assuming that the algorithm is always given one instance of each type.*

Note that this is very strong in the sense that in $d = 2$ it simplifies the problem dramatically, as the difficult part is to find two such examples. However, we believe that in any reasonable learning problem one would always have examples of both types of instances and thus difficulties that arise from finding such instances should only be of theoretical interest. Note also that two points is sufficient for any $d > 0$ so this does not grow with dimension or desired accuracy. Alternatively, an equivalent assumption is that the measure of the positive examples must be bounded away from both 0 and 1, another seemingly reasonable assumption.

Given this assumption, we can consider the set $\hat{W} = W^\sigma \bigcap f_-(x_-)$ which is convex and bounded with radius less than σ. Then we can apply our results from the previous section to prove complexity results.

We begin with the simpler result which shows that one can use margin analysis for active learning on convex instance spaces with full support.

Theorem 12 *Assume that $X = B^d$ and that all of X is in the smooth support of D. Then its sure sample complexity is $O(d \log(d\phi/(A\epsilon)))$, where A is the geometric margin.*

Proof: This follows from Theorem 5. First we note that we can run the active bisection algorithm starting with $\hat{W} = W^\sigma \bigcap f_-(x_-)$ for $\sigma = A\epsilon/2$. This guarantees that the error from using W^σ instead of W will have error at most $\epsilon/2$ by Lemma 11. We run the bisection algorithm to accuracy $\hat{A}\epsilon/(2\sigma)$ which requires a running time of the correct order and guarantees, via Lemma 10 that the error will be less than $\epsilon/2$. This completes the proof. QED

Similarly, we can apply the Active Center of Gravity method to any instance space with full support on the surface of a convex set. This includes the obvious generalization of the case when the distribution is uniform on the surface of the hypersphere.

Theorem 13 *Assume that $X = B^d$ and that all of the boundary of X is in the smooth support of D. Then its sure sample complexity is $O(d \log(d\phi/(A\epsilon)))$, where A is the geometric margin.*

Proof: This proof is similar to that in the previous theorem except we use the Active Center of Gravity Algorithm. The only change is that we need to show that every w is fully covered by D. To do this, note that the points in X which cover w are precisely those which intersect the hyperplane $w \cdot x = 1$. Now we assume that $w = \lambda e_1$ where e_1 is the first unit vector and $0 < \lambda < 1$. In this case, the set of x such that $w \cdot x = 1$ is given by the vectors of the form (λ, v) where $v \in \Re^{d-1}$ and v satisfies $\sum_{i=2}^d v_i^2 = \sqrt(1 - \lambda)$. Now we choose d linearly independent vectors x^i as follows. For $2 \leq i \leq d$ let $x^i = \lambda e_1 + \sqrt(1 - \lambda)e_i$ and $x^1 = \lambda e_1 - \sqrt(1 - \lambda)e_2$. Now, it is easy to see that for $2 \leq i \leq d$ the x^i are linearly independent as they contain different unit vectors. It remains to note that x^1 and x^2 are linearly independent by inspection, completing the proof. QED

3.1 Some algorithmic comments

Note that if we consider computational complexity then the Active Center of Gravity Algorithm is not efficiently implementable as it is difficult to compute the center of gravity of a polytope, although there is some interesting computational work on its approximation [DFK91, BV02].

Instead, one could use the ellipsoid method [**?**] instead of the center of gravity method in the algorithm. Unfortunately, the ellipsoid method only guarantees that it will reduce the volume by $1 - 1/d$ per separating hyperplane, instead of the constant e^{-1} of the center of gravity method. This adds another factor of d to the number of instances the algorithm must label changing the complexities to $O(d^2 \log(d/(A\epsilon)))$ in Theorems 8 and 13. This is still essentially an exponential improvement over non-active learning.

However, the ellipsoid method is known to be inefficient in practice for solving linear programs, so one might be hesitant to apply it to active learning. We see two possible responses to this argument. The first is that perhaps the ellipsoid method isn't that inefficient, it's just that its competitors, interior point methods and simplex methods are amazingly efficient at solving linear programs, as they perform dramatically better in practice then their formal analysis would suggest.

