On the Splitting Index of Disjunctions

Sitan Chen

May 14, 2014

1 Introduction

This project will focus on *active learning*, the variant of the supervised PAC model in which the learner is given a pool of unlabeled examples drawn from some distribution $\mathcal{D}$ and is allowed to query for the labels of any examples it chooses from this pool. The motivation for this is that unlabeled examples are often less costly than labeled ones and the former can allow us to reduce the hypothesis space across which we need to search. Specifically, we will study the relationship between two different flavors of active learning, one relating to aggressively querying points which efficiently split the hypothesis space and the other to exploiting some regularity condition that gives an estimate of the "unlabeled error" of a hypothesis. We formalize these two notions, due to Dasgupta and Balcan/Blum respectively, in section 2. In section 3, we show a lower bound on label complexity for actively learning monotone disjunctions. In section 4, in contrast, we show that if we assume a certain large-margin regularity condition on the distribution, the savings no label complexity become essentially optimal. In section 5, we generalize the framework of section 4 to determine sufficient conditions for a regularity condition to yield an splittable hypothesis space. Throughout, we'll be working in the *realizable case*, that is, there is always an underlying concept with zero true error that we want to learn.

2 Preliminaries

2.1 Splitting Index

The motivation for defining splitting index comes from one of the earliest positive examples of an exponential reduction in the number of labeled examples needed via active learning: learning linear separators in $\mathbb{R}$. Indeed, a binary search over a set of $O(1/\epsilon)$ unlabeled examples allows us to only require $O(\log 1/\epsilon)$ labels in order to output a hypothesis with error ϵ, yielding an exponential saving on label complexity compared to the PAC model.

Dasgupta formalizes this searching strategy as follows. For an underlying distribution $\mathcal{D}$ on the instance space X, equip the hypothesis space $\mathcal{H}$ with the Hamming metric $d(h, h') = \Pr_{x \leftarrow \mathcal{D}}(h(x) \neq h'(x))$. The goal is to eliminate enough candidate hypotheses from $\mathcal{H}$ that the diameter of the resulting space is no more than ϵ.

For a given $x \in X$, define $\mathcal{H}_x^+$ to be the set of hypotheses in which x is a positive example and $\mathcal{H}_x^-$ to be the set in which x is a negative example. For a finite $Q \subset \mathcal{H}^2$, we say that $x \in X$ *ρ-splits* Q if labelling x reduces the number of pairs of hypotheses we need to distinguish in Q by at least ρ (the *splitting index*), i.e.

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

As a remark, later we will also say that $x \in X$ "splits" or "distinguishes" edge (h, h') if $h(x) \neq h'(x)$.

Finally, because we want to distinguish between hypotheses that are more than ϵ-far, for any such Q we will focus on Q_ϵ, defined to be the pairs $(h, h') \in Q$ such that $d(h, h') > \epsilon$.

Definition 1. *$S \subset \mathcal{H}$ is (ρ, ϵ, τ)-splittable if for each finite $Q \subset \mathcal{H}_x^+$,*

$$\Pr_{x \leftarrow \mathcal{D}}(x \text{ } \rho\text{-splits } Q_\epsilon) \geq \tau.$$

2.2 Compatibility Notions

Balcan/Blum [2] propose an alternative strategy of using unlabeled examples given a so-called *compatibility notion* akin to an estimate of the "error rate on unlabeled examples" to allow the learner to whittle down the version space to those with low unlabeled error.

Definition 2. *A* compatibility notion *is any function $\mathcal{C} \times X \to \{0, 1\}$. The* true unlabeled error rate $\mathrm{err}_u(h)$ *of a hypothesis h with distribution $\mathcal{D}$ is $1 - \chi(h, \mathcal{D}) := 1 - \mathbb{E}_{x \leftarrow \mathcal{D}}(\chi(h, x))$. Let $\mathcal{C}_{\mathcal{D},\chi}(\tau) = \{h \in \mathcal{C} : \mathrm{err}_u(h) \leq \tau\}$. The* empirical unlabeled error rate *and $\mathcal{C}_{S,\chi}(\tau)$ depending on a sample S rather than the distribution $\mathcal{D}$ can be defined analogously.*

Example 1. A monotone disjunction is a *two-sided disjunction h with respect to distribution* $\mathbb{D}$ if the support of $\mathbb{D}$ solely consists of points inside h or inside $[n] - h$, the hypothesis consisting of all variables which h does not contain.

