Coarse sample complexity bounds for active learning

Sanjoy Dasgupta
UC San Diego
dasgupta@cs.ucsd.edu

Abstract

We characterize the sample complexity of active learning problems in terms of a parameter which takes into account the specific target hypothesis and the distribution over the input space. For homogeneous (through the origin) linear separators in $\mathbb{R}^d$, we show that under a fairly wide range of data distributions, the number of labels needed by an active learner to achieve an error rate $\leq \epsilon$ is just $\tilde{O}(d \log^2 1/\epsilon)$, exponentially smaller than the usual $\Omega(d/\epsilon)$ sample complexity of supervised learning.

1 Introduction

The goal of active learning is to learn a classifier in a setting where data comes unlabeled, and any labels must be explicitly requested and paid for. The hope is that an accurate classifier can be found by buying just a few labels.

So far the most encouraging theoretical results in this field are [6, 5], both of which apply to the hypothesis class of homogeneous (i.e. through the origin) linear separators, in the specific case where the data is distributed uniformly over the unit sphere in $\mathbb{R}^d$. They show that if the labels correspond perfectly to one of the hypotheses (i.e. the separable case), then just $O(d \log d/\epsilon)$ labels are needed to learn a classifier with error less than ϵ. This is exponentially smaller than the usual $\Omega(d/\epsilon)$ sample complexity of learning linear classifiers in a supervised setting.

However, generalizing this result is non-trivial. Suppose, for instance, that the hypothesis class is expanded to include *non-homogeneous* linear separators. Then even in just two dimensions, and under the same benign input distribution, we will see that there are some target hypotheses for which active learning does not help much, for which $\Omega(1/\epsilon)$ labels are needed. In fact, in this example the label complexity of active learning depends heavily on the specific target hypothesis, and ranges all the way from $O(\log 1/\epsilon)$ to $\Omega(1/\epsilon)$.

In this paper, we consider arbitrary hypothesis classes $\mathcal{H}$ of VC dimension $d < \infty$, and learning problems which are separable. We characterize the sample complexity of active learning in terms of a parameter which takes into account: (1) the distribution $\mathbb{P}$ over the input space $\mathcal{X}$; (2) the specific target hypothesis $h^* \in \mathcal{H}$; and (3) the desired accuracy ϵ.

Specifically, we observe that distribution $\mathbb{P}$ induces a natural topology on $\mathcal{H}$, and we define a *splitting index* ρ which captures the relevant local geometry of $\mathcal{H}$ in the vicinity of h^*, at scale ϵ. We show that this quantity coarsely describes the sample complexity of active learning: any active learning scheme requires $\Omega(1/\rho)$ labels and there is a generic active learner which always uses at most $\tilde{O}((d/\rho) \log^2(1/\epsilon))$ labels.[1]

This ρ is always at least ϵ; if it is ϵ we get approximately the usual sample complexity of supervised learning. But sometimes ρ is a constant, and in such instances active learning gives an exponential improvement in the number of labels needed.

We derive splitting indices for various hypothesis classes. For homogeneous linear separators under an input distribution which is uniform over the unit sphere in $\mathbb{R}^d$, we easily find ρ to be a constant – perhaps the

[1] The $\tilde{O}(\cdot)$ notation is used here to hide factors of the form $\text{polylog}(d, 1/\delta, 1/\rho, \log 1/\epsilon, \log 1/\tau)$, where δ and τ will be introduced later.

most direct proof yet of the efficacy of active learning in this case. But we also show, for the first time, that this $\tilde{O}(d \log^2 1/\epsilon)$ label complexity holds in substantially more general situations. Specifically, we consider input distributions which are a multiplicative factor λ away from uniform, and we find that ρ remains a constant (independent of λ), provided the amount of unlabeled data is increased by a factor of $\tilde{O}(\lambda^2)$.

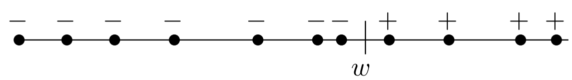
2 Sample complexity bounds

2.1 Motivating examples

Linear separators in $\mathbb{R}^1$

Our first example is taken from [3, 4]. Suppose the data lie on the real line, and the classifiers are simple thresholding functions, $\mathcal{H} = \{h_w : w \in \mathbb{R}\}$:

$$h_w(x) = \begin{cases} 1 & \text{if } x \geq w \\ 0 & \text{if } x < w \end{cases}$$



VC theory tells us that if the underlying distribution $\mathbb{P}$ is separable (can be classified perfectly by some hypothesis in $\mathcal{H}$), then in order to achieve an error rate less than ϵ, it is enough to draw $m = O(1/\epsilon)$ random labeled examples from $\mathbb{P}$, and to return any classifier consistent with them. But suppose we instead draw m *unlabeled* samples from $\mathbb{P}$. If we lay these points down on the line, their hidden labels are a sequence of 0's followed by a sequence of 1's, and the goal is to discover the point w at which the transition occurs. This can be accomplished with a simple binary search which asks for just $\log m = O(\log 1/\epsilon)$ labels. Thus, in this case active learning gives an *exponential* improvement in the number of labels needed.

Can we always achieve a label complexity proportional to $\log 1/\epsilon$ rather than $1/\epsilon$? A natural next step is to consider linear separators in *two* dimensions.

Linear separators in $\mathbb{R}^2$

Let $\mathcal{H}$ be the hypothesis class of linear separators in $\mathbb{R}^2$, and suppose the input distribution $\mathbb{P}$ is some density supported on the perimeter of the unit circle. It turns out that the positive results of the one-dimensional case do not generalize: there are some target hypotheses in $\mathcal{H}$ for which $\Omega(1/\epsilon)$ labels are needed to find a classifier with error rate less than ϵ, no matter what active learning scheme is used.

To see this, consider the following possible target hypotheses (Figure 1):

- h_0: all points are positive.
- h_i $(1 \leq i \leq 1/\epsilon)$: all points are positive except for a small slice B_i of probability mass ϵ.

The slices B_i are explicitly chosen to be disjoint, with the result that $\Omega(1/\epsilon)$ labels are needed to distinguish between these hypotheses. For instance, suppose nature chooses a target hypothesis at random from among the $h_i, 1 \leq i \leq 1/\epsilon$. Then, to identify this target with probability at least $1/2$, it is necessary to query points in at least (about) half the B_i's.

Thus for these particular target hypotheses, active learning offers little improvement in sample complexity over regular supervised learning. What about other target hypotheses in $\mathcal{H}$, for instance those in which the positive and negative regions are more evenly balanced? Consider the following active learning scheme:

1. Draw a pool of $O(1/\epsilon)$ unlabeled points.
2. From this pool, choose query points at random until at least one positive and one negative point have been found. (If all points have been queried, then halt.)
3. Apply binary search to find the two boundaries between positive and negative on the perimeter of the circle.

