Active Learning for Smooth Problems

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

Abstract

Recently it was shown that the true sample complexity of active learning is asymptotically better than that for passive learning. In many interesting cases, the improvement has been shown to be exponential [BHW08]; however, there are artificial examples in which the improvement is small. In this paper we provide a basis for a exponential improvements in active learning. We show that exponential improvements arise when the underlying learning problem is "smooth," i.e., the hypothesis class, the instance space and the distribution can all be described by smooth functions. This provides a unified and simplified analysis for most known examples and significantly extends the class learning problems that are "actively learnable at an exponential rate."

1 Introduction

In many problems in machine learning there is a large amount of unlabeled data available which can be labeled at a cost. For example, in spam detection one has many unlabeled emails which could be classified by humans, an expensive task. Thus, there has been much interest in "active learning" in which one tries to minimize the number of labels requested.

Until recently, the benefits of active learning have been unclear. While there are some examples in which active learning provides substantial benefits, there were others where it appeared that active learning provided no significant advantage over ordinary learning [BBZ07, CAL94, Das05b, Das05a, DHM07, Han07a, Han07b]; in particular the worst case complexity appears to be the same. However, in a recent breakthrough [BHW08], it was shown that active learning is always (under mild assumptions) asymptotically better than ordinary learning and in many cases it is "exponentially" better. In fact, in all "reasonable problems" that have been studied it is exponentially better and only on "artificial" examples is it only polynomially better. However, up until now, there has been no formalization of "reasonable problems" and our understanding of exponential improvements is merely anecdotal.

In this paper we provide a formalization of "reasonable problems" by showing that smoothness of the underlying learning problem leads to exponential improvements. We provide a formal definition of smoothness and show that many realistic models are "smooth." Thus, we provide a simplified and unified analysis of these problems. In addition, our approach shows that one can significantly generalize many of these problems, showing the reach of exponential improvements for active learning.

The intuition for our analysis comes from nonlinear optimization where the asymptotic behavior is controlled by the smoothness of the function. For example, if functions are not smooth (but perhaps Lipschitz continuous) one must resort to exhaustive search, even asymptotically; however, if the function to be optimized is smooth then typically one can exponentially improve the asymptotic convergence rate.

This paper is organized as follows. In the next section we review the basic formulation of active learning and then in Section 3 describe our formulation of smoothness and present our main result. Section 4 shows the ease and generality with with this result can be applied and we conclude in Section 5.

2 Active Learning

A learning problem is defined by a hypothesis class C, an instance space X, and a distribution D on X. We will assume that there are only 2 labels, $\{-,+\}$ and a hypothesis $c \in C$ is a subset of X which labels all instances in this subset with $+$ and all others with $-$. Throughout this paper we will assume that the learning problem is realizable, that is there exists some $c^* \in C$ such that all instances are labeled correctly by c^*.[1]

There exists a pseudo-metric d defined by $d(c,c') = P[c \Delta c']$ where $c \Delta c'$ is the symmetric difference,

$$c \Delta c' = (c \setminus c') \bigcup (c' \setminus c).$$

Using this one defines the error rate of a hypothesis c with respect to a true hypothesis c^* to be

$$err(c; c^*) = d(c^*, c).$$

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

[1] Note that we do not need any assumptions on the VC dimension of C.; however, using the techniques in [BDL98] one can show that the VC-exponents they hypotheses that we study are finite (S. Ben-David and M. Lindenbaum, 1998.)

It is also convenient to define the ball of radius r around c to be

$$B_r(c) = \{c' \in C \mid d(c, c') \leq r\}.$$

We assume that the algorithm has access to an infinite sequence of examples, $x_1, x_2, x_3, \ldots$ sampled i.i.d. from X according to D. An active learning algorithm proceeds by choosing some i and requesting the label of x_i and proceeding sequentially. We will be interested in the asymptotic behavior of the error of the chosen hypothesis as a function of the number of labels requested.

Formally, we define the asymptotic complexity of active learning as follows:

Definition 1 (Balcan, et. al.) *A function $S(\epsilon, \delta, c^*)$ is a sample complexity for a learning problem (C,D) if there exists an active learning algorithm $A(t, \delta)$ that outputs a classifier c_t after making at most t label requests, such that for any target function $c^* \in C$, $\epsilon \in (0, 1/2)$, $\delta \in (0, 1/4)$, and for any*

$$t > S(\epsilon, \delta, c^*),$$

then

$$P[err(c_t) \leq \epsilon] \geq 1 - \delta.$$

Note that, as discussed in detail in [BHW08], this definition only requires that the active learner have a small error, not that it be able to guarantee this small error. The main result in that paper is that under this measure, the complexity of active learning is always asymptotically better than that of ordinary learning in which the learner simply requests the labels of every sample in the order in which they arrive. Note that for problems with finite VC dimension, the sample complexity is typically $O(1/\epsilon)$ and their result shows that the sample complexity of active learning is $o(1/\epsilon)$.

Of particular interest to our analysis is their definition of learning problems which are actively learnable at an exponential rate [BHW08], i.e. problems on which active learning shows an exponential improvement over standard learning.

Definition 2 (Balcan et. al.) *We say that (C,D) is* actively learnable at an exponential rate *if there exists an active learning algorithm achieving sample complexity*

$$S(\epsilon, \delta, c^*) = \gamma_{c^*} polylog(1/(\epsilon\delta))$$

for some finite $\gamma_{c^*} = \gamma(c^*, D)$ independent of ϵ and δ.

