Multi-Instance Learning with Any Hypothesis Class

Sivan Sabato SIVAN_SABATO@CS.HUJI.AC.IL
School of Computer Science & Engineering
The Hebrew University
Jerusalem 91904, Israel

Naftali Tishby TISHBY@CS.HUJI.AC.IL
School of Computer Science & Engineering
The Hebrew University
Jerusalem 91904, Israel

Editor: XXX

Abstract

In the supervised learning setting termed Multiple-Instance Learning (MIL), the examples are bags of instances, and the bag label is a function of the labels of its instances. Typically, this function is the Boolean OR. The learner observes a sample of bags and the bag labels, but not the instance labels that determine the bag labels. The learner is then required to emit a classification rule for bags based on the sample. MIL has numerous applications, and many heuristic algorithms have been used successfully on this problem, each adapted to specific settings or applications. In this work we provide a unified theoretical analysis for MIL, which holds for any underlying hypothesis class, regardless of a specific application or problem domain. We show that the sample complexity of MIL is only poly-logarithmically dependent on the size of the bag, for any underlying hypothesis class. In addition, we introduce a new PAC-learning algorithm for MIL, which employs a regular supervised learning algorithm as an oracle. We prove that efficient PAC-learning for MIL can be generated from any efficient non-MIL supervised learning algorithm that handles one-sided error. The computational complexity of the resulting algorithm is only polynomially dependent on the bag size.

1. Introduction

We consider the learning problem termed Multiple-Instance Learning (MIL), first introduced in Dietterich et al. (1997). MIL is a generalization of the classical supervised classification problem. As in classical supervised classification, in MIL the learner receives a sample of labeled examples drawn i.i.d from an arbitrary and unknown distribution, and its objective is to discover a classification rule with a small expected error over the same distribution. In MIL additional structure is assumed, whereby the examples are received as *bags* of *instances*, such that each bag is composed of several instances. It is assumed that each instance has a true label, however the learner only observes the labels of the bags. In classical MIL the label of a bag is the Boolean OR of the labels of the instances the bag contains. Various generalizations to MIL have been proposed (see e.g. Raedt, 1998; Weidmann et al., 2003). Here we consider both classical MIL and the more general problem where OR can be replaced with an arbitrary Boolean function, known to the learner in advance. We term the latter problem *generalized MIL*.

It is possible, in principle, to view MIL as a regular supervised classification task, where a bag is a single example, and the instances in a bag are merely part of its internal representation. Such treatment, however, would not take advantage of the special structure of a MIL problem and its possible connections to the related non-MIL classification problem. As we show in this work, these connections are strong and useful.

MIL has been used in numerous applications. In Dietterich et al. (1997) the drug design application motivates this setting. In this application, the goal is to predict which molecules would bind to a specific binding site. Each molecule has several possible conformations (shapes) it can take. If at least one of the conformations binds to the binding site, then the molecule is labeled positive. However, it is not possible to experimentally identify which conformation was the successful one. Thus, a molecule can be thought of as a bag of conformations, where each conformation is an instance in the bag representing the molecule. This application employs the hypothesis class of Axis Parallel Rectangles (APRs), and had made APRs the hypothesis class of choice in several theoretical works that we mention below. There are many other applications for MIL, including image classification (Maron and Ratan, 1998), web index page recommendation (Zhou et al., 2005) and text categorization (Andrews, 2007).

Previous theoretical analysis of the computational aspects of MIL has been done in two main settings. In some works (Auer et al., 1998; Blum and Kalai, 1998; Long and Tan, 1998), it is assumed that all the instances are drawn i.i.d from a single distribution over instances, so that the instances in each bag are statistically independent. Under this independence assumption, learning from an i.i.d. sample of bags is as easy as learning from an i.i.d. sample of instances with one-sided label noise. This is stated in the following theorem.

Theorem 1 (Blum and Kalai, 1998) *If a hypothesis class $\mathcal{H}$ is PAC-learnable in polynomial time from one-sided random classification noise, then the hypothesis class $\mathcal{H}$ is PAC-learnable in polynomial time in MIL under the independence assumption. The learning is polynomial in the bag size and in the sample size.*

The assumption of statistical independence of the instances in each bag is, however, very limiting, and it is irrelevant to many applications. More generally, one wishes to learn from an i.i.d. sample of bags drawn from an arbitrary distribution *over bags*, thus the instances within a bag may be statistically dependent. For the hypothesis class of APRs and an arbitrary distribution over bags, it is shown in Auer et al. (1998) that if there exists a PAC-learning algorithm for MIL with APRs, and this algorithm is polynomial in both the size of the bag and the dimension of the Euclidean space, then it is possible to polynomially PAC-learn DNF formulas, a problem which is solvable only if $\mathcal{RP} = \mathcal{NP}$ (Pitt and Valiant, 1986). In addition, if it is possible to improperly learn MIL with APRs (that is, to learn a classifier which is not itself an APR), then it is possible to improperly learn DNF formulas, a problem which has not been solved to this date for general distributions. This result implies that it is not possible to PAC-learn MIL on APRs using an algorithm which is efficient in both the bag size and the problem dimensionality. It does not, however, preclude the possibility of performing MIL efficiently under more restrictive assumptions.

In practice, numerous algorithms have been proposed for MIL, each focusing on a different specialization of this problem. Dietterich et al. (1997) propose several heuristic algorithms for finding an APR that predicts the label of an instance and of a bag. Diverse Density (Maron and Lozano-Pérez, 1998) and EM-DD (Zhang and Goldman, 2001) employ assumptions on the structure of the bags of instances. DP-Boost (Andrews and Hofmann, 2003), mi-SVM and MI-SVM (Andrews et al., 2002), and Multi-Instance Kernels (Gärtner et al., 2002) are approaches for learning MIL using margin-based objectives. Some of these methods work quite well in practice. However, no generalization guarantees have been provided for any of them.

In this work we analyze MIL and generalized MIL in a general framework, independent of a specific application, and provide results that hold for any underlying hypothesis class. We assume some fixed hypothesis class defined over instances. We investigate the relationship between learning with respect to this hypothesis class in the classical supervised learning setting with no bags, and learning with respect to the same hypothesis class in MIL. We address both sample complexity and computational feasibility.