Nonetheless, it would be interesting to try to apply ideas that have been developed for linear programming (and combinatorial optimization) to develop practical algorithms for active learning. The most promising ideas appear to be column generation and Dantzig-Wolfe decomposition, which have been shown to be extremely effective in practice.[4] We leave a detailed analysis of these issues to future work.

[4] See, e.g., the textbook by Schrijver [Sch00] for these and other potentially relevant methods.

4 Conclusions

We have shown how large margin conditions, which are fundamental in many areas of learning theory, guarantee an exponential improvement of active learning for halfspaces over ordinary supervised learning. They also, provide a unified analysis of many problems. In addition our analysis shows tantalizing connections between active learning and convex (and combinatorial) optimization. One obvious direction for future research would be the development of active learning algorithms that extend these connections. For example, one might like to extend the ideas of column generation of Dantzig-Wolf decomposition to active learning.

Also, given that our methods appear to be robust it seems natural that they should be extendable to the agnostic case. In the agnostic setting, optimization frameworks are quite natural. Essentially this would correspond to adding an objective function to our analysis.

Another set of open problems come from generalizing our analysis. Can one further reduce the requirements on the smooth support of the instance distribution? For example, we believe that one can remove a convex set from the smooth support and not change our main results (Theorems 5 and 8).

A fourth area for continued analysis is the extension of margin methods to nonlinear hypotheses. In this case one can show that asymptotically the analysis becomes essentially linear, but for non-asymptotic analysis the curvature of hypotheses leads to complications.

Lastly, we view this work as a step towards understanding active learning for kernel SVMs. One main obstacle to applying our analysis to kernel SVMs is that the problem is only linear in a high dimensional space. However, Balcan et. al. [BBZ07] have shown that under a large margin assumption it may be possible to actively learn with sample complexities that are dimension independent. We expect that this approach may allow one to bound the sure sample complexity of active learning with kernel SVMs in a dimension independent manner, for large margins.

Acknowledgments

This work has been supported in part by NSF under grants ITR-0325453 and CDI-0835706.

References

[BBZ07] M.F. Balcan, A. Broder, and T. Zhang. Margin based active learning. *Proceedings of the Twentieth Annual Conference on Learning Theory (COLT 2008)*, 2007.

[BHW08] M. Balcan, S. Hanneke, and J. Wortman. The true sample complexity of active learning. Forthcoming in the Proceedings of the Twenty-First Annual Conference on Learning Theory (COLT 2008), 2008.

[BV02] Dimitris Bertsimas and Santosh Vempala. Solving convex programs by random walks. In *STOC '02: Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 109–115, New York, NY, USA, 2002. ACM.

[Das05] S. Dasgupta. Coarse sample complexity bounds for active learning. *Advances in Neural Information Processing Systems*, 18:2, 2005.

[DFK91] Martin Dyer, Alan Frieze, and Ravi Kannan. A random polynomial-time algorithm for approximating the volume of convex bodies. *J. ACM*, 38(1):1–17, 1991.

[DZW$^+$07] S.A. Danziger, J. Zeng, Y. Wang, R.K. Brachmann, and R.H. Lathrop. Choosing where to look next in a mutation sequence space: Active Learning of informative p53 cancer rescue mutants. *Bioinformatics*, 23(13):i104, 2007.

[Grü60] B. Grünbaum. Partitions of mass-distributions and of convex bodies by hyperplanes. *Pacific J. Math*, 10(4):1257–1261, 1960.

[Nem94] A. Nemirovski. Efficient methods in convex programming. *Lecture Notes, Faculty of IE & M, Technion*, 1994.

[Sch00] A. Schrijver. *Theory of Linear and Integer Programming*. Wiley-Interscience series in discrete mathematics and optimization. John Wiley and Sons, Chichester, 2000.

[TC01] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118. ACM New York, NY, USA, 2001.

[TK02] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002.

[Vap00] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.