In [BHW08] several examples of problems which are actively learnable at an exponential rate were presented; however, no general analysis that sheds light on why they are actively learnable at an exponential rate was given. In addition, each example required its own unique analysis. In the next section we will resolve some of these issues by providing general smoothness conditions under which problems are actively learnable at an exponential rate which encompass all of the examples in that paper, provides a single direct method of proof and extends these examples significantly.

Note that our analysis actually works for the "verifiable sample complexity" which is a stronger requirement than the "true sample complexity."

An important technical tool that we will use in our analysis is Hanneke's disagreement coefficient [Han07a] which is defined as follows. For any hypothesis class C, define the region of disagreement as

$$DIS(C) = \{x \in X \ : \ \exists c_1, c_2 \in C \ s.t. \ x \in c_1 \text{ and } x \notin c_2\}.$$

One then defines the disagreement coefficient of a hypothesis to be

$$\theta_c = \sup_{r>0} \frac{P[DIS(B(c, r))]}{r}.$$

Hanneke [Han07a] has shown that the algorithm of of Cohn, Atlas, and Ladner [CAL94] has a sample complexity at most

$$\theta_c polylog(1/(\epsilon\delta))$$

when run with concept class C for target function $c \in C$.

3 Smooth Hypothesis Classes

Our goal in this section is to show that "smooth hypothesis classes" are actively learnable at an exponential rate. Our key tool will be Hanneke's disagreement coefficient, through the following simple lemma.

Lemma 3 *Let*

$$\overline{\theta}_c = \limsup_{r \to 0} \frac{P[DIS(B(c, r))]}{r}.$$

If $\overline{\theta}_c < \infty$ then $c \in C$ is actively learnable at an exponential rate.

Proof: In order to guarantee finiteness of the coefficient one only needs to consider the limit as $r \to 0$ since

$$\frac{P[DIS(B(c, r))]}{r} < 1/r.$$

■

We present our main result in several levels of complexity. We begin with the simplest setting.

3.1 Smooth Euclidean Domains

We first present our analysis in Euclidean spaces. Let $X \subset \Re^n$ for some $n > 0$ and assume that X is compact and of full dimension. Let D be a non-atomic measure with nonzero density on X. We define the hypotheses via an auxiliary function, $f(x, h)$ and assume that the hypotheses are given by

$$c(h) = \{x \in X \mid f(x, h) \leq 0\},$$

for $h \in H$, where H is an open subset of an m dimensional Euclidean space. We require that $f(x, h)$ is C^1.

The following assumptions simplify the analysis. Denote the boundary of $c(h)$ by $\partial c(h)$.

Assumption 1

- *Transversality:*

$$|| \bigtriangledown_x f(x, h)|| > 0$$

on $\partial c(h)$. (Note that here and throughout the paper $||\cdot||$ will represent the Euclidean norm.)

- *Non-degeneracy: for all $dh \in \Re^n$ there exists some $x \in \partial c(h)$ such that*

$$| \bigtriangledown_h f(x,h) \cdot dh| > 0.$$

- *Clone free: for all $h \in H$ there is no $h' \in Closure(H)$ such that $c(h) = c(h')$ and $h \neq h'$.*

Transversality assumes that the boundary of $c(h)$ is a "regular point," which will allow us to use linear approximations. We do not believe that removing this assumption would change our results, although it would complicate the proof.

Non-degeneracy implies that any small change in h leads to changes in $c(h)$. Once again, this simplifies the analysis but is probably not required. For example, in most cases one could reparameterize H so that this condition would be satisfied.

Lastly, clone-freeness says that hypotheses are unique, even in a limiting sense. As we discuss later, one can allow a finite number of clones while maintaining the result. We expect that one can also allow an infinite number of clones, subject to some regularity conditions.

As these three conditions are quite mild we do not lose much by assuming them.

The intuition behind our result is straightforward and follows from the fact that smooth functions can be locally well approximated with linear functions. Consider the case where the spaces are one dimensional, $X \subset \Re$ and $H \subset \Re$, and the elements of H are convex. Thus, $c(h^*)$ is an interval and let $c_\pm(h^*)$ be the endpoints of the interval. The key simplification is that for small dh differentiability implies that $c_\pm(h^* + dh) = c_\pm(h^*) + c'_\pm(h^*)dh$. The import of this is that the endpoints move monotonically, so the maximum disagreement region is the same as the maximum distance and $\overline{\theta}_c = 1$. In addition this argument directly generalizes for higher dimensional X (and 1 dimensional H) since in the linear approximation, the boundary of $c(h)$ moves monotonically. For higher dimensional $H \in \Re^m$, the same argument applies in each dimension and patching them together loses a factor of m showing that $\overline{\theta}_c \leq m^{3/2}$.

Fleshing out this argument yields:

Theorem 4 *Assume the following:*

1. *X is a compact full dimensional subset of $\Re^n$.*
2. *H is an open subset of $\Re^m$.*
3. *D is a non-atomic measure with C^0 density function $p(x) > 0$ on X.*
4. *C is generated by a continuously differentiable function, $f(x,h)$.*
5. *$f(x,h)$ satisfies the non-degeneracy and transversality conditions and is clone free.*

Then for any $c \in C$,

$$\overline{\theta}_c \leq 2m^{3/2}$$

and thus this learning problem is actively learnable at an exponential rate.