Our sample complexity analysis shows that for binary hypothesis and thresholded real-valued hypotheses, the sample size required in generalized MIL grows only logarithmically with the bag size. We also provide poly-logarithmic sample complexity results for the case of margin learning. From this analysis it is possible to derive distribution-free generalization bounds for previously proposed algorithms for MIL.

Regarding the computational aspect, we provide a new learning algorithm with provable guarantees for classical MIL. Given a non-MIL learning algorithm for the hypothesis class, which can handle one-sided errors, we improperly learn MIL with the same hypothesis class. The construction is simple to implement, and provides a computationally efficient PAC-learning of MIL, with only a polynomial dependence of the run time on the bag size.

The structure of the paper is as follows. In Section 2 the problem is formally defined and notation is introduced. In Section 3 the sample complexity of generalized MIL for binary hypotheses is analyzed. Section 4 provides the learning algorithm for classical MIL. In Section 5 we analyze the sample complexity of generalized MIL with real-valued functions and for large-margin learning. We conclude in Section 6. Appendix A includes technical proofs that have been skipped in the text. A preliminary version of this work has been published as Sabato and Tishby (2009).

2. Problem Setting and Notation

Let $\mathcal{X}$ be the input space, also called the domain of instances. A bag is a set of instances from $\mathcal{X}$. The domain of labels is $\{-1, +1\}$. Throughout this work we assume for simplicity that all bags are of the same size r for some natural r, and that the instances in each bag are ordered. Thus the domain of bags is $\mathcal{X}^r$. Bags are denoted by $\bar{\mathbf{x}} = (x[1], \ldots, x[r]) \in \mathcal{X}^r$ where each $x[j]$ is an instance in the bag.

Denote the label of an instance $x \in \mathcal{X}$ by the probabilistic function $L(x)$. We assume a conditional probability distribution $D_{Y|X}$, such that $\forall x \in \mathcal{X}, y \in \{-1, +1\}$, $\mathbb{P}[L(x) = y] = D_{y|x}$. For any bag $\bar{\mathbf{x}} \in \mathcal{X}^r$, the label of the bag is determined from the labels of its instances using a fixed Boolean function $f : \{-1, +1\}^r \to \{-1, +1\}$. Thus $L(\bar{\mathbf{x}}) = f(L(x[1]), \ldots, L(x[r]))$. Importantly, the identity of f is known to the learner a-priori, thus each f defines a different generalized MIL problem (In classical MIL, f is the Boolean OR). We further assume a probability distribution $D_{\bar{\mathbf{X}}}$ over bags in $\mathcal{X}^r$. The learner receives as input a sample of labeled bags $\{(\bar{\mathbf{x}}_1, y_1), \ldots, (\bar{\mathbf{x}}_m, y_m)\}$ such that the bags $\bar{\mathbf{x}}_i$ are drawn independently from $D_{\bar{\mathbf{X}}}$, and each $y_i \in \{-1, +1\}$ is drawn according to the distribution of $L(\bar{\mathbf{x}}_i)$. We let D be the distribution over $\mathcal{X}^r \times \{-1, +1\}$ determined by $D_{\bar{\mathbf{X}}}$ and $D_{Y|X}$ as described herein.

The goal of the learner is to find a classification rule that would classify new bags drawn according to the same (unknown) joint distribution D with low error. We point out that it is not generally possible to find a low-error classification rule for instances based on a bag sample. As a simple counter-example, assume that f is the Boolean OR, and that every bag includes both a positive instance and a negative instance. In this case all bags are labeled as positive, and it is not possible to distinguish the two types of instances by observing only bag labels.

For a function $g : \mathcal{Z} \to \mathcal{T}$, we also use its vector extension $g : \mathcal{Z}^k \to \mathcal{T}^k$ defined as $g(\mathbf{a}) \triangleq (g(a[1]), \ldots, g(a[k]))$. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\langle \mathbf{x}, \mathbf{y} \rangle$ denotes the inner product of $\mathbf{x}$ and $\mathbf{y}$. For a natural number k, we denote by $[k]$ the set $\{1, \ldots, k\}$. $\log$ denotes a base 2 logarithm. For two sets A and B, B^A denotes the set of functions from A to B.

$\mathcal{H}$ denotes a hypothesis class that labels instances in $\mathcal{X}$. Hypotheses may be binary, so that $\mathcal{H} \subseteq \{-1, +1\}^{\mathcal{X}}$, or they may be real-valued, so that $\mathcal{H} \subseteq [-1, +1]^{\mathcal{X}}$. The relevant assumptions on $\mathcal{H}$ will be specified in context. We define the bag-labeling operator, which maps a hypothesis over instances into a hypothesis over bags, as follows:

Definition 2 *Let $f : Y^r \to Y$ for some set Y. The* bag-labeling operator, *denoted by $\phi_r^f : Y^{\mathcal{X}} \to Y^{\mathcal{X}^r}$, maps hypotheses over instances to hypotheses over bags as follows:*

$$\forall h \in Y^{\mathcal{X}}, \bar{\mathbf{x}} \in \mathcal{X}^r, \quad \phi_r^f(h)(\bar{\mathbf{x}}) \triangleq f(h(\bar{\mathbf{x}})) \equiv f(h(x[1]), \ldots, h(x[r])).$$

The set of hypotheses over bags generated from $\mathcal{H}$ by ϕ_r^f is denoted by $\phi_r^f(\mathcal{H}) \triangleq \{\phi_r^f(h) \mid h \in \mathcal{H}\}$.

3. Sample Complexity of Binary Hypotheses Classes

In this section we consider binary hypothesis classes $\mathcal{H} \subseteq \{-1,+1\}^{\mathcal{X}}$, and bound the VC-dimension of $\phi_r^f(\mathcal{H})$ as a function of the VC-dimension of $\mathcal{H}$ and of the bag size r. We show that the VC-dimension of $\phi_r^f(\mathcal{H})$ is at most logarithmic in r, and at most linear in the VC-dimension of $\mathcal{H}$, for any Boolean bag-labeling function f. Since the bounds on sample complexity are proportional to the VC-dimension of the problem (Vapnik and Chervonenkis, 1971), it follows that the sample complexity of MIL grows only logarithmically with the size of the bag. Thus MIL is feasible even for quite large bags, and can sometimes be used to accelerate even single-instance learning (Sabato et al., 2010). We further show lower bounds on the VC-dimension of MIL, indicating that the dependence of the upper bound on r and on the VC-dimension of $\mathcal{H}$ is imperative for a large class of Boolean bag-labeling functions. We also show a matching lower bound for VC-dimension of classical MIL with separating hyperplanes.