If we define the compatibility notion $\chi(h, x) = I[h(x) \neq [n] - h]$, then the concept class $\mathcal{C}_n$ of two-sided disjunctions is precisely all monotone disjunctions with zero true unlabeled error with respect to this χ. We will refer to the variables in h the *positive indicators* and the variables in $[n] - h$ the *negative indicators*.

We have a straightforward active learning algorithm exploiting two-sidedness: because any two variables which appear in the same example must be the same type of indicator, if we draw the *commonality graph* G with vertices consisting of the n variables and edges connecting variables which both occur in one of the examples, then it suffices to obtain one label for each of the k connected components of G. An Occam-like argument shows that with $\frac{1}{\epsilon}\left(\ln|\mathcal{C}_n| + \ln\frac{2}{\delta}\right)$ unlabeled examples S and $\frac{1}{\epsilon}\left(\ln|\mathcal{C}_{S,\chi}(0)| + \ln\frac{2}{\delta}\right)$ labels, where $|\mathcal{C}_{S,\chi}| = 2^k$.

Later in this paper, we will look at learning two-sided disjunctions using the splitting index, which turns out to give an exponential improvement on the dependence on $\frac{1}{\epsilon}$.

Example 2. In co-training, every example presented to the learner has two pieces of information $x = (x_1, x_2)$; the hope is that these two views are redundant enough that we can decompose the underlying concept c^* into (c_1^*, c_2^*) such that $c_1^*(x_1) = c_2^*(x_2)$. If we define the compatibility notion $\chi((h_1, h_2), (x_1, x_2)) = I[h_1(x_1) = h_2(x_2)]$, in the co-training setting we aim to find a concept with low true unlabeled error with respect to this χ.

3 A Lower Bound for Monotone Disjunctions

We show that active learning cannot give exponential savings in label complexity for learning monotone disjunctions under the uniform distribution within sufficiently small error, by giving an upper bound on the splitting index. For convenience, because we will refer to concepts in this setting as sets of variables, for concepts h we will use h also to denote $|h|$.

Theorem 1. *If $\mathcal{C}_n$ is the concept class of all monotone disjunctions $\{0,1\}^n \to \{0,1\}$, then for $\epsilon = \frac{1}{2^{n/2}+1}$, $\mathcal{C}_n$ is not (ρ, ϵ, τ)-splittable for*

$$\rho = \frac{2\epsilon}{(1-2\epsilon)\log(1/2\epsilon)} + \eta', \tau = \epsilon + \eta''$$

for any constants $\eta', \eta'' > 0$.

Proof. We begin by determining the distance in general between two hypotheses:

Lemma 1. *The Hamming distance $d(h, h')$ between two hypotheses is*

$$d(h, h') = \frac{1}{2^h} + \frac{1}{2^{h'}} - \frac{2}{2^{h \cup h'}}.$$

Proof. If $h(x) \neq h'(x)$ for $x \in \{0,1\}^n$, then either x contains 1) at least one point in $h - h'$ and no points in h', or 2) at least one point in $h' - h$ and no points in h. Denoting $|h \cap h'|$ by I and $|[n] - h \cup h'|$ by R, we have

$$d(h, h') = \frac{2^R(2^{h-I} + 2^{h'-I} - 2)}{2^{R+h+h'-I}} = \frac{1}{2^h} + \frac{1}{2^{h'}} - \frac{2}{2^{h \cup h'}}.$$

□

For $\epsilon = \frac{1}{2^{n/2+1}}$, we construct a collection of edges Q that is split by the desired factor with probability $\frac{1}{2^{n/2}}$ as follows: first, for convenience call a pair of distinct hypotheses $(h, h \cup g)$ for $h, g \subseteq [n]$ the (g, h)-*edge*.

Let $\tilde{h}$ be the hypothesis consisting of variables $x_{n/2+2}, ..., x_n$. Now for every $1 \leq i \leq n/2$, for each collection $g \subseteq \{x_1, ..., x_{n/2+1}\}$ of i variables, and each variable $x \in [n] - (g \cup \tilde{h})$, include the $(g, \tilde{h} \cup \{x\})$-edge in Q. We verify that each such edge is of length at least ϵ: denoting $\tilde{h} \cup \{x\}$ by h, we have

$$d(h, h \cup g) = \frac{1}{2^h} - \frac{1}{2^{h+g}} = \frac{1}{2^{n/2}}\left(1 - \frac{1}{2^g}\right) \geq \frac{1}{2^{n/2+1}}.$$

Also add the $([n] - \tilde{h}, \tilde{h})$-edge; this certainly is at least ϵ-long.