Proof: In the following we will fix h^* and $c^* = c(h^*)$. First note that by transversality all the zeros of $f(x,h)$ for fixed h are regular and therefore since X is compact, the boundary ∂c is compact. Thus, any real valued continuous function on ∂c will attain a minimum on ∂c. In addition, by transversality, this boundary is a continuous function of h, so in addition this implies that $d(c(h), c(h'))$ is continuous in h.

We first apply these facts to see that transversality implies that for any dh such that $||dh|| = 1$,

$$\min_{dh \ s.t. \ ||dh||=1} \max_{x \in \partial c^*} | \bigtriangledown_h f(x,h) \cdot dh|$$

is nonzero. This, combined with clone freeness implies that for any $s > 0$ there exists some $r_+ > 0$ such that for all $r < r_+$, $B_r(c^*) \subseteq c(B_s(h^*))$, where

$$c(B_s(h^*)) = \{c(h') \mid h' \in B_s(h^*)\}.$$

Thus, small balls around c^* are contained in small balls around h^*.

This can be seen by defining

$$\tau(r) = max_{h' \in B_r(h^*)} d(c(h), c(h')).$$

By compactness and continuity this value exists and is attained by some $h' \in B_r(h^*)$. By non-degeneracy $\tau(r) > 0$ for $s > 0$ and since $B_r(h^*) \subseteq B_{r'}(h^*)$ for $s \leq s'$ we see that $\tau(r)$ is monotonic in r. It is also continuous in r by transversality. Lastly, note that by definition $\tau(0) = 0$. Thus, for any $s > 0$ we can find r_+ such that $\tau(r_+) = s$.

Next, by compactness we can find a finite covering of ∂c^* by balls of radius $\epsilon > 0$ with centers $x \in X_P \subset \partial c^*$, for any $\epsilon > 0$. This is used to create a partition of a neighborhood of ∂c^* into a union of "small" closed neighborhoods. Each cell of the partition is the set of points that are no farther from the center of one of the small balls, $x \in X_P$, than they are from any other center. Thus, each cell is a closed set which is contained in a ball of radius ϵ and the cells only intersect each other on their boundaries, which are sets of measure 0 with respect to D.

Now, we will look at the effect of changes in ∂c^* in each cell. Consider a specific cell, denoted by A_x and centered at $x \in X_P$. On A_x we can approximate $f(x,h)$ by its Taylor expansion around x, the center of the ball that defined A_x. Thus,

$$f(x + dx, h^* + dh) =$$
$$\bigtriangledown_x f(x,h) \cdot dx + \bigtriangledown_h f(x,h) \cdot dh + o(||x|| + ||h||),$$

where we have used the fact that $f(x,h^*) = 0$ and $\bigtriangledown_x$ denotes the x components of the gradient while $\bigtriangledown_h$ the h components. Setting this to 0 gives us the approximate boundary of $c(h^* + dh)$ which is the hyperplane in dx defined by

$$\bigtriangledown_x f(x,h^*) \cdot dx = - \bigtriangledown_h f(x,h^*) \cdot dh + o(||dx|| + ||dh||).$$

Thus, for small changes in h, locally the tangent plane to the boundary simply shifts a small amount but remains parallel to the original one.

If we want to approximate $d(c(h^*), c(h^* + dh))$ we can add up, over all the cells, the measure between the original tangent planes and the shifted ones. We can approximate the density in each cell by a constant density, chosen to be its value at x, since the function is continuous. Also, since the density is nonzero and the boundary is compact, $p(x)$

has a lower bound on ∂c^*, so the relative approximation is arbitrarily accurate, $o(\epsilon)$, for small enough $\epsilon > 0$.

The key point of this approach is that the movement of the hyperplane only depends on $\bigtriangledown_h f(x,h) \cdot dh$ and not on dh more generally. Thus, for each neighborhood the contribution to the error can be written as $|a^x \cdot dh|$, which is linear in dh, for some constant a^x. By adding all these together, we see that for small dh,

$$d(c(h^*), c(h^* + dh)) = \sum_{x \in X_P} |a^x \dot{dh}| + o(||dh||).$$

First consider the case where $m = 1$. In this case we can write

$$d(c(h^*), c(h^* + dh)) = \sum_{x \in X_P} |a^x| \, |dh| + o(||dh||) = a|dh|,$$

where $a = \sum_{x \in X_P} |a^x|$. Thus, the ball $B_r(c(h^*))$ corresponds to to the parameters arising from $|dh| < r/a$. Next, we note that

$$DIS(A_x \bigcup B_r(c(h^*)))$$

is just the measure between the tangent planes in A_x. Adding all of these up we see that

$$P[DIS(A_x \bigcup B_r(c(h^*)))] = 2r$$

which implies that $\overline{\theta}_{c^*} = 2r/r = 2$.

When $m > 1$ the analysis is more complicated due to the absolute values. Consider the approximation to the distance function,

$$\hat{d}(dh) = \sum_{x \in X_P} |a^x \cdot dh|.$$

This function is convex, as it is a sum of convex functions. It is also symmetric as $\hat{d}(dh) = \hat{d}(-dh)$. This implies that

$$\hat{B}_r = \{dh \in \Re^n \mid c(h^* + dh) \in B_r(c^*)\}$$

is convex and symmetric about the origin.