3.1 VC-Dimension Upper Bound

The following theorem establishes a VC-Dimension upper bound for generalized MIL.

Theorem 3 *Let $f : \{-1,+1\}^r \to \{-1,+1\}$ be an r-ary Boolean function. Let ϕ_r^f be the bag-labeling operator defined in Def. 2. Let $\mathcal{H} \subseteq \{-1,+1\}^X$ be a hypothesis class with a finite VC-dimension d, and denote the VC-dimension of $\phi_r^f(\mathcal{H})$ by d_r. Then*

$$d_r \leq \max\{16, 2d\log(2er)\}.$$

Proof For a hypothesis h, denote by $h_{|A}$ its restriction to a set A, and for a set of hypotheses $\mathcal{J}$, denote by $\mathcal{J}_{|A}$ the restriction of each of its members to A, so that $\mathcal{J}_A \triangleq \{h_{|A} \mid h \in \mathcal{J}\}$.

Since d_r is the VC-dimension of $\phi_r^f(\mathcal{H})$, there exists a set of bags $S = \{\bar{\mathbf{x}}_i\}_{i\in[d_r]} \subseteq \mathcal{X}^r$ that is shattered by $\phi_r^f(\mathcal{H})$, so that $|\phi_r^f(\mathcal{H})_{|S}| = 2^{d_r}$. Let $S^{\cup} = \{x_i[j]\}_{i\in[m],j\in[r]}$ be the set of instances of bags in S. Clearly $|\phi_r^f(\mathcal{H})_{|S}| \leq |\mathcal{H}_{|S^{\cup}}|$, therefore $2^{d_r} \leq |\mathcal{H}_{|S^{\cup}}|$. Applying Sauer's lemma (Sauer, 1972; Vapnik and Chervonenkis, 1971) to $\mathcal{H}$ we get

$$2^{d_r} \leq |\mathcal{H}_{|S^{\cup}}| \leq \left(\frac{e|S^{\cup}|}{d}\right)^d \leq \left(\frac{erd_r}{d}\right)^d,$$

Where e is the base of the natural logarithm. It follows that $d_r \leq d(\log(er) - \log d) + d\log d_r$. To provide an explicit bound for d_r, we bound $d\log d_r$ by dividing to cases:

1. Either $d\log d_r \leq \frac{1}{2}d_r$, thus $d_r \leq 2d(\log(er) - \log d) \leq 2d\log(er)$,
2. or $\frac{1}{2}d_r < d\log d_r$. In this case,
   (a) either $d_r \leq 16$,
   (b) or $d_r > 16$. In this case $\sqrt{d_r} < d_r/\log d_r < 2d$, thus $d\log d_r = 2d\log\sqrt{d_r} \leq 2d\log 2d$. Substituting in the implicit bound we get $d_r \leq d(\log(er) - \log d) + 2d\log 2d \leq 2d\log(2er)$.

Combining the cases we have $d_r \leq \max\{16, 2d\log(2er)\}$. ■

3.2 VC-Dimension Lower Bounds

In this section we show lower bounds on the VC-dimension of MIL, indicating that the dependence on d and r in Theorem 3 is tight in two important settings. We start with a lower bound with respect to a worst-case hypothesis class, for any bag-labeling function which is sensitive to its inputs in a specific sense defined in the following theorem. Functions that satisfy this requirement include the Boolean OR, AND, and Parity, and all their variants that stem from negating some of the inputs.

Theorem 4 *Let $f : \{-1,+1\}^r \to \{-1,+1\}$ be an r-ary Boolean function. Assume that there exist two Boolean vectors $\mathbf{c}, \mathbf{a} \in \{-1,+1\}^r$ such that*

$$\forall j \in [r], y \in \{-1,+1\}, \qquad f(c[1],\ldots,c[j]\cdot y,\ldots,c[r]) = a[j]\cdot y.$$

For any natural d and any instance domain $\mathcal{X}$ with $|\mathcal{X}| \geq rd\lfloor\log(r)\rfloor$, there exists a hypothesis class $\mathcal{H}$ with a VC dimension at most d, such that the VC dimension of $\phi^f_r(\mathcal{H})$ is at least $d\lfloor\log(r)\rfloor$.

Proof Let $S \subseteq \mathcal{X}^r$ be a set of $d\lfloor\log(r)\rfloor$ bags, such that all the instances in all the bags are distinct elements of $\mathcal{X}$. Divide S into d mutually exclusive subsets, each with $\lfloor\log(r)\rfloor$ bags. Denote bag p in subset t by $\bar{\mathbf{x}}_{(p,t)}$. We define the hypothesis class

$$\mathcal{H} \triangleq \{h[k_1,\ldots,k_d] \mid \forall i \in [d], k_i \in [2^{\lfloor\log(r)\rfloor}]\},$$

where $h[k_1,\ldots,k_d]$ is defined as follows (see illustration in Table 1): For $x \in \mathcal{X}$ which is not an instance of any bag in S, $h[k_1,\ldots,k_d] = -1$. For $x = x_{(p,t)}[j]$, let $b_{(p,n)}$ be bit p in the binary representation of the number n. We define

$$h[k_1,\ldots,k_d](x_{(p,t)}[j]) = \begin{cases} c[j]\cdot a[j]\cdot(2b_{(p,j-1)}-1) & j = k_t, \\ c[j] & j \neq k_t, \end{cases}$$

t	p	Instance label $h(x_{(p,t)}[r])$								Bag label $\phi^{\mathrm{OR}}_r(h)(\bar{\mathbf{x}}_i)$
1	1	$-$	$-$	$-$	$+$	$-$	$-$	$-$	$-$	$+$
	2	$-$	$-$	$-$	$+$	$-$	$-$	$-$	$-$	$+$
	3	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$
2	1	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$+$	$+$
	2	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$+$	$+$
	3	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$+$	$+$
3	1	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$
	2	$-$	$+$	$-$	$-$	$-$	$-$	$-$	$-$	$+$
	3	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$

Table 1: An example of the hypotheses $h = h[4,8,3]$, with $f = \mathrm{OR}$ (so that $\mathbf{c}$ and $\mathbf{a}$ are all -1 vectors), $r = 8$, and $d = 3$. Each line represents a bag in S, each column represents an instance in the bag.