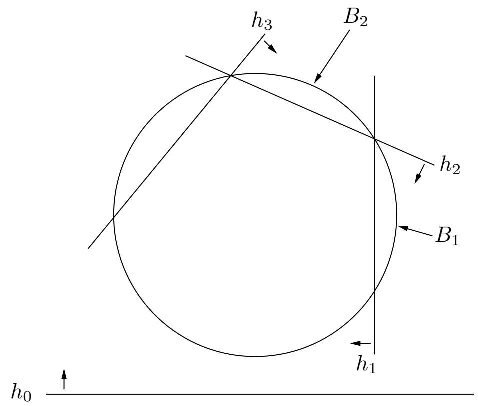


Figure 1: $\mathbb{P}$ is supported on the circumference of a circle. Each B_i is an arc of probability mass ϵ.

For any $h \in \mathcal{H}$, define

$$i(h) \;=\; \min\{\text{positive mass of } h, \text{ negative mass of } h\}.$$

It is not hard to see that when the target hypothesis is h, step (2) asks for $O(1/i(h))$ labels (with probability at least 9/10, say) and step (3) asks for $O(\log 1/\epsilon)$ labels.

Thus even within this simple hypothesis class, the label complexity of active learning can run anywhere from $O(\log 1/\epsilon)$ to $\Omega(1/\epsilon)$, depending on the specific target hypothesis.

Linear separators in $\mathbb{R}^3$

In our two previous examples, the amount of unlabeled data needed was $O(1/\epsilon)$, exactly the usual sample complexity of supervised learning. We next turn to a case where it helps to have significantly more unlabeled data than this.

Consider the distribution of the previous, two-dimensional example: for concreteness, fix $\mathbb{P}$ to be the uniform distribution on the unit circle in $\mathbb{R}^2$. Now lift it into three dimensions by adding to each point $x = (x_1, x_2)$ a third coordinate $x_3 = 1$. Let $\mathcal{H}$ consist of *homogeneous* (i.e. through the origin) linear separators in $\mathbb{R}^3$. Clearly the bad cases of the previous example persist. And as before, the hard part is finding an initial positive point and negative point; thereafter binary search is possible.

Suppose, now, that a trace amount $\tau \ll \epsilon$ of a second distribution $\mathbb{P}'$ is mixed in with $\mathbb{P}$ (Figure 2, left), giving an overall combined distribution of $(1-\tau)\mathbb{P} + \tau\mathbb{P}'$. This $\mathbb{P}'$ is uniform on the circle

$$x_1^2 + x_2^2 = 1, \quad x_3 = 0.$$

It represents such a tiny fraction of the data that it has almost no influence on the error rate of a classifier. However, it is a source of good query points, for the following reason: the "bad" linear separators in $\mathcal{H}$ cut off just a small portion of $\mathbb{P}$ but nonetheless divide $\mathbb{P}'$ perfectly in half (Figure 2, right). This permits a three-stage algorithm which uses $\mathbb{P}'$ to help locate positive and negative points of $\mathbb{P}$:

1. By running the two-dimensional algorithm (given above) on points from $\mathbb{P}'$, approximately identify (within arc-length $\epsilon\pi/2$, say) the two places at which the target hypothesis h^* cuts $\mathbb{P}'$.

2. Let the midpoints of these (approximate) positive and negative $\mathbb{P}'$-intervals be $p' = (p_1, p_2, 0)$ and $n' = (n_1, n_2, 0)$. Look at the points directly above them, $p = (p_1, p_2, 1)$ and $n = (n_1, n_2, 1)$. If the positive region of $\mathbb{P}$ has $\mathbb{P}$-mass at least ϵ, then it must contain p and all points within arc-length $\epsilon\pi/2$ of it; likewise with n and the negative region. So query a point within arc-length $\epsilon\pi/2$ of each of p, n; if they are both positive or both negative, stop and announce the all-positive or all-negative hypothesis. Otherwise continue.

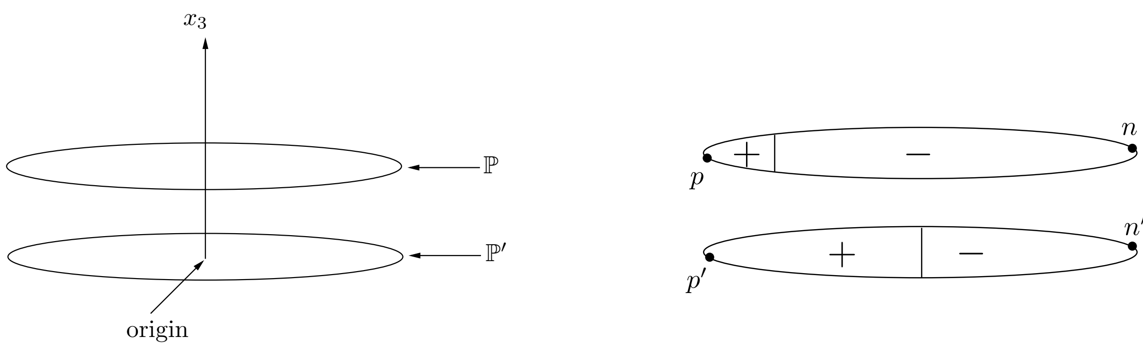


Figure 2: *Left:* The same distribution $\mathbb{P}$ as in the 2-d case, now lifted to 3-d, and with trace amounts of another distribution $\mathbb{P}'$ mixed in. *Right:* A bad target hypothesis cuts off a small portion of $\mathbb{P}$ but exactly half of $\mathbb{P}'$.

3. Armed with this pair of positive and negative points from $\mathbb{P}$, do binary search on $\mathbb{P}$, as in step (3) of the 2-d algorithm.

Steps (1) and (3) each use $O(\log 1/\epsilon)$ labels.

This $O(\log 1/\epsilon)$ label complexity is made possible by the presence of $\mathbb{P}'$ and is therefore only achievable if the amount of unlabeled data is $\Omega(1/\tau)$, which could potentially be enormous. With less unlabeled data, the usual $\Omega(1/\epsilon)$ label complexity applies.

2.2 Basic definitions

The sample complexity of supervised learning is commonly expressed as a function of the error rate ϵ and the underlying distribution $\mathbb{P}$. As the previous three examples demonstrate, for active learning it is also important to take into account the target hypothesis and the amount of unlabeled data. The main goal of this paper is to present one particular formalism by which this may be accomplished.

Let $\mathcal{X}$ be an instance space with underlying distribution $\mathbb{P}$. Let $\mathcal{H}$ be the hypothesis class, a set of functions from $\mathcal{X}$ to $\{0, 1\}$ whose VC dimension is $d < \infty$.

Our analysis is for a non-Bayesian setting, with no measure (prior) on the space $\mathcal{H}$. In the absence of such a measure, there is no natural notion of the "volume" of the current version space. However, the distribution $\mathbb{P}$ does induce a natural distance function on $\mathcal{H}$, a pseudometric:

$$d(h, h') = \mathbb{P}\{x : h(x) \neq h'(x)\}.$$

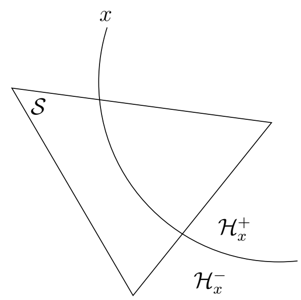
We can likewise define the notion of neighborhood: $B(h, r) = \{h' \in \mathcal{H} : d(h, h') \leq r\}$.