Now, for any such convex set, Lemma 12, in the appendix, shows that there exists a set of vectors

$$v_1, v_2, \ldots, v_m \in \hat{B}_r$$

such that for any $b \in \hat{B}_r$ we can write $b = \sum_i \beta_i v_i$ where $|\beta_i| \leq m^{1/2}$. Now consider a cell A_x. The intersection of the disagreement region and that cell is simply the area between the two extreme values of the tangent surfaces over $b \in \hat{B}_r$ which arise from some b^x on the boundary of $\hat{B}_r$ and its reflection $-b^x$. The measure of this region is simply

$$2|a^x \cdot b^x| = 2|a^x \cdot (\sum_i \beta_i v_i)| \leq 2 \sum_i |\beta_i| \, |a^x \cdot v_i|,$$

so we have the following bound

$$P[DIS(B_r(c^*))] \leq 2m^{1/2} \sum_{x \in X_P} \sum_i |a^x \cdot v_i|.$$

Now, by the definitions of v_i and $B_r(c^*)$, we have

$$r \geq \sum_{x \in X_P} |a^x \cdot v_i|,$$

for all i. Summing over all i yields

$$2mr \geq \sum_{x \in X_P} \sum_i |a^x \cdot v_i| \geq m^{-1/2} P[DIS(B_r(c^*))].$$

Thus we see that

$$\overline{\theta}_{c^*} = P[DIS(B_r(c^*))]/r \leq 2m^{3/2} r/r = 2m^{3/2}$$

proving the theorem when $\epsilon \to 0$, which corresponds to the case when $r \to 0$, since all the error terms are $o(\epsilon)/\epsilon$ and vanish in the limit.

■

In addition, the theorem is almost tight. Let $X = [0,1]$, D be the uniform distribution,

$$H = \{h \in \Re^m \mid 0 < h_i < 1/(2m+1)\}$$

and $c(h) = \bigcup_i c_i(h_i)$ where $c_i(h_i)$ is the interval of length h_i centered at $i/(m+1)$. Let $c = c(h)$ where $h_i = 1/4m$. It is easy to see that for small r the ball of radius r centered at $c(h)$, $B_r(c(h))$ corresponds to the region in H where

$$\sum_i |h_i' - h| < r.$$

The disagreement region,

$$DIS(B_r(c(h)))$$

corresponds to the region in H where $|h_i' - h_i| < r$. Thus,

$$P[DIS(B_r(c(h)))] = 2rm,$$

so the disagreement coefficient is $2rm/r = m$ and taking the limit as $r \to 0$ yields $\overline{\theta}_c = 2m$.

Now, for general problems, since X is full dimensional, we can consider some open ball, B contained in X. By the assumptions on D we can choose B small enough so that the distribution is essentially constant, up to small error terms that can be made arbitrarily small.

Now we consider a small line segment contained in B and a cylindrical neighborhood of that line segment which is also contained in B. Then we can use this same construction by replacing intervals of the line $c_i(h_i)$ by "intervals" of the cylinder which leads to the same result. We state this in a Theorem.

Theorem 5 *For any X, D there exist a parameter space $H \in \Re^m$ and function $f(x,h)$ which satisfies the assumptions of Theorem 4 such that $\overline{\theta}_c = 2m$ for the induced set of hypotheses.*

While the example used to prove this theorem is quite specialized, we claim that the result is common – $\overline{\theta}_c = 2m$ will commonly arise. For example, it can be shown that if we take C to be the set of axis oriented ellipsoids in $\Re^m$ where the parameter space is a subset of $\Re^m$ (the length of the axes) then a (tedious) calculation shows that in this case $\overline{\theta}_c = 2m$. Similarly for generalized hypercubes centered at the origin one gets $\overline{\theta}_c = 2m$. In some sense it appears that $\overline{\theta}_c$ will be equal to the number of "relevant" parameters, although the formal definition of "relevance" appears to be technically complex and we do not pursue it further here.

In fact, we conjecture that $\overline{\theta}_c = 2m$ is the true upper bound too, as opposed to our result of $2m^{3/2}$. This can be seen in the proof of Lemma 12 in which one could tighten the constraints on the β_i's significantly, a result which should be usable to tighten our upper bound.

3.1.1 Example: Circles and Ellipsoids

An interesting hypothesis class that satisfies these assumptions are hyperspheres in $\Re^n$. Let

$$X = \{x \in \Re^n \mid \sum_{i=1}^{n} x_i^2 \leq 1\}$$

and

$$H = \{h \in \Re^{n+1} \mid \sum_{i=1}^{n} h_i^2 < 1 - h_{n+1}^2, \ \ 0 < h_{n+1} < 1\}.$$

Now define

$$f(x,h) = \sum_{i=1}^{n} (x_i - h_i)^2 - h_{n+1}^2.$$

Then $c(h)$ is the hypersphere centered at $(h_1, \ldots, h_n)$ of radius h_{n+1}. Our analysis above showed that this class is actively learnable at an exponential rate. In fact, one can extend this result to other interesting hypothesis classes. For example, one can easily modify this example to allow for axis oriented ellipsoids and, more generally, arbitrary ellipsoids.

3.2 Piecewise Smoothness and Clones

Note that our proof of Theorem 4 only relies on properties of small neighborhoods of $\partial c(h)$. Thus, as long as our function $f(x,h)$ is locally smooth, our analysis extends directly.