We now show that S is shattered by $\phi^f_r(\mathcal{H})$, indicating that the VC-dimension of $\phi^f_r(h)$ over $\mathcal{X}$ is at least $|S| = d\lfloor\log(r)\rfloor$. To complete the proof, we further show that the VC dimension of $\mathcal{H}$ is no more than d.

- S is shattered by $\phi^f_r(\mathcal{H})$: Let $\{y_{(p,t)}\}_{p\in\lfloor\log(r)\rfloor, t\in[d]}$ be some labeling over $\{-1,+1\}$ for the bags in S. For each $t \in [d]$ let

$$k_t \triangleq 1 + \sum_{p=1}^{\lfloor\log(r)\rfloor} \frac{y_{(p,t)}+1}{2}\cdot 2^{p-1}.$$

Then for all $p \in [\lfloor \log(r) \rfloor]$ and $t \in [d]$,

$$\phi_r^f(h[k_1,\ldots,k_d])(\bar{\mathbf{x}}_{(p,t)}) = f(c[1],\ldots,c[k_t]\cdot a[k_t]\cdot(2b_{(p,k_t-1)}-1),\ldots,c[r]) = 2b_{(p,k_t-1)}-1 = y_{(p,t)}.$$

Thus $h[k_1,\ldots,k_d]$ labels S according to $\{y_{(p,t)}\}$.

- The VC-dimension of $\mathcal{H}$ is no more than d: Let $A \subseteq \mathcal{X}$ of size $d+1$. If there is an element in A which is not an instance in S then this element is labeled -1 by all $h \in \mathcal{H}$, therefore A is not shattered. Otherwise, all elements in A are instances in bags in S. Since there are d subsets of S, there exist two elements in A which are instances of bags in the same subset t. Denote these instances by $x(p_1,t)[j_1]$ and $x(p_2,t)[j_2]$. Consider all the possible labelings of the two elements by hypotheses in $\mathcal{H}$. If A is shattered, there must be four possible labelings for these elements. However, by the definition of $h[k_1,\ldots,k_d]$ it is easy to see that if $j_1 = j_2 = j$ then there are at most two possible labelings by hypotheses in $\mathcal{H}$, and if $j_1 \neq j_2$ then there are at most three possible labelings. Thus A is not shattered by $\mathcal{H}$, hence the VC-dimension of $\mathcal{H}$ is no more than d.

■

Theorem 7 below provides a lower bound for the VC-dimension of MIL for the common case where f is the Boolean OR and the hypothesis class is the class of separating hyperplanes in $\mathbb{R}^n$, denoted by $\mathcal{W}_n \triangleq \{\mathbf{x} \mapsto \mathrm{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) \mid \mathbf{w} \in \mathbb{R}^n\}$. We denote the VC-dimension of $\phi_r^{\mathrm{OR}}(\mathcal{W}_n)$ by $d_{r,n}$. The lower bound is proved using two lemmas: Lemma 5 provides a lower bound for $d_{r,3}$, and Lemma 6 links $d_{r,n}$ for small n with $d_{r,n}$ for large n. The resulting general lower bound is then stated in Theorem 7.

Lemma 5 *Let $d_{r,n}$ the VC-dimension of $\phi_r^{\mathrm{OR}}(\mathcal{W}_n)$ as defined above. Then $d_{r,3} \geq \lfloor \log(2r) \rfloor$.*

Proof Denote $L \triangleq \lfloor \log(2r) \rfloor$. We will construct a set S of L bags of size r that is shattered by $\mathcal{W}_3$. The construction is illustrated in Figure 1.

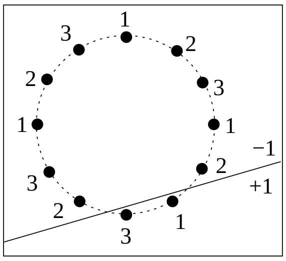

Figure 1: An illustration of the constructed shattered set, with $r = 4$ and $L = \log 4 + 1 = 3$. Each dot corresponds to an instance. The numbers next to the instances denote the bag to which an instance belongs, and match the sequence N defined in the proof. In this illustration bags 1 and 3 are labeled as positive by the bag-hypothesis represented by the solid line.

Let $\mathbf{n} = (n_1,\ldots,n_K)$ be a sequence of indices from $[L]$, created by concatenating all the subsets of $[L]$ in some arbitrary order, so that $K = L2^{L-1}$, and every index appears $2^{L-1} \leq r$ times in $\mathbf{n}$. Define a set $A = \{\mathbf{a}_k \mid k \in [K]\} \subseteq \mathbb{R}^3$ where $\mathbf{a}_k \triangleq (\cos(2\pi k/K), \sin(2\pi k/K), 1) \in \mathbb{R}^3$, so that $\mathbf{a}_1,\ldots,\mathbf{a}_K$ are equidistant on a unit circle on a plane embedded in $\mathbb{R}^3$. Define the set of bags $S = \{\bar{\mathbf{x}}_1,\ldots,\bar{\mathbf{x}}_L\}$ such that $\bar{\mathbf{x}}_i = (x_i[1],\ldots,x_i[r])$ where $\{x_i[j] \mid j \in [r]\} = \{a_k \mid n_k = i\}$.

We now show that S is shattered by $\mathcal{W}_3$: Let $(y_1,\ldots,y_L)$ be some binary labeling of L bags, and let $Y = \{i \mid y_i = +1\}$. By the definition of $\mathbf{n}$, there exist j_1, j_2 such that $Y = \{n_k \mid j_1 \leq k \leq j_2\}$. Clearly, there exists a hyperplane $\mathbf{w} \in \mathbb{R}^3$ that separates the vectors $\{\mathbf{a}_k \mid j_1 \leq k \leq j_2\}$ from the rest of the vectors in A. Thus $\mathrm{sign}(\langle \mathbf{w}, \mathbf{a}_k \rangle) = +1$ if and only if $j_1 \leq k \leq j_2$. It follows that $\phi_r^{\mathrm{OR}}(\mathbf{w})(\bar{\mathbf{x}}_i) = +1$ if and only if there is a $k \in \{j_1, \ldots, j_2\}$ such that $\mathbf{a}_k$ is an instance in $\bar{\mathbf{x}}_i$, that is such that $n_k = i$. This condition holds if and only if $i \in Y$, hence $\mathbf{w}$ classifies S according to the given labeling. It follows that S is shattered by $\mathcal{W}_3$, therefore $d_r^3 \geq |S| = \lfloor \log(2r) \rfloor$. ■