We will be dealing with a *separable* learning scenario, in which all labels correspond perfectly to some concept $h^* \in \mathcal{H}$, and the goal is to find $h \in \mathcal{H}$ such that $d(h^*, h) \leq \epsilon$. To do this, it is sufficient to whittle down the version space to the point where it has diameter at most ϵ, and to then return any of the remaining hypotheses. Likewise, if the diameter of the current version space is more than ϵ then any hypothesis chosen from it will have error more than $\epsilon/2$ with respect to the worst-case target. Thus, in a non-Bayesian setting, active learning is about *reducing the diameter* of the version space.

If our current version space is $\mathcal{S} \subset \mathcal{H}$, how can we quantify the amount by which a point $x \in \mathcal{X}$ reduces its diameter? Let $\mathcal{H}_x^+$ denote the classifiers that assign x a value of 1:

$$\mathcal{H}_x^+ = \{h \in \mathcal{H} : h(x) = 1\}$$

and let $\mathcal{H}_x^-$ be the remainder, which assign it a value of 0. We can think of x as a cut through hypothesis space:



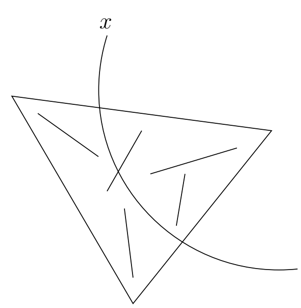
In this example, x is clearly helpful, but it doesn't reduce the diameter of $\mathcal{S}$. And we cannot say that it reduces the *average* distance between hypotheses, since again there is no measure on $\mathcal{H}$. What x seems to be doing is to reduce the diameter in a certain "direction". Is there some notion in arbitrary metric spaces which captures this intuition?

Consider any finite $Q \subset \binom{\mathcal{H}}{2}$. We will think of an element $\{h, h'\} \in Q$ as an *edge* between *vertices* h and h'. For us, each such edge will represent a pair of hypotheses which need to be distinguished from one another: that is, they are relatively far apart, so there is no way to achieve the target accuracy if both of them remain in the version space. We would hope that for any finite set of edges Q, there are queries that will remove a substantial fraction of them.

To this end, a point $x \in \mathcal{X}$ is said to *ρ-split* Q if its label is guaranteed to reduce the number of edges by a fraction $\rho > 0$, that is, if:

$$\max\left\{\left|Q \cap \binom{\mathcal{H}_x^+}{2}\right|, \left|Q \cap \binom{\mathcal{H}_x^-}{2}\right|\right\} \le (1-\rho)|Q|.$$

For instance, in the figure below, the edges are 3/5-split by x.



If our target accuracy is ϵ, we only care about edges of length more than ϵ. So define

$$Q_\epsilon = \{\{h, h'\} \in Q : d(h, h') > \epsilon\}.$$

Finally, we say that a subset of hypotheses $\mathcal{S} \subset \mathcal{H}$ is *(ρ, ϵ, τ)-splittable* if for all finite edge-sets $Q \subset \binom{\mathcal{S}}{2}$,

$$\mathbb{P}\{x : x\ \rho\text{-splits } Q_\epsilon\} \ge \tau.$$

Paraphrasing, at least a τ fraction of the distribution $\mathbb{P}$ is useful for splitting $\mathcal{S}$.[2] This τ gives a sense of how many unlabeled samples are needed. If τ is miniscule, then there are good points to query, but these points will emerge only in an enormous pool of unlabeled data.

It will soon transpire that the parameters ρ, τ play roughly the following roles:

$$\#\text{ labels needed} \propto 1/\rho, \quad \#\text{ unlabeled points needed} \propto 1/\tau$$

[2]Whenever an edge of length $l \ge \epsilon$ can be constructed in $\mathcal{S}$, then by taking Q to consist solely of this edge, we see that $\tau \le l$. Thus we typically expect τ to be at most about ϵ, although of course it might be a good deal smaller than this.

A first step towards understanding them is to establish a trivial lower bound on ρ.

Lemma 1 *Pick any $0 < \alpha, \epsilon < 1$, and any set $\mathcal{S}$. Then $\mathcal{S}$ is $((1-\alpha)\epsilon, \epsilon, \alpha\epsilon)$-splittable.*

Proof. Pick any finite edge-set $Q \subset \binom{\mathcal{S}}{2}$. Let Z denote the number of edges of Q_ϵ cut by a point x chosen at random from $\mathbb{P}$. Since the edges have length at least ϵ, this x has at least an ϵ chance of cutting any of them, whereby $\mathbb{E}Z \geq \epsilon|Q_\epsilon|$. Now,

$$\epsilon|Q_\epsilon| \;\leq\; \mathbb{E}Z \;\leq\; \mathbb{P}(Z \geq (1-\alpha)\epsilon|Q_\epsilon|) \cdot |Q_\epsilon| \;+\; (1-\alpha)\epsilon|Q_\epsilon|,$$

which upon rearrangement reads $\mathbb{P}(Z \geq (1-\alpha)\epsilon|Q_\epsilon|) \geq \alpha\epsilon$, as claimed. ∎

Thus, ρ is always $\Omega(\epsilon)$; but of course, we hope for a much larger value.

We will now see that the splitting index roughly characterizes the sample complexity of active learning.

2.3 Lower bound

We start by showing that if some region of the hypothesis space has a low splitting index, then it must contain hypotheses which are not amenable to active learning.

Theorem 2 *Fix any hypothesis space $\mathcal{H}$ and distribution $\mathbb{P}$. Suppose that for some $0 < \rho, \epsilon < 1$ and some $0 < \tau < 1/2$, set $\mathcal{S} \subset \mathcal{H}$ is not (ρ, ϵ, τ)-splittable. Then any active learning strategy which, with probability $> 3/4$ (taken over the random sampling of data), achieves an accuracy of $\epsilon/2$ on all target hypotheses in $\mathcal{S}$, must either draw $\geq 1/\tau$ unlabeled samples, or must request $\geq 1/\rho$ labels.*

Proof. Let $Q_\epsilon \subset \binom{\mathcal{S}}{2}$ be the finite set of edges of length $> \epsilon$ which defies splittability, and let $\mathcal{V} \subset \mathcal{S}$ be its vertex set:

$$\mathcal{V} = \{h : \{h, h'\} \in Q_\epsilon \text{ for some } h' \in \mathcal{H}\}.$$

We'll show that in order to distinguish between hypotheses in $\mathcal{V}$, either $1/\tau$ unlabeled samples or $1/\rho$ queries are needed.

So pick less than $1/\tau$ unlabeled samples. With probability at least $(1-\tau)^{1/\tau} \geq 1/4$, none of these points ρ-splits Q_ϵ; put differently, each of these potential queries has a bad outcome ($+$ or $-$) in which less than $\rho|Q_\epsilon|$ edges are eliminated. In this case there must be a target hypothesis in $\mathcal{V}$ for which at least $1/\rho$ labels are required. ∎

In our examples, we will apply this lower bound through the following simple corollary.