By construction, the points which split at least one of the edges in Q are precisely all nonempty subsets of $\{x_1, ..., x_{n/2}\}$. In particular, there is exactly one point, which only splits the $([n] - \tilde{h}, \tilde{h})$ edge, namely that consisting of variables $x_1, ..., x_{n/2+1}$, and it satisfies the disjunctions in all other edges in Q. Q is of size

$$1 + \sum_{i=1}^{n/2} \binom{n/2+1}{i} \left(\frac{n}{2} + 1 - i\right) = \frac{1}{2}\left(2^{n/2} - 1\right)(n+2) = \log\left(\frac{1}{2\epsilon}\right)\left(\frac{1-2\epsilon}{2\epsilon}\right),$$

and we get the desired bound on the splitting index of $\mathcal{C}_n$. □

We now make use of the following lower bound [4]:

Proposition 1. *(Dasgupta) For hypothesis space $\mathcal{H}$, if there exist constants $0 < \rho, \epsilon < 1$ and $0 < \tau < 1/2$ such that some $S \subset \mathcal{H}$ is not (ρ, ϵ, τ)-splittable, then any active learning algorithm predicting any target hypothesis in S with confidence $3/4$ and accuracy $\epsilon/2$ requires either $1/\tau$ unlabeled examples or $1/\rho$ labeled examples.*

We conclude that any algorithm that actively learns any monotone disjunction $\{0,1\}^n \to \{0,1\}$ to within $3/4$ confidence and $O(2^{-n})$ error requires at least $1/\epsilon$ unlabeled examples or $\Omega(1/\epsilon)$ labeled examples.

4 Splitting Index of Two-Sided Disjunctions

We next show that imposing a strong regularity condition on the distribution $\mathcal{D}$, namely two-sidedness as defined in section 2, yields a high splitting index and label-efficient active learnability of monotone disjunctions.

Because two-sidedness is a condition on $\mathbb{D}$ depending on the underlying concept c^*, we'll need to modify the definition of splitting index. We say a point x splits an edge $(h, h') \in \binom{\mathcal{C}_n}{2}$ not only if $h(x) = 1$ and $h'(x) = 0$ or if $h'(x) = 1$ and $h(x) = 0$, but also if x does not exist in the support of any distribution $\mathbb{D}'$ for which h (alternatively h') is two-sided with respect to $\mathbb{D}'$, in which case we will write $h(x) = -1$. Furthermore, define the modified Hamming metric $\tilde{d}(h, h') = \Pr_{x \leftarrow \mathcal{D}}(h(x) \neq h'(x) \vee h(x) = -1 \vee h'(x) = -1)$.

We will study the case where $\mathbb{D}$ is the uniform distribution over all x such that $c^*(x) \neq [n] - c^*(x)$, i.e. all nonempty subsets of c^* and of $[n] - c^*$. For convenience, denote c^* and $[n] - c^*$ by c^+ and c^-.

Theorem 2. *If $\mathcal{C}_n$ is the concept class of all two-sided disjunctions $\{0,1\}^n \to \{0,1\}$ and the distribution is $\mathbb{D}$ as defined above with respect to some underlying two-sided disjunction c^* with k positive indicators and $n-k$ negative indicators, then for all $0 < \epsilon < \frac{1}{2}$, $\mathcal{C}_n$ is $(1/4, \epsilon, \epsilon/2)$-splittable.*

Proof. Again, we first compute the modified Hamming metric $\tilde{d}$:

Lemma 2. *For monotone disjunctions h_1, h_2,*

$$\tilde{d}(h_1, h_2) = 1 - \frac{2^{h_1 \cap h_2 \cap c^+} + 2^{h_1 \cap h_2 \cap c^-} + 2^{c^+ - h_1 \cup h_2} + 2^{c^- - h_1 \cup h_2} - 3}{2^{c^+} + 2^{c^-} - 1}.$$

Proof. We want to show, using the variables in Figure 1, that there are $2^a + 2^b + 2^g + 2^h - 3$ points in the support of $\mathbb{D}$ such that either $h_1(x) = h_2(x) = 1$ or $h_1(x) = h_2(x) = 0$. In the first case, x must be a subset of a or of b; in the second, x must be a subset of g or of h, giving the desired result. □