Lemma 6 *Let k, n, r be natural number such that $k \leq n$. Then $d_{r,n} \geq \lfloor n/k \rfloor d_{r,k}$.*

Proof For a vector $\mathbf{x} \in \mathbb{R}^k$ and a number $t \in \{0, \ldots, \lfloor n/k \rfloor\}$ define the vector $s(\mathbf{x}, t) \triangleq (0, \ldots, 0, x[1], \ldots, x[k], 0, \ldots, 0) \in \mathbb{R}^n$, where $x[1]$ is at coordinate $kt + 1$. Similarly, for a bag $\bar{\mathbf{x}}_i = (\mathbf{x}_i[1], \ldots, \mathbf{x}_i[r]) \in (\mathbb{R}^k)^r$, define the bag $s(\bar{\mathbf{x}}_i, t) \triangleq (s(\mathbf{x}_i[1], t), \ldots, s(\mathbf{x}_i[r], t)) \in (\mathbb{R}^n)^r$.

Let $S_k = \{\bar{\mathbf{x}}_i\}_{i \in [d_{r,k}]} \subseteq (\mathbb{R}^k)^r$ be a set of bags with instances in $\mathbb{R}^k$ that is shattered by $\phi_r^{\mathrm{OR}}(\mathcal{W}_k)$. Define S_n, a set of bags with instances in $\mathbb{R}^n$: $S_n \triangleq \{s(\bar{\mathbf{x}}_i, t)]\}_{i \in [d_{r,k}], t \in [\lfloor n/k \rfloor]} \subseteq (\mathbb{R}^n)^r$. Then S_n is shattered by $\mathcal{W}_n$: Let $\{y_{(i,t)}\}_{i \in [d_{r,k}], t \in [\lfloor n/k \rfloor]}$ be some labeling for S_n. S_k is shattered by $\mathcal{W}_k$, hence there are separators $\mathbf{w}_1, \ldots, \mathbf{w}_{\lfloor n/k \rfloor} \in \mathbb{R}^k$ such that $\forall i \in [d_{r,k}], t \in \lfloor n/k \rfloor, \quad \phi_r^{\mathrm{OR}}(\mathbf{w}_t)(\bar{\mathbf{x}}_i) = y_{(i,t)}$.

Set $\mathbf{w} \triangleq \sum_{t=0}^{\lfloor n/k \rfloor} s(\mathbf{w}_t, t)$. Then $\langle \mathbf{w}, s(\mathbf{x}, t) \rangle = \langle \mathbf{w}_t, \mathbf{x} \rangle$. Therefore

$$\begin{aligned}\phi_r^{\mathrm{OR}}(\mathbf{w})(s(\bar{\mathbf{x}}_i, t)) &= \mathrm{OR}(\mathrm{sign}(\langle \mathbf{w}, s(\mathbf{x}_i[1], t) \rangle), \ldots, \mathrm{sign}(\langle \mathbf{w}, s(\mathbf{x}_i[r], t) \rangle)) \\ &= \mathrm{OR}(\mathrm{sign}(\langle \mathbf{w}_t, \mathbf{x}_i[1] \rangle), \ldots, \mathrm{sign}(\langle \mathbf{w}_t, \mathbf{x}_i[r] \rangle)) = \phi_r^{\mathrm{OR}}(\mathbf{w}_t)(\bar{\mathbf{x}}_i) = y_{(i,t)}.\end{aligned}$$

S_n is thus shattered, hence $d_{r,n} \geq |S_n| = \lfloor n/k \rfloor d_{r,k}$. ■

The desired theorem is an immediate consequence of the two lemmas above:

Theorem 7 *Let $\mathcal{W}_n$ be the class of separating hyperplanes in $\mathbb{R}^n$ as defined above. The VC-dimension of $\phi_r^{\mathrm{OR}}(\mathcal{W}_n)$ is at least $\lfloor n/3 \rfloor \lfloor \log 2r \rfloor$.*

4. PAC-Learning for MIL

In the previous section we addressed the sample complexity of generalized MIL, showing that it grows only logarithmically with the bag size. We now turn to consider the computational aspect of MIL, focusing on classical MIL, in which the bag-labeling function is the Boolean OR. We provide a simple algorithm for MIL which uses as an oracle a learning algorithm which operates on single instances. In this section we assume real-valued hypotheses, that is $\mathcal{H} \subseteq [-1, +1]^{\mathcal{X}}$. The bag-labeling function is accordingly generalized to a $\max$ instead of OR. The related sample complexity analysis of MIL for real-valued hypotheses is deferred to Section 5.

The proposed algorithm, named `MILearn`, uses an algorithm $\mathcal{A}$ as a black-box. $\mathcal{A}$ operates on single instances and returns a hypothesis in $\mathcal{H}$, while `MILearn` operates on a sample of bags and selects a bag-hypothesis from $\phi_r^{\max}(\mathcal{H})$. We show that if $\mathcal{A}$ perfoms one-sided or agnostic learning of single instances with respect to $\mathcal{H}$, then `MILearn` is a *weak learner* for MIL with respect to $\phi_r^{\max}(\mathcal{H})$. `MILearn` can thus be used as the building block in a boosting algorithm (Freund and Schapire, 1997). The boosting algorithm returns a linear combination of bag-hypotheses that classifies unseen bags with high accuracy. Furthermore, if $\mathcal{A}$ is efficient then our algorithm is also efficient, with a polynomial dependence on the bag size.

We describe the weak learner in Section 4.1. We then proceed to explain the boosting construction in Section 4.2, and conclude the section with a short discussion of implications in Section 4.3.