Corollary 3 *Suppose that in some neighborhood $B(h_0, \Delta)$, there are hypotheses $h_1, \ldots, h_N$ such that:*

1. *$d(h_0, h_i) > \epsilon$ for all i; and*
2. *the "disagree sets" $\{x : h_0(x) \neq h_i(x)\}$ are disjoint for different i.*

Then for any $\tau > 0$ and any $\rho > 1/N$, the set $B(h_0, \Delta)$ is not (ρ, ϵ, τ)-splittable. Any active learning scheme which achieves an accuracy of $\epsilon/2$ on all of $B(h_0, \Delta)$ must use at least N labels for some of the target hypotheses, no matter how much unlabeled data is available.

In this case, the distance metric on $h_0, h_1, \ldots, h_N$ can accurately be depicted as a *star* with h_0 at the center and with spokes leading to each h_i. Each query only cuts off one spoke, so N queries are needed.

2.4 Upper bound

We now show a loosely matching upper bound on sample complexity, via a simple algorithm which repeatedly halves the diameter of the remaining version space. For some ϵ_0 less than half the target error rate ϵ, it starts with an ϵ_0-cover of $\mathcal{H}$: that is, a set of hypotheses $S_0 \subset \mathcal{H}$ such that any $h \in \mathcal{H}$ is within distance ϵ_0 of S_0. It is well-known that there exists such an S_0 of size $\leq 2((2e/\epsilon_0)\ln(2e/\epsilon_0))^d$ (see, for instance, Theorem

5 of [8]), which in turn is at most $1/\epsilon_0^{2d}$ if $\epsilon_0 \leq 1/(32e)$. The ϵ_0-cover serves as a surrogate for the hypothesis class – for instance, the final hypothesis is chosen from it.

The algorithm below is hopelessly intractable and is not intended for deployment. Its purpose is merely to demonstrate the upper bound.

Let S_0 be an ϵ_0-cover of $\mathcal{H}$
for $t = 1, 2, \ldots, T = \lg 2/\epsilon$:
  $S_t = \text{split}(S_{t-1}, 1/2^t)$
return any $h \in S_T$

Each iteration of the inner loop reduces the diameter of the version space by a factor of two, as per the following *split* procedure.

function split(S, Δ)
Let $Q_0 = \{\{h, h'\} \in \binom{S}{2} : d(h, h') > \Delta\}$
Repeat for $t = 0, 1, 2, \ldots$:
  Draw m unlabeled points $x_{t1}, \ldots, x_{tm}$
  Find the x_{ti} which maximally splits Q_t
  Ask for its label
  Let Q_{t+1} be the remaining edges
until $Q_{t+1} = \emptyset$
return remaining hypotheses in S

We first show that for appropriately large m, *split* works as required. Clearly, upon termination the remaining hypotheses in S have diameter at most Δ. We need to bound the total number of queries made.

Lemma 4 *Suppose set $S \subset \mathcal{H}$ is (ρ, Δ, τ)-splittable. Define Q_0 as in the split procedure. Then with probability at least $1 - (1/\rho)(\ln |Q_0|)e^{-m\tau}$, split$(S, \Delta)$ will terminate after making at most $(1/\rho)\ln |Q_0|$ queries.*

Proof. Pick any $t' > 0$. Taking probabilities over the random samples of unlabeled points,

$$\begin{aligned}\mathbb{P}(\text{for some } t < t', \text{ there is no } x_{ti} \text{ which } \rho\text{-splits } Q_t) &\leq \sum_{t=0}^{t'-1} \mathbb{P}(\text{there is no } x_{ti} \text{ which } \rho\text{-splits } Q_t) \\ &\leq t' \cdot (1-\tau)^m \leq t' e^{-m\tau}.\end{aligned}$$

If this bad event does not occur, then at time t', the number of remaining edges is $|Q_{t'}| \leq |Q_0|(1-\rho)^{t'} < |Q_0|e^{-\rho t'}$. For $t' = (1/\rho)\ln |Q_0|$, this is < 1. ∎

We now spell out the conditions needed for all calls to *split* to succeed.

Lemma 5 *Suppose that the S_t are $(\rho, 1/2^{t+1}, \tau)$-splittable for all $t = 0, 1, \ldots, \lg 2/\epsilon - 1$. Then with probability at least $1 - (1/\rho)(\ln |S_0|^2)(\lg 2/\epsilon)e^{-m\tau}$, the total number of unlabeled points drawn will be*

$$M \leq \frac{m}{\rho}\left(\ln |S_0|^2\right)\left(\lg \frac{2}{\epsilon}\right),$$

and the total number of labels requested will be at most $(1/\rho)(\ln |S_0|^2)(\lg 2/\epsilon)$.

Proof. On any call to *split*, $|Q_0| \leq |S_0|^2$. The rest follows by applying the previous lemma, and summing over all calls to *split*. ∎

At the end, what remains of the starting ϵ_0-cover S_0 is a subset S_T of diameter at most $\epsilon/2$. We need to confirm that this subset is not empty – specifically, that it contains the element of S_0 closest to the target hypothesis.

Lemma 6 *Suppose the target hypothesis is some h^* and that its closest element in S_0 is h_0. If at most M unlabeled points are drawn, then the probability that $h_0 \in S_T$ is at least $1 - M\epsilon_0$.*

Proof. The probability that a randomly chosen unlabeled point separates h^* from h_0 is $d(h^*, h_0) \leq \epsilon_0$. Therefore, with probability at least $1 - M\epsilon_0$, none of the unlabeled points distinguishes between h^* and h_0 and so none of the queries can either. ∎

If the bad event in this lemma doesn't occur, then at termination S_T has diameter at most $\epsilon/2$ and contains h_0. Therefore, any $h \in S_T$ is a good approximation to h^*:

$$d(h^*, h) \;\leq\; d(h^*, h_0) + d(h_0, h) \;\leq\; \epsilon_0 + \epsilon/2 \;<\; \epsilon$$

(assuming $\epsilon_0 \leq \epsilon/2$). Setting m, ϵ_0 appropriately yields a bound on sample complexity which takes into account both locality and scale.

Theorem 7 *Let the target hypothesis be some $h^* \in \mathcal{H}$. Pick any target accuracy $\epsilon > 0$ and confidence level $\delta > 0$. Suppose $B(h^*, 4\Delta)$ is (ρ, Δ, τ)-splittable for all $\Delta \geq \epsilon/2$. Then there is an appropriate choice of ϵ_0 and m for which, with probability at least $1 - \delta$, the algorithm will return a hypothesis of error at most ϵ, with the following sample complexity:*

$$\#\ \textit{unlabeled points} \;\leq\; \tilde{O}\left(\frac{d}{\rho\tau} \cdot \log\frac{1}{\epsilon} \cdot \log\frac{1}{\epsilon\tau}\right), \qquad \#\ \textit{labels} \;\leq\; \tilde{O}\left(\frac{d}{\rho} \cdot \log\frac{1}{\epsilon} \cdot \log\frac{1}{\epsilon\tau}\right).$$

Proof. Let h_0 be the closest representative to h^* in S_0. If h_0 is never eliminated and if $\epsilon_0 \leq \epsilon/2$, then

$$S_t \subset B(h_0, 1/2^t) \subset B(h^*, 1/2^t + \epsilon_0) \subset B(h^*, 1/2^{t-1})$$

for all $t \leq T - 1$. Therefore each S_t is $(\rho, 1/2^{t+1}, \tau)$-splittable. The rest follows from the previous lemmas, by choosing

$$\frac{1}{\epsilon_0} \;=\; \max\left\{\frac{2}{\epsilon},\ \tilde{\Theta}\left(\frac{d}{\rho\tau\delta} \cdot \log\frac{1}{\epsilon} \cdot \log\frac{1}{\tau}\right)\right\}$$

and $m = \tilde{O}(1/\tau)$. ∎

This theorem makes it possible to derive label complexity bounds which are fine-tuned to the specific target hypothesis. At the same time, it is quite loose in that no attempt has been made to optimize logarithmic factors.