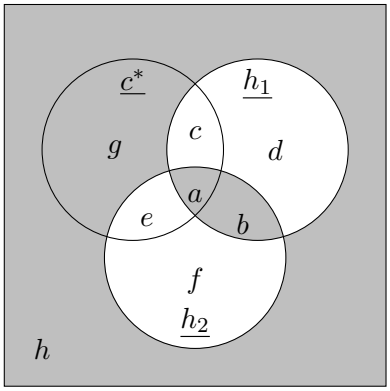


Figure 1: Relevant sets of variables in computing $\tilde{d}(h_1, h_2)$

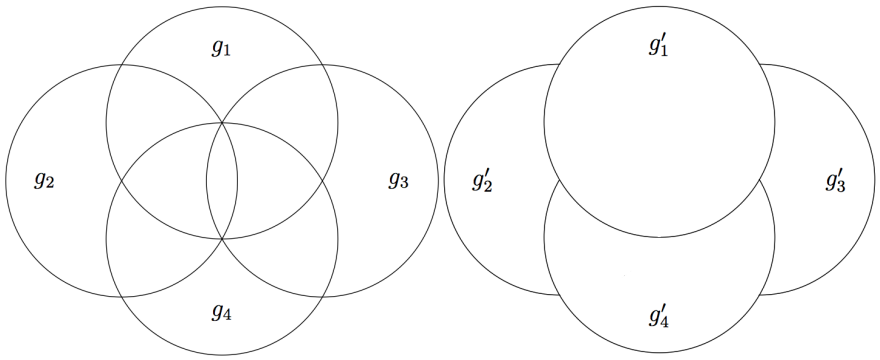
Let Q consist of the pairs (g_i, h_i) for some $\{g_i\}_{i \in [N]}$ and $\{h_i\}_{i \in [N]}$. Denote $g_i \cap c^+$ and $g_i \cap c^-$ by g_i^+ and g_i^-; do the same for h_i^+ and h_i^-.

We first consider the case in which for all i, i) $g_i \cap h_i = \emptyset$, ii) g_i^+ and g_i^- are nonempty, iii) $g_i \cap g_j = \emptyset$ for all $j \neq i$. In this case, $\tilde{d}(g_i, h_i) \geq \frac{1}{2}$ for each i.

For each $N/2 \leq i \leq N$, define S_i^+ (resp. S_i^-) to be the set of all points $x \subseteq c^+$ (resp. $x \subseteq c^-$) for which there are exactly i edges (g_j, h_j) such that x contains some variable in g_j^+ (resp. g_j^-); these are the edges split by x. We show that half of all points 1/4-split, in fact 1/2-split, Q by showing that $\sum_{i=n/2}^{N} |S_i^+| \geq 2^{k-1}$ and $\sum_{i=N/2}^{N} |S_i^-| \geq 2^{N-k-1}$. We know that

$$|S_i^+| = 2^{k-g_1^+-\cdots-g_N^+} \cdot \sum_{j_1,\ldots,j_i \in [N]} \left(2^{g_{j_1}^+} - 1\right) \cdots \left(2^{g_{j_i}^+} - 1\right).$$

Substituting a_j for each $2^{g_j^+} - 1$, we see that $\frac{1}{2^{k-g_1^+-\cdots-g_N^+}} \sum_{i=N/2}^{N} |S_i^+|$ is the sum of all terms in the polynomial $\prod_{j=1}^{N}(a_j + 1)$ of total degree at least $N/2$. The sum of the terms of degree exactly $N/2 + j$ are certainly at least that of the terms of degree exactly $N/2 - j$ for $j \geq 0$, and $\prod_{j=1}^{N}(a_j + 1) = 2^{g_1^+ + \cdots + g_N^+}$, so we get the desired bound on $\sum_{i=N/2}^{N} |S_i^+|$. The same method works for showing the bound on $\sum_{i=N/2}^{N} |S_i^-|$. In fact, it will turn out that for all but a small pathological class of Q's outlined at the end of this proof, we get a splitting index of 1/2.



(a) Q (b) Q'

Figure 2: Transformation to relax constraint iii)

Next, we show that constraint iii) can be assumed of Q without loss of generality. We are still keeping the assumption that edges consist of disjoint hypotheses, so it still holds that $\tilde{d}(g_i, h_i) \geq \frac{1}{2}$. For any such Q satisfying only i) and ii), we can construct a Q' satisfying i), ii), and iii) for which the number of points splitting Q' is at most the number splitting Q (see Figure 2): for each variable $x \in \cup_{i=1}^{N} g_i$, pick a single index j such that $x \in g_j$ and assign x to the set g_j'. Then $\cup_{i=1}^{N} g_i' = \cup_{i=1}^{N} g_i$, but $Q' = \{(g_i', h_i)\}$ satisfies conditions i), ii), and iii). For any point which 1/2-splits Q', by construction at least half of the h_i in Q contain some variable in this point, so this point at least 1/2-splits Q as well, as desired.