4.1 The Weak Learner

We start with some notation. A labeled and weighted sample of instances is a set of triplets $(w, x, y) \in \mathbb{R}^+ \times \mathcal{X} \times \{-1, +1\}$, where w is the weight of the instance, x is the instance, and y is the instance label. A labeled and weighted sample of bags is a set of triplets $(w, \bar{\mathbf{x}}, y) \in \mathbb{R}^+ \times \mathcal{X}^r \times \{-1, +1\}$, where w is the weight of the bag, $\bar{\mathbf{x}}$ is the bag, and y is the bag label. The *edge* of a hypothesis is a measure of how successful the hypothesis is in classifying with respect to a distribution. For an instance hypothesis $h : \mathcal{X} \to [-1, +1]$ and a distribution D over $\mathcal{X} \times \{-1, +1\}$, the edge of h with respect to D is

$$\Gamma(h, D) \triangleq \mathbb{E}_{(X,Y)\sim D}[Y \cdot h(X)].$$

Note that if $h(x)$ is interpreted as the probability of h to emit 1 for input x, then $\frac{1-\Gamma(h,D)}{2}$ is the expected error of h on D. For a weighted and labeled instance sample $S = \{(w_i, x_i, y_i)\}_{i\in[m]}$, D_S is the probability distribution over $\mathcal{X} \to [-1, +1]$ defined by $\mathbb{P}_{D_S}[(X, Y) = (x_i, y_i)] = w_i / \sum_{j=1}^{m} w_j$. Where it is clear from context, we use S interchangeably with D_S. Thus $\Gamma(h, S) \equiv \Gamma(h, D_S)$. $\Gamma(h, D)$ and $\Gamma(h, S)$ are defined similarly for a bag hypothesis $h \in \phi_r^{\max}(\mathcal{H})$, a distribution D over $\mathcal{X}^r \times [-1, +1]$, and a labeled and weighted sample of bags S.

The proposed algorithm `MILearn`, listed as Algorithm 1 below, accepts as input a bag sample denoted S_B, and assumes access to an algorithm $\mathcal{A}$. $\mathcal{A}$ receives a labeled and weighted instance sample and returns an instance hypothesis $h \in \mathcal{H}$. We denote by $\mathcal{A}(S) \in \mathcal{H}$ the result of running $\mathcal{A}$ with input S. h_{pos} denotes the constant positive hypothesis: $\forall x \in \mathcal{X}, \quad h_{\text{pos}}(x) = +1$. For simplicity we assume $h_{\text{pos}} \in \mathcal{H}$. The output of `MILearn` is a bag-hypothesis in $\phi_r^{\max}(\mathcal{H})$ that classifies S_B with an edge that depends on the best achievable edge for S_B, as we presently show.

`MILearn` is a simple algorithm: It constructs a sample of instances S_I from the instances that make up bags in S_B, labeling each instance in S_I with the label of the bag it came from. The weight of an instance with a positive label is set to be the weight of the bag it came from, divided by r, and the weight of an instance with a negative label is set to be the same as the weight of the bag it came from. Having constructed S_I, `MILearn` runs $\mathcal{A}$ on S_I. It then selects whether to return $\phi_r^{\max}(\mathcal{A}(S_I))$ or $\phi_r^{\max}(h_{\text{pos}})$, whichever provides the better edge on S_B.

Algorithm 1: `MILearn`

Assumptions:

- Access to an algorithm $\mathcal{A}$, that receives a weighted instance sample and returns a hypothesis in $\mathcal{H}$.
- $h_{\text{pos}} \in \mathcal{H}$.

Input: $S_B \triangleq \{(w_i, \bar{\mathbf{x}}_i, y_i)\}_{i\in[m]}$ – a labeled and weighted sample of bags;

Output: $h_M \in \phi_r^{\max}(\mathcal{H})$.

```
$\alpha_{(+1)} \leftarrow \frac{1}{r}, \alpha_{(-1)} \leftarrow 1.$
$S_I \leftarrow \{(\alpha_{y_i} \cdot w_i, x_i[j], y_i)\}_{i\in[m],j\in[r]}.$
$h_I \leftarrow \mathcal{A}(S_I).$
if $\Gamma(\phi_r^{\max}(h_I), S_B) \geq \Gamma(\phi_r^{\max}(h_{\text{pos}}), S_B)$ then
    $h_M \leftarrow \phi_r^{\max}(h_I),$
else
    $h_M \leftarrow \phi_r^{\max}(h_{\text{pos}}).$
```

We now prove that MILearn provides guarantees for the edge of the resulting hypothesis, depending on the properties on $\mathcal{A}$. Before stating the result, we define some auxiliary notation. For a distribution D over $\mathcal{X} \times \{-1,+1\}$, we denote by $\Omega(D)$ the set of hypotheses which have only one-sided error on D. We specifically require that such hypotheses err only on *positive* examples in S. Formally,

$$\Omega(D) \triangleq \{h \in [-1,+1]^{\mathcal{X}} \mid \mathbb{P}_D[h(X) = -1 \mid Y = -1] = 1\}.$$

If D is a distribution over $\mathcal{X}^r \times \{-1,+1\}$, then

$$\Omega(D) \triangleq \{h \in [-1,+1]^{\mathcal{X}} \mid \mathbb{P}_D[\phi_r^{\max}(h)(X) = -1 \mid Y = -1] = 1\}.$$

The definition for D and h defined on bags is similar.

In the following theorem we compare the edge achieved using MILearn to the best possible edge for the sample S_B. The best edge for S_B achievable by a hypothesis in $\mathcal{H}$ is denoted γ^*, and the best edge achievable by a hypothesis in $\mathcal{H}$ with one-sided error is denoted γ_+^*. Formally:

$$\gamma^* \triangleq \max_{h \in \mathcal{H}} \Gamma(\phi_r^{\max}(h), S_B), \tag{1}$$

$$\gamma_+^* \triangleq \max_{h \in \mathcal{H} \cap \Omega(S_B)} \Gamma(\phi_r^{\max}(h), S_B). \tag{2}$$

Theorem 8 *Let $\mathcal{H} \subseteq [-1,+1]^{\mathcal{X}}$ be a set of instance hypotheses. Let h_M be the hypothesis returned by* MILearn *when receiving S_B as input, and let $\gamma \triangleq \Gamma(h_M, S_B)$. Then*

(a) If for any instance sample S, $\Gamma(\mathcal{A}(S), S) \geq \max_{h \in \mathcal{H} \cap \Omega(S)} \Gamma(h, S) - \eta$ for some $\eta > 0$, then

$$\gamma \geq \frac{\gamma_+^* - r^2\eta}{2r-1}. \tag{3}$$

(b) If for any instance sample S, $\Gamma(\mathcal{A}(S), S) \geq \max_{h \in \mathcal{H}} \Gamma(h, S) - \eta$ for some $\eta > 0$, and $\gamma^* - \eta \geq 1 - \frac{1}{r^2}$, then

$$\gamma \geq \frac{1 - r^2(1 - \gamma^* + \eta)}{2r-1} \geq 0. \tag{4}$$

Proof We prove part (a) of the theorem, and defer the similar proof of part (b) to Appendix A. Denote the total weight of examples in a sample S by $W(S)$. In addition, denote $W_+ \triangleq \sum_{i:y_i=+1} w_i$ and $W_- \triangleq \sum_{i:y_i=-1} w_i$, where $\{w_i\}$ are the weights of the bags in S_B. We assume w.l.o.g. that $W(S_B) \equiv W_+ + W_- = 1$. Let $h_+^* \triangleq \operatorname{argmax}_{h \in \mathcal{H} \cap \Omega(S_B)} \Gamma(\phi_r^{\max}(h), S_B)$. S_I and h_I are as defined in steps 2 and 3 of MILearn. The proof of Theorem 8(a) employs the following technical lemmas. Their proofs are provided in Appendix A.