In a hypothesis class with a wide range of label complexities, such as our earlier two-dimensional example, it may not be the case that all neighborhoods $B(h, r)$ are splittable with constant ρ. Thus even for a good target hypothesis h^*, the search process might not be terribly efficient until the version space has been narrowed down somewhat. This could happen in a two-stage process:

1. A relatively inefficient phase in which, for instance, query points are chosen at random until the version space is contained within some $B(h^*, \Delta_0)$.
2. Efficient search down to $B(h^*, \epsilon)$.

Correspondingly, when deriving splitting indices, we will sometimes just show that the sets $B(h^*, 4\Delta)$ are (ρ, Δ, τ)-splittable for Δ *small enough*, that is, for $\Delta \leq \Delta_0$. However, the theorem above also requires splitting indices for larger values of Δ, and for these, we can appeal to Lemma 1. The resulting upper bound on label complexity will then be proportional to $(d/\rho + d/\Delta_0)$ rather than just d/ρ.

3 Examples

3.1 Linear separators in $\mathbb{R}^1$

Returning to our first example, let the instance space $\mathcal{X}$ be the real line, and let $\mathcal{H} = \{h_w : w \in \mathbb{R}\}$, where each h_w is a threshold function: $h_w(x) = \mathbf{1}(x \geq w)$. Suppose $\mathbb{P}$ is the underlying distribution on $\mathcal{X}$; for simplicity we'll assume it's a density, although the discussion can easily be generalized. It induces a distance measure on $\mathcal{H}$,

$$d(h_w, h_{w'}) \;=\; \mathbb{P}\{x : h_w(x) \neq h_{w'}(x)\} \;=\; \mathbb{P}\{x : w \leq x < w'\} \;=\; \mathbb{P}[w, w')$$

(assuming $w' \geq w$). The splitting indices are easily obtained and reflect the fact that active-learning $\mathcal{H}$ is just a binary search.

Claim 8 *$\mathcal{H}$ is $(\rho = 1/2, \epsilon, \epsilon)$-splittable for any $\epsilon > 0$.*

Proof. Pick any accuracy $\epsilon > 0$ and consider any finite set of edges $Q = \{\{h_{w_i}, h_{w'_i}\} : i = 1, \ldots, n\}$, where without loss of generality (1) the w_i are in nondecreasing order, (2) each $w_i \leq w'_i$, and (3) each edge has length greater than ϵ: $\mathbb{P}[w_i, w'_i) > \epsilon$. Pick w so that $\mathbb{P}[w_{\lceil n/2 \rceil}, w] = \epsilon$. It is not hard to see that any query $x \in [w_{\lceil n/2 \rceil}, w]$ must eliminate at least half the edges in Q, because we can be sure that

$$w_1, w_2, \ldots, w_{\lceil n/2 \rceil} \;\;\leq\;\; x \;\;<\;\; w'_{\lceil n/2 \rceil}, \ldots w'_n.$$

If x is positive then the hypotheses on the right-hand side are invalidated; if x is negative, it is the left-hand batch. Either way, at least half the edges are removed. ∎

3.2 Intervals on the line

The next case we consider is very similar to our earlier example of 2-d linear separators. The hypotheses correspond to intervals on the real line: $\mathcal{X} = \mathbb{R}$ and $\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}\}$, where

$$h_{a,b}(x) = \mathbf{1}(a \leq x \leq b).$$

Once again assume $\mathbb{P}$ is a density. The distance measure it induces is

$$d(h_{a,b}, h_{a',b'}) \;=\; \mathbb{P}\{x : x \in [a,b] \cup [a',b'], x \notin [a,b] \cap [a',b']\} \;=\; \mathbb{P}([a,b]\Delta[a',b']),$$

where $S\Delta T$ denotes the symmetric difference $(S \cup T) \setminus (S \cap T)$.

It turns out that even in this very simple class, some target hypotheses are much easier to active-learn than others. Let's start by looking at the troublesome ones.

Hypotheses that are not amenable to active-learning. For any $\epsilon' > \epsilon$, choose $\lfloor 1/\epsilon' \rfloor$ disjoint closed intervals on the real line, each with probability mass $> \epsilon$, and let $\{h_i : i = 1, ..., \lfloor 1/\epsilon' \rfloor\}$ denote the hypotheses which take value 1 on these intervals. Let h_0 be the concept which is everywhere-zero. Then the h_i exactly satisfy the conditions of Corollary 3; their star-shaped configuration forces $\rho \leq 1/\lfloor 1/\epsilon' \rfloor$, and active learning doesn't help much in choosing amongst them.

Hypotheses that are amenable to active learning. The bad hypotheses we've seen have intervals with very small probability mass. We'll now show that the larger concepts are not so bad.

Claim 9 *Pick any $\epsilon > 0$. If $\mathbb{P}[a,b] > 4\epsilon$ then the set $B(h_{a,b}, 4\epsilon)$ is $(1/32, \epsilon, \epsilon/8)$-splittable.* [3]

Proof. Pick any $\epsilon > 0$ and any $h_{a,b}$ such that $\mathbb{P}[a,b] > 4\epsilon$. Consider a finite set of edges Q whose endpoints are in $B(h_{a,b}, 4\epsilon)$ and which all have length $> \epsilon$.

In the figure below, all lengths denote probability masses. Any concept in $B(h_{a,b}, 4\epsilon)$ — more precisely, the interval corresponding to that concept — must lie within the outer box

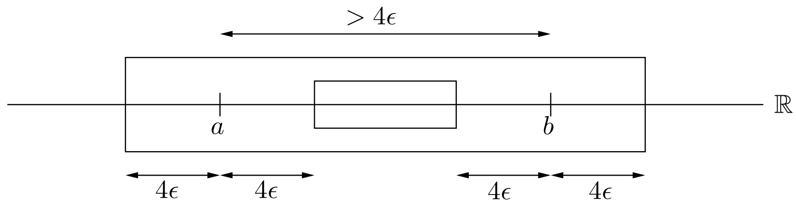
$$\big[\inf\{x : \mathbb{P}[x,a] \leq 4\epsilon\},\; \sup\{x : \mathbb{P}[b,x] \leq 4\epsilon\}\big]$$

[3] A different argument shows $B(h_{a,b}, 4\epsilon)$ is $(1/4, \epsilon, \epsilon/2)$-splittable if $\mathbb{P}[a,b] > 8\epsilon$. We have presented the weaker bound because it also demonstrates that in this case, a good strategy is simply to choose a query point at random from the *region of uncertainty*, as defined in [3]: the portion of $\mathcal{X}$ on which there is still some disagreement in the remaining version space.

and must contain the inner box

$$\big[\sup\{x : \mathbb{P}[a, x] \leq 4\epsilon\},\ \inf\{x : \mathbb{P}[x, b] \leq 4\epsilon\}\big].$$

Of course this inner box might be empty.