We now partially relax condition i): pairs of hypotheses can intersect, but not in all of c^+ or all of c^-. $\tilde{d}(h, h')$ is then still at least $\frac{1}{2}$. Letting ℓ_i denote $g_i \cap h_i$, the points which split (g_i, h_i) are all points $x \subset c^+$ (resp. $x \subseteq c^-$) containing variables in at least one of $g_i^+ - \ell_i^+$ and $h_i^+ - \ell_i^+$ (resp. $g_i^- - \ell_i^-$ and $h_i^- - \ell_i^-$), so

$$|S_i^+| = 2^{k-g_1^+ \cup h_1^+ - \cdots - g_N^+ \cup h_N^+} \cdot \sum_{j_1,\ldots,j_i \in [N]} \left(2^{g_{j_1}^+ \cup h_{j_1}^+} - 2^{\ell_{j_1}^+}\right) \cdots \left(2^{g_{j_i}^+ \cup h_{j_i}^+} - 2^{\ell_{j_i}^+}\right),$$

and by the same method, this time with polynomial $\prod_{j=1}^{N} \left(a_j + 2^{\ell_j}\right)$, we get $\sum_{i=N/2}^{N} |S_i^+| \geq 2^{k-1}$ and a similar bound for $\sum_{i=N/2}^{N} |S_i^-|$.

Thus far, we have only dealt with Q where the lengths of edges are at least $1/2$; it remains to completely remove dependence on i) and ii). Examining our expression for $\tilde{d}$ in Lemma 2, we see that if $g_i \neq h_i$, $\tilde{d}(g_i, h_i) < \frac{1}{2}$ only if 1) $g_i^+ = h_i^+ = \emptyset$, 2) $g_i^- = h_i^- = \emptyset$, 3) $g_i^+ \cap h_i^+ = c^+$, or 4) $g_i^- \cap h_i^- = c^-$. Obviously, these cases are mutually exclusive because g_i and h_i are distinct.

Say that each edge (g_i, h_i) falls in one of these categories. By our discussion above about constraint iii), say that g_i and h_i intersect nontrivially in at most one of c^+ and c^-. Let $(a_i, b_i) = (g_i - h_i, h_i - g_i)$. We first need to bound the lengths of the edges: if $a_i, b_i \subseteq c^-$ (resp. $a_i, b_i \subseteq c^+$), then $\tilde{d}(h_1, h_2) = \frac{2^{n-k} - 2^{n-k-a-b}}{2^{n-k} + 2^k - 1}$ (resp. $\tilde{d}(h_1, h_2) = \frac{2^k - 2^{k-a-b}}{2^{n-k} + 2^k - 1}$).

Without loss of generality, say that there are more i for which $a_i, b_i \subseteq c^+$, say that for each such i, $|a_i| \geq |b_i|$, and say that these i are $1, ..., m$, where $m \geq N/2$.

For a point x, if at least half of $a_1, ..., a_m$ contain variables belonging to x, then x certainly $1/4$-splits Q. Letting S_i denote points for which there exist exactly i members of $a_1, ..., a_m$ containing variables inside x, we reuse the argument above to show that $\sum_{m/2}^{m} |S_i| = 2^{k-1}$. Comparing the probability over all points in $\mathbb{D}$ that Q is $1/4$-split, which we just showed is at least $\frac{2^{k-1}}{2^k + 2^{n-k} - 1}$, to the minimum length

$$\frac{2^{k-1}}{2^k + 2^{n-k} - 1} \leq \min_{1 \leq i \leq m} \frac{2^k - 2^{k-a_i-b_i}}{2^k + 2^{n-k} - 1} \leq \frac{2^k}{2^k + 2^{n-k} - 1}$$

of the edges in Q, we conclude that $\mathcal{C}_n$ is indeed $(1/4, \epsilon, \epsilon/2)$-split. □

The main conclusion we can draw from this is that, by the following theorem of Dasgupta [4], we incur an exponential improvement on label complexity. Dasgupta makes use of the intuition that labeled examples represent comparisons in a binary search, defining a procedure for halving the diameter of the search space by successively drawing sufficiently many unlabeled points and getting the label of the one which maximally splits the space of remaining hypotheses. This yields an upper bound on label complexity that is linear in $1/\rho$ but polylogarithmic in $1/\epsilon$:

Theorem 3. *(Dasgupta) If the Hamming ball $B(c^*, 4\Delta)$ is (ρ, Δ, τ)-splittable for all $\Delta \geq \epsilon/2$, then there exists an active learning algorithm which learns c^* to within confidence δ and error ϵ given $\tilde{O}\left(\frac{d}{\rho\tau} \log \frac{1}{\epsilon} \log \frac{1}{\epsilon\tau}\right)$ unlabeled examples and $\tilde{O}\left(\frac{d}{\rho} \log \frac{1}{\epsilon} \log \frac{1}{\epsilon\tau}\right)$ labeled examples, where d is the VC dimension of the concept class.*

Remark 1. *Dasgupta uses the $\tilde{O}(\cdot)$ to conceal extra $\mathsf{polylog}(d, 1/\delta, 1/\rho, \log 1/\epsilon, \log 1/\tau)$ so that sample complexity depends polylogarithmically on $1/\delta$ as usual.*

Corollary 1. *The class $\mathcal{C}_n$ of monotone disjunctions $\{0,1\}^n \to \{0,1\}$ under the uniform distribution $\mathbb{D}$ over all x such that $c^*(x) \neq [n] - c^*(x)$ for underlying two-sided disjunction c^* is actively learnable given $\tilde{O}\left(\frac{n}{\epsilon}\right)$ unlabeled examples and $\tilde{O}\left(n \log^2 \frac{1}{\epsilon}\right)$ labeled examples.*

5 Splitting Index and Compatibility Notions

In this section we examine the relationship between compatibility notions and splitting index by abstracting away some of the properties of the regularity condition of two-sidedness which allow the hypothesis space to be cut so efficiently.

Given concept class $\mathcal{C}_n$ and instance space X, fix some compatibility notion $\chi : \mathcal{C}_n \times X \to \{0,1\}$. Denote the underlying concept by c^*, and consider a (not necessarily uniform) distribution $\mathcal{D}$ for

which the unlabeled error rate of c^* is zero, i.e. $\chi(c^*, \mathcal{D}) = 1$. Furthermore, as in the previous section, equip $\mathcal{C}_n$ with a pseudometric $\tilde{d} : \mathcal{C}_n \times \mathcal{C}_n \to \{0, 1\}$ given by

$$\tilde{d}(h, h') = \Pr_{x \leftarrow \mathcal{D}}(\chi(x, h) = 0 \vee \chi(x, h') = 0 \vee h(x) + h'(x) = 1).$$

The first property we consider is the similarity between unlabeled and true error rate. In the case of two-sided disjunctions, recall, we had that

$$\tilde{d}(c^*, h) = \frac{2^{c^+} - 2^{h^+} + 2^{c^-} - 2^{c^- - h^-}}{2^{c^+} + 2^{c^-} - 1}$$

$$1 - \chi(h, \mathcal{D}) = \frac{2^{c^+} - 2^{h^+} - 2^{c^+ - h^+} + 2^{c^-} - 2^{h^-} - 2^{c^- - h^-} + 2}{2^{c^+} + 2^{c^-} - 1}.$$

A bit of manipulation gives that $\tilde{d}(h, c^+) + \tilde{d}(h, c^-) = 2 - \frac{1}{2^{c^+} + 2^{c^-} - 1} - \chi(h)$, but given an edge (h, h') such that $\tilde{d}(h, h') > \epsilon$, the triangle inequality only tells us that $\max(1 - \chi(h, \mathcal{D}), 1 - \chi(h, \mathcal{D})) > \epsilon - 1 - \frac{1}{2^{c^+} + 2^{c^-} - 1}$. Instead, we make the following observation that we can interpret as saying unlabeled error rate is a good approximation of true error rate.

Observation 1. *There is some constant $\alpha > 0$ such that $\alpha\tilde{d}(c^*, h) < 1 - \chi(h, \mathcal{D}) < \tilde{d}(c^*, h)$.*

In general, call a compatibility notion α-*faithful* if the above observation holds. For example, the χ defined for two-sided disjunctions is 1/2-faithful.

Now consider any set of edges $Q = \{(g_i, h_i\}_{1 \le i \le N}$ such that $\tilde{d}(g_i, h_i) > \epsilon$ for all i.