Lemma 9 *For any instance hypothesis h, $\Gamma(\phi_r^{\max}(h), S_B) \geq W(S_I)\Gamma(h, S_I) + (1-r)W_-$.*

Lemma 10 *If the condition on $\mathcal{A}$ in (a) holds, then $\Gamma(h_I, S_I) \geq \Gamma(h_+^*, S_I) - \eta$.*

Lemma 11 *For any instance instance hypothesis h,*

$$W(S_I)\Gamma(h, S_I) \geq \tfrac{1}{r}\Gamma(\phi_r^{\max}(h), S_B) + (\tfrac{1}{r} - 1)W_+ - (r - \tfrac{1}{r}) \sum_{y_i=-1} w_i \phi_r^{\max}(h)(\bar{\mathbf{x}}_i).$$

For $h \in \Omega(S_B)$, $\sum_{y_i=-1} w_i \phi_r^{\max}(h)(\bar{\mathbf{x}}_i) = -W_-$. Therefore, from Lemma 11,

$$W(S_I)\Gamma(h, S_I) \geq \tfrac{1}{r}\Gamma(\phi_r^{\max}(h), S_B) + (\tfrac{1}{r} - 1)W_+ + (r - \tfrac{1}{r})W_-. \tag{5}$$

Applying Lemma 9, Lemma 10 and Eq. (5) sequentially, we have

$$\begin{aligned}\Gamma(\phi_r^{\max}(h_I), S_B) &\geq W(S_I)\Gamma(h_I, S_I) + (1-r)W_- \geq W(S_I)(\Gamma(h_+^*, S_I) - \eta) + (1-r)W_- \\ &\geq \tfrac{1}{r}\Gamma(\phi_r^{\max}(h_+^*), S_B) + (\tfrac{1}{r} - 1)W_+ + (1 - \tfrac{1}{r})W_- - r\eta = \tfrac{1}{r}\gamma_+^* + (1 - \tfrac{1}{r})(1 - 2W_+) - r\eta,\end{aligned} \tag{6}$$

where the last equality follows from the assumption that $W_+ + W_- = 1$. By step 4 of `MILearn`,

$$\gamma = \max\{\Gamma(\phi_r^{\max}(h_I), S_B), \Gamma(\phi_r^{\max}(h_{\text{pos}}), S_B)\} \geq \max\left\{\tfrac{1}{r}\gamma_+^* + (1 - \tfrac{1}{r})(1 - 2W_+) - r\eta,\ 2W_+ - 1\right\}.$$

It is easy to verify that for any $W_+ \in [0,1]$, $\gamma \geq \frac{\gamma_+^* - r^2\eta}{2r-1}$. ■

Theorem 8 guarantees that if $\mathcal{A}$ performs approximate ERM with respect to $\mathcal{H}$ on its non-bag input sample, then `MILearn` achieves an approximation to the optimal edge of a hypothesis in $\phi_r^{\max}(\mathcal{H})$ on its bag input sample. It is also easy to see that the time complexity of `MILearn` is bounded by $O(c(\mathcal{A}) + rm)$, where $c(\mathcal{A})$ is an upper bound on the time complexity of $\mathcal{A}$. In addition, the results of Theorem 8 can easily be extended to the case where instead of access to an approximate ERM algorithm, we have access to a PAC-learning algorithm, with no assumption on its internal mechanism.

Definition 12 (One-sided and agnostic PAC-learning algorithms) *Let $\mathcal{B}(\epsilon, \delta, S)$ be an algorithm that accepts as input $\delta, \epsilon \in (0,1)$, and a labeled sample $S \in (\mathcal{X} \times \{-1,+1\})^m$, and emits as output a hypothesis $h \in \mathcal{H}$.*

$\mathcal{B}$ is a one-sided PAC-learning algorithm *for $\mathcal{H}$ with complexity $c(\epsilon, \delta)$ if $\mathcal{B}$ runs for no more than $c(\epsilon, \delta)$ steps, and for any probability distribution D over $\mathcal{X} \times \{-1,+1\}$, if S is an i.i.d. sample from D of size $c(\epsilon, \delta)$ then with probability at least $1 - \delta$ over S and the randomization of $\mathcal{B}$,*

$$\Gamma(\mathcal{B}(S), D) \geq \sup_{h \in \mathcal{H} \cap \Omega(D)} \Gamma(h, D) - 2\epsilon.$$

Similarly, $\mathcal{B}$ is an agnostic PAC-learning algorithm *for $\mathcal{H}$ if with probability at least $1 - \delta$ over S and the randomization of $\mathcal{B}$,*

$$\Gamma(\mathcal{B}(S), D) \geq \sup_{h \in \mathcal{H}} \Gamma(h, D) - 2\epsilon.$$

With access to a one-sided or agnostic PAC-learning algorithm, we can construct an algorithm $\mathcal{A}$ such that the required guarantees for Theorem 8 would hold with high probability over the randomization of $\mathcal{A}$. Let S be the input to $\mathcal{A}$. $\mathcal{A}$ creates an unweighted sample $\tilde{S}$ from S by drawing $c(\eta/2, \delta)$ labeled instances independently according to D_S, and returns $\mathcal{B}(\tilde{S})$. If $\mathcal{B}$ is a one-sided PAC-learning algorithm, then with probability at least $1 - \delta$, $\Gamma(\mathcal{A}(S), S) \geq \max_{h \in \mathcal{H}} \Gamma(h, S) - \eta$. Similarly, if $\mathcal{B}$ is an agnostic PAC-learning algorithm, then with probability at least $1 - \delta$, $\Gamma(\mathcal{A}(S), S) \geq \max_{h \in \mathcal{H} \cap \Omega(S)} \Gamma(h, S) - \eta$. Thus, by a slight modification of the proof of Theorem 8, we get the following theorem.