An important motivating example in active learning theory is that of axis oriented rectangles. The general version of this hypothesis class is defined by $X = [0,1]^n$, $H \subset [0,1]^{2n}$ and

$$c(h) = \{x \in X \mid h_i \leq x_i \leq h_{n+i} \ \ \forall i\}.$$

However, for this hypothesis class there is no function $f(x,h)$ which generates it and is C^1 everywhere. One can find such a function which is C^0 everywhere and C^1 a.e.; however, this does not appear to be sufficient, since the accuracy of the approximations might not hold uniformly on the whole surface of a hypothesis. Thus, we consider the following requirements on $f(x,h)$ to avoid these difficulties.

First we define a smooth open partition of $\partial c(h)$ to be a finite collection of connected co-dimension 1 open sets such that the closure of their union is $\partial c(h)$ and the boundary of each set is C^1.

For example it is easy to see that one can partition the boundary of any axis oriented rectangle into $2n$ open subsets of the defining hyperplanes.

It is clear that our proof can be applied to each element of the partition separately. Since there are only a finite number of elements of the partition one can easily guarantee uniformity of the approximations. In addition, since all of our analysis is local, we only need to require that the function $f(x,h)$ is C^1 on a neighborhood of the surface on each element of the partition.

For example, for the axis oriented rectangles we can define $f(x,h)$ to be the (oriented) distance of x to the nearest edge of the rectangle.

In addition, note that our restriction on clones is stronger than necessary as we can allow a finite number of clones. The only change in our analysis would be that the analysis would have to be carried out for each clone, but once again, if there are only a finite number of clones then the approximations will hold uniformly.

Thus, we have the following theorem:

Theorem 6 *Assume the following:*

1. *X is a compact full dimensional subset of $\Re^n$.*
2. *H is an open subset of $\Re^m$.*
3. *D is a non-atomic measure with C^0 density function $p(x) > 0$ on X.*
4. *C is generated by a continuously differentiable function, $f(x,h)$.*
5. *C is generated by $f(x,h)$ where*
   (a) *f is C^0.*
   (b) *There exists a smooth open partition of $\partial c(h)$ such that on the relative interior of each element of the partition, f is C^1.*
   (c) *f has a bounded number of clones and satisfies the non-degeneracy and transversality conditions.*

Then for any $c \in C$,

$$\overline{\theta}_c \leq 2m^{3/2}$$

and thus this learning problem is actively learnable at an exponential rate.

Thus, axis oriented rectangles as well as direct generalizations to many classes of polytopes are all actively learnable at an exponential rate.

3.3 Smooth Manifolds

Recall that a manifold is a space which is locally Euclidean [War83]. For example, the surface of sphere, S^n, and the set of rotations, $SO(n)$, are both manifolds. For any point $m \in M$ of a manifold there is a chart, g, which is a function from a neighborhood of m to a Euclidean space. The set of charts is known as the atlas. We will consider only compact manifolds, in which case there is a finite set of charts which cover the manifold. A manifold is C^1 if the composition of charts is continuously differentiable and on such a manifold, differentiation is well defined.

Given a manifold and a finite atlas we can decompose the manifold into a finite set of pieces and then analyze them locally using the charts using the standard tools from multivariate calculus. Following from our decomposition approach in the previous section, it is easy to see that this corresponds to a partition and our analysis can be applied directly to each element of the partition, i.e., each chart in a finite atlas. Thus we get the following useful extension of our main theorem.

Theorem 7 *Assume the following:*
Assume the following:

1. *X is a compact full dimensional subset of a C^1 manifold.*
2. *H is a C^1 manifold without a boundary of dimension m.*
3. *D is a non-atomic measure with C^0 density function $p(x) > 0$ on X.*

4. *There exists a finite set of charts on $X \times H$ such that C is generated by $f(x,h)$ where (in local coordinates on each chart)*

   (a) *f is C^0 everywhere.*
   (b) *For every $h \in H$ there exists an open partition of $\partial c(h)$ such that f is C^1.*
   (c) *f has a finite number of clones and satisfies the non-degeneracy and transversality conditions.*

Then for any $c \in C$,

$$\overline{\theta}_c \leq 2m^{3/2}$$

and thus this learning problem is actively learnable at an exponential rate.

Proof: In order to prove this it suffices to construct a finite partition of ∂c^*, as in the proof of Theorem 4, such that each element of the partition is contained in a single chart and then apply essentially the identical proof. To do this we take a finite covering of H by open charts, which exists since ∂c^* is compact. Then we can choose $\epsilon > 0$ sufficiently small such that the fraction of partition elements which are not contained in a single chart are sufficiently small that they can be ignored with only a small loss of accuracy. ■

Note that we do not require that H be open, only that it not contain a boundary. This is useful since many manifolds are closed but do not have a boundary, such as the boundary of an n-dimensional hypersphere, such as the circle.

We have presented the coordinate representation of the theorem; however, one could present it in a "coordinate free" manner. For a good introduction to these methods see [AM94].

Example: Linear Separators and S^n

An important set of hypothesis classes arise from linear separators, as in SVMs e.g., [Joa02]. In this case the hypothesis class is typically generated by linear separators, $b^t x + c \geq 0$, or their unions and/or intersections. These can be easily shown to be smooth by letting $H \subset S^n \times \Re^n$ where $h = (b, c)$ where $b \in S^n$ is an element from the surface of the unit n-sphere and $c \in \Re^n$.

Thus, for any smooth distribution over say $[0,1]^n$ the class of linear separators will be actively learnable at an exponential rate.