Proposition 2. *For every $(g_i, h_i) \in Q$ either $\Pr_{x \leftarrow \mathcal{D}}(\chi(g_i, x) = 1) < 1 - \frac{\alpha\epsilon}{2}$ or $\Pr_{x \leftarrow \mathcal{D}}(\chi(h_i, x) = 1) < 1 - \frac{\alpha\epsilon}{2}$.*

Proof. By the triangle inequality, $\epsilon < \tilde{d}(g_i, h_i) \le \tilde{d}(g_i, c^*) + \tilde{d}(h_i, c^*) < \frac{1}{\alpha}(2 - \chi(g_i, \mathcal{D}) - \chi(h_i, \mathcal{D}))$ so that $\chi(g_i, \mathcal{D}) + \chi(h_i, \mathcal{D}) < 2 - \epsilon\alpha$, and the desired result follows by definition. □

In other words, for any edge $(g_i, h_i) \in Q$, with probability more than $\alpha\epsilon/2$, a random point drawn from $\mathcal{Q}$ splits (g_i, h_i). The assumption of α-faithfulness, however, is insufficient on its own to guarantee a high splitting index. For convenience, define $\mathrm{IC}(h)$ to be the set of $x \in X$ such that $\chi(h, x) = 0$: consider the compatibility notion for which $\mathrm{IC}(h_1) \cap \mathrm{IC}(h_2) = \emptyset$ for all $h_1 \neq h_2$.

The next observation we make about χ defined for two-sided disjunctions guarantees that $\mathrm{IC}(h_1)$ and $\mathrm{IC}(h_2)$ are actually very close.

Observation 2. *There are sets $\bigcup_{i=1}^{k} H_i = \mathcal{C}_n$ and a constant β such that for each $i \in [h]$, for all $h_1, h_2 \in H_i$, $\frac{\Pr(\mathrm{IC}(h_1) \cap \mathrm{IC}(h_2))}{\Pr(\mathrm{IC}(h_1))} \ge \beta$.*

Proof. In particular, $k = 3$, $\beta = 1/2$, and the H_1, H_2, H_3 are the sets of hypotheses h for which, respectively, 1) $h \subseteq c^+$, 2) $h \subseteq c^-$, and 3) $h \cap c^+$ and $h \cap c^-$ are both nonempty. We show the claim for 1); the other two cases follow in the exact same manner. We will use the variables from Figure 1. The points incompatible with both h_1 and h_2 are those containing either 1) a point in g and either a point in a or points in both c and e, or 2) a point in h and either a point in b or points in both c and e. Some computation shows that $|\mathrm{IC}(h_1) \cap \mathrm{IC}(h_2)| = (2^g - 1)(2^{a+c+e} - 2^c - 2^e + 1)$ while $|\mathrm{IC}(h_1)| = (2^{g+e} - 1)(2^{a+c} - 1)$. The result follows from the fact that c and e are nonempty. □

In general, call a compatibility notion (k, β)-*local* if it satisfies the above observation. Morally, locality ensures that points incompatible with some hypothesis are incompatible with many others.

The properties of faithfulness and locality turn out to be sufficient for a high splitting index.

Theorem 4. *If there exist constants α, β, k for which a compatibility notion χ is α-faithful and (k, β)-local and c^* is perfectly compatible under distribution $\mathcal{D}$, $\mathcal{C}_n$ is $\left(\frac{\beta-p}{2k}, \epsilon, \frac{\alpha p \epsilon}{2}\right)$-splittable for all $\epsilon > 0$ and $0 \leq p \leq \beta$*

Proof. We make use of the following fact:

Lemma 3. *If $\mathcal{S} = \{S_i\}_{i \in I}$ is a (finite) family of sets such that $|S_i \cap S_j| \geq \beta |S_i|$ for all $S_i, S_j \in \mathcal{S}$, then for all $0 \leq p \leq \beta$ at least $p \max_i |S_i|$ belong to at least $\lfloor (\beta - p)|I| \rfloor$ sets in $\mathcal{S}$.*

Proof. Pick the largest set of $\mathcal{S}$, call it S, fix any $0 < p < \beta$, and consider the intersections of S with the remaining sets in $\mathcal{S}$. For convenience, denote the number of sets in $\mathcal{S}$ that a point $s \in S$ belongs to by $n(s)$, and if $n(s) \geq \lfloor (\beta - p)|I| \rfloor$, we refer to it as"good." Say that fewer than p of the points in S are good. Then $\sum_{s \in S} n(s) < (p + (1-p)(\beta - p))|I||S| = (\beta - p(\beta - p))|I||S|$. However, by the assumption that $|S_i \cap S_j| \geq \beta S_i$ and that in particular, $|S \cap S_i| \geq \beta |S|$, we also get that $\sum_{s \in S} n(s) \geq \beta |I||S|$. Contradiction! □

Now take any collection of edges $Q = \{(g_i, h_i\}_{1 \leq i \leq N}$ of minimum length ϵ. By our discussion above of α-faithfulness, for every i either $\mathrm{IC}(g_i)$ or $\mathrm{IC}(h_i)$ has density of at least $\frac{\alpha \epsilon}{2}$; say that the hypothesis in each edge for which this is true is h_i.