Any edge $\{h_{a',b'}, h_{a'',b''}\} \in Q$ has length more than ϵ, so $[a', b']\Delta[a'', b'']$ (which is either a single interval or a union of two intervals) has total length $> \epsilon$ and lies between the inner and outer boxes.

Now pick x at random from the distribution $\mathbb{P}$ restricted to the space between the two boxes. This space has probability mass at most 16ϵ, of which at least ϵ is occupied by $[a', b']\Delta[a'', b'']$. Therefore the probability that x separates $h_{a',b'}$ from $h_{a'',b''}$ is at least $1/16$.

Now let's look at all the edges in Q. The expected number of edges split by our x is at least $|Q|/16$, and therefore the probability that more than $|Q|/32$ edges are split is at least $1/32$. So, for x chosen at random from the *entire* distribution $\mathbb{P}$,

$$\mathbb{P}\{x : x\ \tfrac{1}{32}\text{-splits } Q\} \ \geq \ \frac{1}{32} \cdot \mathbb{P}\{x : x \text{ lies between the two boxes}\} \ \geq \ \frac{4\epsilon}{32} \ = \ \frac{\epsilon}{8},$$

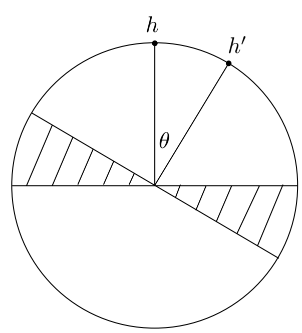
as claimed. ▮

This means that for any target hypothesis $h_{a,b}$, efficient active learning is possible once the version space has been whittled down to $B(h_{a,b}, \mathbb{P}[a, b]/4)$. And, as in the discussion after Theorem 7, the initial whittling phase can be managed by random sampling, using $\tilde{O}(1/\mathbb{P}[a, b])$ labels: not too bad when $\mathbb{P}[a, b]$ is large.

3.3 Linear separators under the uniform distribution

The most encouraging theoretical results for active learning to date have been for learning homogeneous (through the origin) linear separators with respect to data drawn from the uniform distribution over the unit sphere in $\mathbb{R}^d$ [6, 5]. We now give a simple analysis showing that in this case, $\mathcal{H}$ is $(\rho = 1/4, \epsilon, \Omega(\epsilon))$-splittable for any $\epsilon > 0$.

Without loss of generality, each hypothesis can be thought of as a unit vector in $\mathbb{R}^d$. For two such hypotheses h, h', consider the plane defined by them:



A point x is classified differently by h and h' precisely if the projection of x into this plane lies in the shaded area. Thus the pseudometric on $\mathcal{H}$ has an intuitive geometric interpretation in terms of angles:

$$d(h, h') \ = \ \mathbb{P}\{x : h(x) \neq h'(x)\} \ = \ \frac{\theta}{\pi}.$$

The $\mathbb{R}^2$ case is straightforward to analyze. We deal with higher dimensions by reducing to the 2-d case, via the method of random projection.

Theorem 10 *There is a constant $c > 0$ such that for any dimension $d \geq 2$, if*

- *$\mathcal{H}$ is the class of homogeneous linear separators in $\mathbb{R}^d$, and*
- *$\mathbb{P}$ is the uniform distribution over the surface of the unit sphere,*

then $\mathcal{H}$ is $(1/4, \epsilon, c\epsilon)$-splittable for all $\epsilon > 0$.

Proof. Choose a finite set Q of edges of length $> \epsilon$ (measured according to the underlying pseudometric). We will show that a point x drawn randomly from $\mathbb{P}$ has an $\Omega(\epsilon)$ chance of eliminating at least a quarter of these edges.

It is convenient to think of x as being chosen in two steps:

1. Pick a random 2-d subspace — that is, a plane — of $\mathbb{R}^d$.
2. Choose x uniformly at random from the unit circle in this plane.

We will see that for most planes, an $\Omega(\epsilon)$ fraction of the points on the corresponding unit circle 1/4-split Q.

Once a random plane has been chosen, the fate of Q (that is, the extent to which it gets split by x) is determined completely by the projection of its vertex set into this plane. So pick any edge $\{h, h'\} \in Q$, and let's look at the projections $\tilde{h}, \tilde{h}' \in \mathbb{R}^2$ of its endpoints. Since $d(h, h') > \epsilon$ we know the angle between $h, h' \in \mathbb{R}^d$ is some $\theta > \epsilon\pi$. By the angle-preserving property of random projection (Lemma 11), there is at least a 3/4 probability that $\tilde{h}$ and $\tilde{h}'$ are separated by an angle $\geq c_0\theta > c_0\epsilon\pi$, where $c_0 > 0$ is some absolute constant. Call $\{h, h'\}$ a *good edge* if this event occurs.

For a random choice of plane, the expected number of good edges is $\geq \frac{3}{4}|Q|$ and thus the probability of having less than $\frac{1}{2}|Q|$ good edges is $\leq 1/2$.

Now consider any plane which has $\geq \frac{1}{2}|Q|$ good edges. We will finish up by showing that at least a $c_0\epsilon$ fraction of the points x on the corresponding unit circle eliminate half or more of the good edges, and thus 1/4-split Q. From this the theorem follows, with $c = c_0/2$.

Let the good edges in the plane be $\{\tilde{h}_1, \tilde{h}'_1\}, \ldots, \{\tilde{h}_n, \tilde{h}'_n\}$, each a projection of an edge from the original high-dimensional space. Assume without loss of generality that the counterclockwise angle from each $\tilde{h}_i$ to $\tilde{h}'_i$ is $\leq \pi$, and of course we know it is $> c_0\epsilon\pi$.

Choose a point x_0 on the unit circle such that $\tilde{h}_i \cdot x_0 \leq 0$ for at least $\lceil n/2 \rceil$ of the $\tilde{h}_i$ and $\tilde{h}_i \cdot x_0 \geq 0$ for at least $\lceil n/2 \rceil$ of them. Querying x_0 will eliminate at least half the $\tilde{h}_i$'s and thus half the good edges. But it is also enough to query any point x whose counterclockwise angle from x_0 is in the range $[0, c_0\epsilon\pi]$ or symmetrically, $[\pi, \pi + c_0\epsilon\pi]$: if querying x_0 eliminates $\tilde{h}_i$, then querying a point in this range is certain to eliminate either $\tilde{h}_i$ or $\tilde{h}'_i$. Thus, a point chosen at random from the unit circle will 1/2-split the good edges (and thus 1/4-split Q) with probability $\geq c_0\epsilon$. ∎

The proof above relies crucially upon the following angle-preserving property of random projections.