Another important case, the analysis of linear separators over the uniform distribution on the surface of the unit sphere [BBL06, BBZ07, Das05b, DHM07, DKM05] appears, at first, to be problematic, since if we take X to be $[0,1]^n$ then the distribution is not smooth, as it is concentrated on the surface of the unit sphere. In fact, in this case the detailed relationship between linear separators and spheres is crucial to the analysis. For example, if the hypothesis class contained separators with "spherical pieces" then problems could arise. The problems are avoided with linear separators, but small perturbations of this hypothesis class could be problematic.

Nonetheless, the analysis of this problem is straightforward when we set X to be the surface of the unit sphere, a smooth manifold. In this space, the projection of linear separators onto the surface are codimension 1 spheres and thus it is easy to show that this class is actively learnable at an exponential rate for any smooth non-vanishing distribution on the sphere.[2]

In fact, this analysis shows that linear separators are actively learnable at an exponential rate over any smooth distribution on any smooth, convex, nowhere flat, sub-manifold of $\Re^n$, such as ellipses. More generally, one can take the union of multiple ellipses as the direct product of several smooth manifolds (which is also a smooth manifold). In addition, one can often reduce the requirement of convexity, as long as the hyperplanes are never tangent to a "flat" piece of the surface.

One can even generalize these arguments to show that even in cases where results don't imply that a class/distribution pair is actively learnable at an exponential rate, small perturbations of either the distribution of the hypothesis class restore the exponential learnability.

3.4 Countable Unions of Hypothesis Classes

Note that our theorem does not include the case where the hypothesis class is a finite union of smooth classes which are actively learnable at an exponential rate. One could modify the proof to cover finite collections, but the extension to countably infinite collections appears to be technically complex. Fortunately, [BHW08] has shown that the extension to countable unions of hypothesis classes is straightforward.

Theorem 8 (Balcan et. al.) *If $C_1, C_2, \ldots$ are all hypothesis classes for X, D which are all actively learnable at an exponential rate then $C = \bigcup_i C_i$ is also actively learnable at an exponential rate.*

Thus, one can combine arbitrary finite collections of hypothesis classes which are individually actively learnable at an exponential rate into a single large class which is actively learnable at an exponential rate.

4 Tools for Constructing Smooth Hypothesis Classes

In this section we present two powerful methods for constructing smooth hypothesis classes.

4.1 Semi-Algebraic Sets

Semi-algebraic sets can be used for constructing many interesting hypothesis classes. (See, e.g. [BDL98].) In order to

[2] To see the problem when X has flat pieces, let X be the surface of the unit square, $X = \partial S$ where

$$S = \{x \in \Re^2 \mid \forall i: \ 0 \leq x_i \leq 1\}$$

with the uniform distribution. Then the halfspace given by

$$HS(h_1) = \{x \in \Re^2 \mid x_1 \leq h_1\}$$

generates the hypothesis

$$c(h_1) = HS(h_1) \bigcap X,$$

and it is easy to see that $c(h_1)$ is discontinuous at $h_1 = 0$; however, if we deformed X slightly so that it was strictly convex, the hypotheses would be continuous for all values of h_1.

construct a semi-algebraic hypothesis class we allow there to be a finite number of defining functions $f_i(x,h)$ and then allow finite unions or intersections of the hypotheses generated.

For example, one can define the class of axis oriented rectangles in $\Re^2$ by considering the intersection of the sets created by the following four functions for $H \subset \Re^4$:

$$f_1(x,h) = h_1 - x_1,$$
$$f_2(x,h) = x_1 - h_2,$$
$$f_3(x,h) = h_3 - x_2,$$
$$f_1(x,h) = x_2 - h_4.$$

In a semi-algebraic set we additionally require X and H be subsets of a Euclidean space (as in Theorem 1) and that each of the functions be a polynomial in x, h. In addition, we require that the functions be non-degenerate; that is, for any $h \in H$ all the f_i's are distinct.

The key issue is whether we can find a smooth partition of the resulting hypothesis class which only relies on a single function in each element of the partition. However, when the functions are polynomials, Bezout's theorem bounds the number of intersections among the surfaces generated by each function which guarantees that such a smooth partition exists and is finite. Thus under very mild conditions semi-algebraic sets of hypotheses can be actively learned at an exponential rate.[3]

Clearly semi-algebraic hypothesis classes are very general. These include most of the previous (non manifold) examples as well as many generalizations.

4.2 Transformations and Lie Groups

Next, we consider a simple, but powerful, method for constructing hypothesis classes. Let $T_h(x)$ be a C^1 mapping from X (or its containing manifold) to itself, whose gradient is Lipschitz continuous. Then given some initial hypothesis defined by $g(x) \leq 0$ where $g(x)$ is also C^1, define

$$f(x,h) = g(T_h(x)).$$

It is easy to see that $f(x,h)$ will be C^1. Thus it can be used to define a hypothesis class which is actively learnable at an exponential rate.

Note that this construction also works if the functions are only smooth on the elements of a smooth partition, so our analysis also applies in that case.

For example, if $X \subset \Re^n$ and $H \subset \Re^n$, T_h could be the set of translations on X,

$$T_h(x) = x + h.$$

Then if $g(x)$ defines a set in X we can use T_h to generate all translations of that set, such as all balls of unit radius or unit hyperspheres.