There is some H_i among the k sets partitioning $\mathcal{C}_n$ in the sense of (k, β)-locality which contains at least $\frac{1}{k}$ of the hypotheses h_i. By approximating the densities of the $\mathrm{IC}(h)$ for $h \in H_i$ arbitrarily closely using an arbitrarily large family $\mathcal{S}$, we can apply the above lemma to conclude that for any $0 < p < \beta$, the probability over points in $\arg\max_{h \in H_i} \Pr(\mathrm{IC}(h))$ that $\Pr_{h \in H_i} x \in \mathrm{IC}(h) \geq (\beta - p)$ is at least p.

Thus, with probability at least $\frac{\alpha p \epsilon}{2}$ over points x chosen according to $\mathcal{D}$, $(\beta - p)$ of the h_i are found to have nonzero unlabeled error. In the worst case, x has a label which does not distinguish edges consisting of hypotheses compatible with x, so we can only rule out $(\beta - p)$ of the edges for sure. □

In particular, any $\mathcal{C}_n$ and $\mathcal{D}$ for which such a compatibility notion exists is $(\Omega(1), \epsilon, \Omega(\epsilon))$-splittable, and we get an exponential saving on label complexity in the active learning setting.

6 Discussion

In this paper we looked at active learning of disjunctions and in particular the effectiveness of greedily querying points which maximally cut the version space. While we showed that active learning doesn't necessarily yield significant savings on label complexity even in the case of learning monotone disjunctions, we concluded that by imposing the large-margin regularity condition of two-sidedness, it is possible to reduce the number of labels needed in the supervised setting *exponentially* using active learning. We then examined the properties of two-sidedness that were sufficient to guarantee $(\Omega(1), \epsilon, \Omega(\epsilon))$-splittability and determined that it suffices for a concept class and distribution to have a corresponding "local" compatibility notion which closely approximates distance to the underlying concept c^*.

As far as the author knows, this is the first attempt to study the relationship between the power of unlabeled examples in the sense of both [4] and [2], i.e. of greedily querying points which split up the version space and of exploiting regularity conditions in the distribution. That said, the properties of faithfulness and locality defined in section 5 are quite strong, and it would be interesting to see if weaker assumptions on χ can still ensure high splitting index; other properties of two-sidedness to use as references could be 1) *robustness*: for $X = \{0,1\}^n$, $\chi(h, x) = \chi(x')$ for

almost all vertices x' adjacent to x on the Boolean hypercube, 2) *monotonicity of examples*: if $\chi(h, x) = 0$, then $\chi(h, y) = 0$ for all $y \supseteq x$, 3) *closure* (strictly weaker than monotonicity): if $\chi(h, x) = 0$ and $\chi(h, x') = 0$, then $\chi(h, x \cup x') = 0$, 4) *monotonicity of hypotheses*: if $\chi(h, x) = 0$, then $\chi(h', x) = 0$ for all $h' \subseteq h$. To generalize 4) to other concept classes $\mathcal{C}$, we'd need some ordering of the hypotheses in $\mathcal{C}$.

It would also be helpful to study other compatibility notions in the case of learning disjunctions. One of several that [1] mentions is the co-training setting defined in Section 2. Some preliminary work suggests that a *time-efficient* active learning algorithm for this may not exist, as there is a natural connection between counting the number of monotone disjunctions consistent with a collection of labels and counting the number of vertex covers of a hypergraph, and even approximate counting the number of vertex covers up to any constant factor is known to be NP-hard. That said, one possible step might be to compute the splitting index for $\mathcal{C}_n$ under distributions for which the target concept is perfectly compatible under this χ.

References

[1] Blue, Avrim and Balcan, Maria-Florina. Open-problems in efficient semi-supervised PAC learning. In *Proceedings of the 20th Annual Conference on Computational Learning Theory (COLT)*, 2007.

[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] M.-F. Balcan, C. Berlind, S. Ehrlich, and Y. Liang. Efficient Semi-supervised and Active Learning of Disjunctions. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.

[4] S. Dasgupta, "Coarse sample complexity bounds for active learning," in NIPS, 2005.