Lemma 11 *For any $d \geq 2$, let x, y be vectors in $\mathbb{R}^d$ separated by an angle of $\theta \in [0, \pi]$. Let $\tilde{x}, \tilde{y}$ be their projections into a randomly chosen two-dimensional subspace. There is an absolute constant $c_0 > 0$ (which does not depend on d) such that with probability at least $3/4$ over the choice of subspace, the angle between $\tilde{x}$ and $\tilde{y}$ is at least $c_0\theta$.*

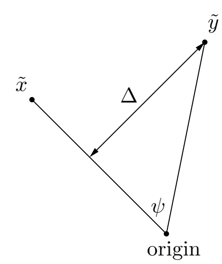
Proof. Assume for convenience that x and y have unit length. Since the subspace into which they are projected is chosen at random, the particular basis in which x and y are initially specified is irrelevant. Assume without loss of generality that

$$x = (1, 0, \ldots, 0) \quad \text{and} \quad y = (\cos\theta, \sin\theta, 0, \ldots, 0).$$

Say the subspace is chosen by selecting orthogonal unit vectors $u = (u_1, \ldots, u_d), v = (v_1, \ldots, v_d) \in \mathbb{R}^d$ at random. The projections of x and y into this plane are $\tilde{x} = t_1$ and $\tilde{y} = t_1 \cos\theta + t_2 \sin\theta$, where

$$t_1 = (u_1, v_1) \quad \text{and} \quad t_2 = (u_2, v_2).$$

We need to lower-bound the angle $\psi \in [0, \pi]$ between $\tilde{x}$ and $\tilde{y}$. If $\psi \geq \pi/2$ we are done (the constant c_0 is smaller than $1/2$), so assume it is acute:



In (the subsequent) Lemma 12, it is shown that for some constants $c_1, c_2 > 0$, there is at least a $3/4$ probability that (1) t_2 has a component orthogonal to t_1 of magnitude at least $c_1/\sqrt{d}$, and vice versa; and (2) $\|t_1\|, \|t_2\| \leq c_2/\sqrt{d}$. Under these conditions, the length Δ in the figure above is at least

$$\Delta \;\geq\; \frac{c_1}{\sqrt{d}} \sin\theta,$$

and

$$\|\tilde{y}\| \;=\; \|t_1 \cos\theta + t_2 \sin\theta\| \;\leq\; \|t_1\| + \|t_2\| \;\leq\; \frac{2c_2}{\sqrt{d}},$$

whereupon

$$\sin\psi \;=\; \frac{\Delta}{\|\tilde{y}\|} \;\geq\; \frac{c_1}{2c_2} \sin\theta.$$

We now consider two cases.

Case one: $\theta \leq \pi/2$

In this case, we can use the inequality $2z/\pi \leq \sin z \leq z$ for $z \in [0, \pi/2]$ to get

$$\psi \;\geq\; \sin\psi \;\geq\; \frac{c_1}{2c_2} \sin\theta \;\geq\; \frac{c_1}{\pi c_2} \theta.$$

Case two: $\theta > \pi/2$

Since $\psi < \pi/2$, we know $\tilde{y} \cdot \tilde{x} = \tilde{y} \cdot t_1 > 0$. Rewriting this,

$$\|t_1\|^2 \cos\theta + (t_1 \cdot t_2) \sin\theta \;>\; 0.$$

Here the cosine is negative and the sine is positive, so $(t_1 \cdot t_2) > 0$ and

$$\sin\theta \;\geq\; \frac{\|t_1\|^2}{(t_1 \cdot t_2)} \cdot (-\cos\theta) \;\geq\; \frac{\|t_1\|^2}{\|t_1\| \cdot \|t_2\|} \cdot (-\cos\theta) \;=\; \frac{\|t_1\|}{\|t_2\|} \cdot (-\cos\theta) \;\geq\; \frac{c_1/\sqrt{d}}{c_2/\sqrt{d}} \cdot (-\cos\theta).$$

For the last inequality we lower-bound the length of t_1 by the magnitude of its component orthogonal to t_2. At this point either $\theta \geq 3\pi/4$, in which case $(-\cos\theta) \geq 1/\sqrt{2}$ and thus $\sin\theta \geq c_1/c_2\sqrt{2}$; or else $\pi/2 < \theta < 3\pi/4$, in which case $\sin\theta > 1/\sqrt{2}$. Either way,

$$\psi \;\geq\; \sin\psi \;\geq\; \frac{c_1}{2c_2} \sin\theta \;\geq\; \frac{c_1}{2c_2} \min\left\{\frac{c_1}{c_2\sqrt{2}}, \frac{1}{\sqrt{2}}\right\} \;\geq\; \frac{c_1}{2\pi c_2} \min\left\{\frac{c_1}{c_2\sqrt{2}}, \frac{1}{\sqrt{2}}\right\} \cdot \theta.$$

Setting

$$c_0 \;=\; \min\left\{\frac{c_1}{\pi c_2}, \frac{c_1^2}{2\sqrt{2}\pi c_2^2}, \frac{c_1}{2\sqrt{2}\pi c_2}\right\}$$

gives the lemma. ▮

Lemma 12 *Let $u = (u_1, u_2, \ldots, u_d)$ and $v = (v_1, v_2, \ldots, v_d)$ be random unit vectors in $\mathbb{R}^d$ which are orthogonal to each other but are otherwise chosen independently. Then for some absolute constants $c_1, c_2 > 0$, with probability at least $3/4$ over the choice of u and v, the two-dimensional vectors*

$$t_1 = (u_1, v_1) \quad \textit{and} \quad t_2 = (u_2, v_2)$$

satisfy the following properties:

1. *t_2 has a component orthogonal to t_1 of magnitude at least $c_1/\sqrt{d}$; and vice versa.*
2. *$\|t_1\|, \|t_2\| \le c_2/\sqrt{d}$.*

Proof. We select u, v in a roundabout fashion, by first choosing a $d \times d$ matrix whose rows constitute a random orthonormal basis of $\mathbb{R}^d$,

$$R \;=\; \begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1d} \\ R_{21} & R_{22} & \cdots & R_{2d} \\ & & \vdots & \\ R_{d1} & R_{d2} & \cdots & R_{dd} \end{bmatrix},$$

and then setting u and v to its first and second rows, respectively. Because of the spherical symmetry of the multivariate Gaussian, R can chosen as follows:

1. Let Z be a $d \times d$ matrix of i.i.d. $N(0, 1/d)$ random variables (normal with mean zero, variance $1/d$); it is almost surely of full rank. Let Z_i denote the i^{th} column of Z.
2. Normalize each Z_i to unit length. Call the resulting vectors $\hat{Z}_i$.
3. Now orthogonalize: for each column $i = 2, 3, \ldots$, subtract from $\hat{Z}_i$ its component in $\mathrm{span}\{\hat{Z}_1, \ldots, \hat{Z}_{i-1}\}$, and renormalize to unit length.

The final matrix R has orthogonal columns: $R^T R = I_d$. But this implies also that $R^{-1} = R^T$ and that $RR^T = I_d$. Thus the *rows* of R also form a random orthonormal basis of $\mathbb{R}^d$. We will choose u, v to be the first two rows.