Another useful set of transformations are the dilations, in which case $H' \subset \Re^n$ and $T'_{h'}$ is defined by

$$T'_{h'}(x)_i = x_i / h'_i.$$

In this case, starting with a hypersphere centered at $x_0 \in X$ one can generate all axis oriented ellipsoids centered at x_0. Similarly, starting with a hypercube centered at x_0 one can generate all axis oriented parallelepipeds centered at x_0.

In addition, one can combine two transformations to create their combination. For example define

$$\hat{T}_{h,h'}(x) = T_h(T'_{h'}(x)).$$

Then beginning with a hypersphere, one can generate all ellipsoids.

A generalization of these two examples is the set of all affine transformations: $X \subset \Re^n$ and $H \subset \Re^{n^2+n}$ where the first n^2 elements of h are the elements of a matrix A_h and the last n elements are the elements of a vector b_h. Then

$$T_h(x) = A_h x + b_h$$

which is an affine transformation, which can be used to generate many interesting hypothesis classes.

However, this approach runs into difficulties for certain important classes of transformations. For example, consider the set of rotations in $\Re^2$. This can be implemented by the matrix

$$M_h = \begin{pmatrix} h_1 & h_2 \\ h_3 & h_4 \end{pmatrix}$$

but then the set $H \subset \Re^4$ is not open. One can also try constructing this as

$$M_h = \begin{pmatrix} h_1 & h_2 \\ -h_2 & h_1 \end{pmatrix}$$

but still the set $H \subset \Re^2$ is not open. The resolution of this issue (and its generalization to more complicated problems) is to realize that the correct representation of H is as a Lie Group.

A Lie Group is a group that is also a smooth manifold. A useful Lie group is the simple orthogonal group $SO(n)$ which is the group of all orthogonal $n \times n$ matrices with unit determinant and it is manifold of dimension $n(n-1)/2$. The action of this group, through matrix multiplication, is that of arbitrary n-dimensional rotations. Locally, one can find $n-1$ coordinates to represent this group, but globally one must use a manifold. Thus if we let H be the Lie Group $SO(n)$ and A_h be the rotation matrix corresponding to $h \in H$ then our results apply directly.

5 Agnostic Active Learning

In some cases, our results extend naturally to the agnostic setting, where the true hypothesis may not be contained in C. Let $c^* \notin C$ be the true hypothesis and $c' \in C$ be the best hypothesis, i.e. the one which minimizes $err(c; c^*)$ over C. Let $\nu = err(c'; c^*)$.

As Hanneke has shown, a key parameter in the number of label requests by the A^2 algorithm [BBL06] for agnostic active learning can be bounded in terms of the maximum disagreement coefficient:

$$\theta(C) = \max_{c \in C} \theta_c.$$

Theorem 9 (Hanneke, 2008) *If $\theta(C)$ is the disagreement coefficient for C, then with probability at least $1-\delta$, given the inputs C, ϵ, and δ, A^2 outputs $c' \in C$ with*

$$err(h', h*) \leq \nu + \epsilon$$

[3] The formal analysis parallels that for sign patterns of polynomials as discussed in [War68, Alo96].

and the number of label requests made by A^2 is at most

$$O\left(\theta(C)^2(\frac{\nu^2}{\epsilon^2}+1)(m\log(1/\epsilon)+\log(1/\delta))\log(1/\epsilon)\right).$$

Thus, if we show that $\theta(C)$ is finite for some hypothesis class and distribution on the instance space, then that problem is agnostically actively learnable at an exponential rate, *assuming that* $\nu/\epsilon = O(1)$, an assumption we maintain throughout this section.[4] For example, Hanneke [Han07a] has shown that for the hypothesis space which consists of all linear separators through the origin where the instances are given by a uniform distribution over the unit sphere, $\theta(C) \leq \pi m^{1/2}$, thus they are, in certain limits, agnostically actively learnable at an exponential rate. We will show how our techniques can extend these results to arbitrary distributions, with non-vanishing densities over arbitrary convex surfaces which don't intersect the origin.

First we note that the extension is a bit more subtle than it might appear at first glance. Although $\overline{\theta}_c \leq m^{3/2}$ for all $c \in C$, subject to our smoothness assumptions, this does not immediately imply that $\theta(C)$ is finite. This is because the divergence of disagreement coefficients can occur over a sequence of finite r's, but for different c's. In the case where C is open, the value of r in which the linear approximation is valid, while finite for every $c \in C$, could be infinite over the entire C.

One way to avoid this problem is to take a compact subset of C and assume some additional differentiability of $f(x,h)$. For example in the example which looks at subintervals of the unit interval, this might correspond to setting a lower limit on the length of a hypothesis interval. An alternative approach relies on the properties of manifolds which can be compact but without a boundary. This is the case that arises for linear separators through the origin as they are completely specified by a point on the surface of the unit hypersphere, a compact manifold without boundary.

First we state a theorem:

Theorem 10 *Let H, X, D, f satisfy the assumptions of Theorem 4 or Theorem 7 and in addition assume that $f(x,h)$ is twice continuously differentiable (on each chart). Let $C' \subseteq C(H)$ be compact. Then $\theta(C') < \infty$ and C' is agnostically actively learnable at an exponential rate, if $\nu/\epsilon = O(1)$.*

Proof: (Sketch:) Let

$$\theta_r(C) = \max_{c \in C} \frac{P[DIS(B(c,r))]}{r}.$$

Our goal is to show that

$$\sup_{r>0} \theta_r(C) < \infty.$$

First we note that for any finite r, $\theta_r(C) \leq 1/r$, so all we need to show is that

$$\limsup_{r \to 0} \theta_r(C) < \infty.$$

This does not follow immediately from our result that $\overline{\theta}_c \leq m^{3/2}$, since the limiting behavior could vary by $c \in C$. However, one can see from the proof of Theorem 4 that for any level of approximation, there is some $\epsilon(c)$ such that the linear approximation is valid and if $f(x,h)$ is C^2 then this function $\epsilon(c(h))$ is continuous in h and by compactness there exists a uniform bound which completes the proof. ■

Thus, we can extend many of our previous results to the agnostic case and extend some of Hanneke's results.

6 Conclusions

Our results imply that many, and perhaps most, learning problems of interest are actively learnable at an exponential rate; however, these results are asymptotic and it would be important to understand how quickly the asymptotic region is reached.

More broadly, we expect that the idea that small perturbations of "bad" problems can make them "good" merits further studies. For example, does this imply that in some stronger sense that for "geometric learning problems" exponential improvements are "generic."

Next, our results for agnostic learning are very preliminary. We conjecture that one does not need to restrict to compact sets of hypotheses for similar agnostic active learning results as smoothness should allow one to provide straightforward conditions, such as bounds on the Lipschitz constants of the function $f(\cdot,\cdot)$, which allow one to consider open sets of hypotheses. A more comprehensive study of agnostic active learning remains to be done.

Lastly, while the CAL algorithm provides exponential convergence we expect that active learning problems which take direct advantage of geometry and smoothness (such as in [BHW08]) could be simpler and more efficient in practice and expect that a geometric analysis of these problems may lead to improved active learning algorithms.

Acknowledgments

I would like to thank Shane Henderson, Adrian Lewis, and Mike Todd for helpful conversations and two anonymous referees for helpful comments. This work has been supported in part by the NSF under grants ITR-0325453 and CDI-0835706.

A Lemma

The key to proving Lemma 12 is the following result by John [Joh48].

Theorem 11 (John, 1948) *Let $B \subset \Re^m$ be a convex set which is symmetric about the origin. Then there exists an ellipsoid, the Löwner-John Ellipsoid, $E \in \Re^n$ such that*

$$m^{-1/2}E \subseteq B \subseteq E$$

where

$$m^{-1/2}E = \{m^{-1/2}x \mid x \in E\}.$$

Lemma 12 *Let $B \subset \Re^n$ be a convex set which is symmetric about the origin, i.e, if $b \in B$ then $-b \in B$. Then there exist a set of m orthogonal vectors, $v_1, v_2, \ldots, v_m \in B$, such that any $b \in B$ can be written $\sum_i \beta_i v_i$ for*

$$|\beta_i| \leq m^{1/2}.$$

[4] Without this assumption, the ν/ϵ term will dominate the $\log(1/\epsilon)$ term, leading to polynomial convergence.

Proof: Proof: By John's lemma, we consider the enclosed ellipsoid, $m^{-1/2}E$. Let v_i be one of the vectors defining the i'th axis of E. Then $v_i \in B$ by John's lemma. Then note that $\sum_i \alpha_i v_i$ for $|\alpha_i| \leq m^{1/2}$ must contain E, completing the proof. ■

References

[Alo96] N. Alon. Tools from higher algebra, Handbook of combinatorics (vol. 2), 1996.

[AM94] R. Abraham and J. Marsden. *Foundations of Mechanics*. Perseus Books, 1994.

[BBL06] N. Balcan, A. Beygelzimer, and J. Langford. Agostic active learning. *23rd International Conference on Machine Learning, Pittsburgh, PA, USA*, 2006.

[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.

[BDL98] S. Ben-David and M. Lindenbaum. Localization vs. Identification of Semi-Algebraic Sets. *Machine Learning*, 32(3):207–224, 1998.

[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.

[CAL94] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

[Das05a] S. Dasgupta. Analysis of a greedy active learning strategy. *Advances in Neural Information Processing Systems 17: Proceedings Of The 2004 Conference*, 2005.

[Das05b] S. Dasgupta. Coarse sample complexity bounds for active learning. *Advances in Neural Information Processing Systems*, 18:2, 2005.

[DHM07] S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. *Neural Information Processing Systems (NIPS)*, 2007.

[DKM05] S. Dasgupta, A.T. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. *Eighteenth Annual Conference on Learning Theory*, 2005.

[Han07a] S. Hanneke. A bound on the label complexity of agnostic active learning. *Proceedings of the 24th International Conference on Machine Learning*, pages 353–360, 2007.

[Han07b] S. Hanneke. Teaching dimension and the complexity of active learning. *Proceedings of the 20th Conference on Learning Theory*, 2007.

[Joa02] T. Joachims. *Learning to Classify Text Using Support Vector Machines – Methods, Theory, and Algorithms*. Kluwer/Springer, 2002.

[Joh48] F. John. Extremum problems with inequalities as subsidiary conditions. In *Studies and Essays presented to R. Courant on his 60th Birthday*, pages 187–204, 1948.

[War68] H.E. Warren. Lower bounds for approximation by nonlinear manifolds. *Trans. Amer. Math. Soc*, 133(1):167–178, 1968.

[War83] F.W. Warner. *Foundations of Differentiable Manifolds and Lie Groups*. Springer, 1983.