Letting R_i denote the i^{th} column of R, by construction

$$R_2 = \frac{Z_2/\|Z_2\| - \text{component in direction } R_1}{\text{some normalization factor} \le 1}.$$

Considering just the first two coordinates, this means that for some scalar α,

$$t_2 \;=\; (R_{12}, R_{22}) \;=\; \frac{(Z_{12}, Z_{22})/\|Z_2\| - \alpha(R_{11}, R_{21})}{\text{some normalization factor} \le 1} \;=\; \frac{(Z_{12}, Z_{22})/\|Z_2\| - \alpha t_1}{\text{some normalization factor} \le 1}.$$

Let $\hat{s} \in \mathbb{R}^2$ be a unit vector orthogonal to t_1. We need to show that t_2 has a component of significant magnitude in the direction of $\hat{s}$. By the formula just derived,

$$|t_2 \cdot \hat{s}| \;\ge\; \frac{|(Z_{12}, Z_{22}) \cdot \hat{s}|}{\|Z_2\|}.$$

Again by spherical symmetry, $(Z_{12}, Z_{22}) \cdot \hat{s}$ is distributed as $N(0, 1/d)$, and thus has magnitude $\geq 0.03/\sqrt{d}$ with probability at least 31/32. And since $\mathbb{E}\|Z_2\|^2 = 1$ we know $\|Z_2\|^2 \leq 32$ with probability at least 31/32. Thus there there is at least a 15/16 probability that

$$|t_2 \cdot \hat{s}| \;\geq\; \frac{0.03}{\sqrt{32d}} \;\stackrel{\text{def}}{=}\; \frac{c_1}{\sqrt{d}}.$$

The same applies for the component of t_1 orthogonal to t_2.

For property (2), symmetry dictates that $\mathbb{E}\|t_1\|^2 = 2/d$. So $\|t_1\|^2 \leq 32/d \stackrel{\text{def}}{=} c_2^2/d$ with probability at least 15/16; and the same holds for $\|t_2\|^2$. ▮

3.4 Linear separators under other distributions

The previous example is encouraging, but a uniform distribution of data is qualitatively quite different from what one expects in typical classifier-learning scenarios. It places most of the probability mass right at the boundary between the two classes, whereas the success of margin-based methods suggests that quite the reverse might be true in many applications.

The work of [6] addresses this issue by considering distributions $\mathbb{P}'$ over $\mathbb{R}^d$ which are within a multiplicative factor $\lambda > 1$ of the uniform distribution $\mathbb{P}$ over the surface of the unit sphere; that is, for any $A \subseteq \mathbb{R}^d$,

$$\frac{1}{\lambda}\mathbb{P}(A) \quad\leq\quad \mathbb{P}'(A) \quad\leq\quad \lambda\mathbb{P}(A).$$

For $\mathcal{H}$ = {homogeneous linear separators in $\mathbb{R}^d$}, they derive label complexity bounds for $\mathbb{P}'$ which are a multiplicative factor of $\lambda^{\Omega(1)}$ higher than those for $\mathbb{P}$.

Our bound is somewhat more reassuring: the ρ value remains the same as for $\mathbb{P}$, provided the amount of *unlabeled* data increases by a factor of λ^2. We show this by a very general lemma.

Theorem 13 *Pick a hypothesis class $\mathcal{H}$ and a distribution $\mathbb{P}$ over the input space. Suppose there is some function $f(\cdot)$ and some constant $\rho > 0$ such that $(\mathcal{H}, \mathbb{P})$ is $(\rho, \epsilon, f(\epsilon))$-splittable for all $\epsilon > 0$. Now consider any distribution $\mathbb{P}'$ within a multiplicative factor $\lambda \geq 1$ of $\mathbb{P}$. Then $(\mathcal{H}, \mathbb{P}')$ is $(\rho, \epsilon, f(\epsilon/\lambda)/\lambda)$-splittable for all $\epsilon > 0$.*

Proof. Let d and d' be the pseudometrics on $\mathcal{H}$ induced by $\mathbb{P}$ and $\mathbb{P}'$, respectively. They are closely related, since for any $h, h' \in \mathcal{H}$,

$$d'(h, h') \quad=\quad \mathbb{P}'\{x : h(x) \neq h'(x)\} \quad\leq\quad \lambda \cdot \mathbb{P}\{x : h(x) \neq h'(x)\} \quad=\quad \lambda \cdot d(h, h').$$

Pick a set of edges Q of d'-length $> \epsilon$. These same edges have d-length $> \epsilon/\lambda$ (by the formula just derived), whereupon

$$\mathbb{P}'\{x : x\ \rho\text{-splits } Q\} \quad\geq\quad \frac{1}{\lambda} \cdot \mathbb{P}\{x : x\ \rho\text{-splits } Q\} \quad\geq\quad \frac{1}{\lambda} \cdot f\left(\frac{\epsilon}{\lambda}\right),$$

and we're done. ▮

Corollary 14 *There is an absolute constant $c > 0$ such that for any dimension $d \geq 2$, if*

- *$\mathcal{H}$ is the class of homogeneous linear separators in $\mathbb{R}^d$, and*
- *the input distribution is within a multiplicative factor $\lambda \geq 1$ of the uniform distribution over the unit sphere in $\mathbb{R}^d$,*

then $\mathcal{H}$ is $(1/4, \epsilon, c\epsilon/\lambda^2)$-splittable for any $\epsilon > 0$.

4 Related work and open problems

There has been a lot of work on a related model in which the points to be queried can be synthetically constructed, rather than merely chosen from unlabeled data [1]. The expanded role of $\mathbb{P}$ in the present model makes it substantially different, although a few intuitions do carry over – for instance, Corollary 3 can be thought of as generalizing the notion of *teaching dimension* [7].

We have already mentioned [6, 4, 5]. One other technique which seems useful for active learning is to look at the unlabeled data and then place bets on certain target hypotheses, for instance those with large margin. This insight is nicely formulated in [2, 9], and it is interesting to consider how it might effectively be combined with the search strategy suggested in this paper.

Some of the positive examples in this paper have the following additional property: a random data point which intersects the version space has a good chance of $\Omega(1)$-splitting it. This permits a naive active learning strategy, earlier suggested in [3]: just pick a random point whose label you are not yet sure of. On what kinds of learning problems will this work well?

Finally, the upper and lower bounds shown in this paper differ by a multiplicative factor of $d \operatorname{poly} \log 1/\epsilon$. This is modest compared to the range in which the label complexity can lie, but there is clearly a lot of room for improvement.

Acknowledgements

I'd like to thank Yoav Freund for introducing me to this field of research; Peter Bartlett, John Langford, Adam Kalai, and Claire Monteleoni for helpful discussions; the anonymous NIPS reviewers for their detailed and perceptive feedback; and the National Science Foundation for support under grant IIS-0347646.

References

[1] D. Angluin. Queries revisited. *Proceedings of the Twelfth International Conference on Algorithmic Learning Theory*, 12–31, 2001.

[2] M.-F. Balcan and A. Blum. A PAC-style model for learning from labeled and unlabeled data. *Eighteenth Annual Conference on Learning Theory*, 2005.

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

[4] S. Dasgupta. Analysis of a greedy active learning strategy. *Advances in Neural Information Processing Systems 17*, 2004.

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

[6] Y. Freund, S. Seung, E. Shamir, and N. Tishby. Selective sampling using the query by committee algorithm. *Machine Learning Journal*, 28:133–168, 1997.

[7] S. Goldman and M. Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50(1):20–31, 1995.

[8] D. Haussler. Decision-theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

[9] J. Shawe-Taylor, P. Bartlett, R. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.