Theorem 13 *Let $\delta, \eta \in (0,1)$. Under the same definitions as in Theorem 8, If $\mathcal{A}$ is created from a one-sided PAC-learning algorithm $\mathcal{B}$ with complexity $c(\epsilon, \delta)$ as described above, then* `MILearn` *uses $O(\max(c(\eta/2, \delta), mr))$ steps, and with probability at least $1 - \delta$ emits a hypothesis with edge γ that satisfies Eq. (3). If $\mathcal{B}$ is an agnostic PAC-learning algorithm, then Eq. (4) holds instead.*

Specifically, if $\mathcal{B}$ is a one-sided PAC-learning algorithm, we can set $\eta \leq \frac{\gamma_+^*}{2r^2}$ to get a guaranteed approximation for γ_+^*. If $\mathcal{B}$ is an agnostic PAC-learning algorithm and $\gamma \geq 1 - \frac{1}{2r^2}$, we can set $\eta = \frac{1}{2r^2}$ to get a positive γ with high probability.

4.2 Boosting with the Weak Learner

Theorem 8 and Theorem 13 show that under suitable conditions, `MILearn` produces a hypothesis whose edge approximates the edge of the best hypothesis in $\phi_r^{\max}(\mathcal{H})$. In this section we conclude that `MILearn` can be used as the weak learner in a boosting algorithm for MIL. The result is a learning algorithm for MIL with guaranteed generalization.

There are plenty of possible boosting algorithms. For concreteness, we base the following discussion on `AdaBoost`* (Rätsch and Warmuth, 2005), since it provides suitable guarantees on the *margin* of its output hypothesis. The margin of a linear combination of hypotheses on a sample $S = \{(x_i, y_i)\}_{i=1}^m$ is defined as follows:

$$M(\alpha, S) = \min_{i\in[m]} y_i \sum_h \mathbb{P}_\alpha[h]h(x_i),$$

where α is some distribution with finite support over hypotheses. The input to `AdaBoost`* is an i.i.d. labeled sample. Like all boosting algorithms, `AdaBoost`* assumes access to a *weak learner*, which is an algorithm that accepts a weighted sample and returns a hypothesis from some fixed hypothesis class. `AdaBoost`* activates its weak learner several times on different weighted samples, and returns as output a linear combination of the returned hypotheses. If the hypothesis returned by the weak learner in each round has edge at least ρ, then after $\frac{2\ln m}{\nu^2}$ iterations, `AdaBoost`* finds a linear combination of hypotheses α, with $M(\alpha, S) \geq \rho - \nu$. The generalization error of α can be bounded using its margin on S, and using d, the complexity of the underlying hypothesis class. The following bound (Schapire et al., 1998; Schapire and Singer, 1999) holds with probability $1 - \delta$ over the training samples:

$$\mathbb{P}[Y \sum_h \mathbb{P}_\alpha[h]h(X) \leq 0] \leq O\left(\left(\frac{d\log^2(m/d)}{mM^2(\alpha, S)} + \log(1/\delta)\right)^{\frac{1}{2}}\right). \tag{7}$$

In our case, the input sample is a labeled sample of bags, the fixed hypothesis class is $\phi_r^{\max}(\mathcal{H})$, and the output of `AdaBoost`* is a linear combination of hypotheses in this class. We will show that if `MILearn` is used as the weak learner, then under suitable assumptions the margin guarantees indeed hold, and a resulting generalization bound with a polynomial dependence on r follows.

For a sample of bags S, let ρ^* be the largest margin that can be achieved for this sample by a linear combination of hypotheses from $\phi_r^{\max}(\mathcal{H})$. Formally, let A be the set of distributions over $\phi_r^{\max}(\mathcal{H})$ with finite support, and define

$$\rho^* \triangleq \max_{\alpha\in A} M(\alpha, S).$$

Let $S_\mathbf{w}$ be the sample S with its bags weighted according to $\mathbf{w} \in (\mathbb{R}^+)^m$. From the Min-Max theorem (von Neumann 1928, and further developed in Rätsch and Warmuth 2005),

$$\rho^* = \min_{\mathbf{w}:\sum w_i=1} \max_h \Gamma(h, S_\mathbf{w}),$$

whenever the maximum on $\phi_r^{\max}(\mathcal{H})$ is defined. Thus, for any weighting of the sample S, there exists a single hypothesis $h \in \phi_r^{\max}(\mathcal{H})$ with edge at least ρ^*. The input to `MILearn` in every iteration of `AdaBoost`* is the weighted sample $S_B \equiv S_\mathbf{w}$. Thus in each round, $\gamma^* \geq \rho^*$, where γ^* is the best achievable edge, defined in Eq. (1).

For instance, assume that the conditions on $\mathcal{A}$ in Theorem 8(b) hold with $\eta = 0$, and suppose $\rho^* \geq 1 - \frac{1}{r^2}$. Then for every input sample $S_\mathbf{w}$, $\gamma^* \geq \rho^*$. Thus by Theorem 8 `MILearn` returns a hypothesis with edge at least $\rho = \frac{1-r^2(1-\rho^*)}{2r-1} \geq 0$. Setting $\nu = \rho/2$, we get that `AdaBoost`* achieves a margin of $\frac{1-r^2(1-\rho^*)}{4r-2}$ after a number of iterations which is polynomial in the bag size r.[1] It is easy to extend

[1] One can also replace the hard margin requirement with a soft margin formulation, following Warmuth et al. (2007) and Shalev-Shwartz and Singer (2008).

this argument also for the case where MILearn only satisfies the conditions with high probability, as in Theorem 13. In this case, the confidence parameter δ used in MILearn should be inversely proportional to the number of iterations of AdaBoost*, which is polynomial in r.

A more specifics analysis can be done in the *realizable case*, where there exists a single hypothesis in $\phi_r^{\max}(\mathcal{H})$ that classifies the training sample perfectly. In this case $\rho^* = \gamma^* = \gamma_+^* = 1$. Assume that $\mathcal{A}$ is a one-sided ERM algorithm, i.e. Theorem 8(a) holds for any weight vector $\mathbf{w}$, with $\eta = 0$. Thus, a margin of $\frac{1}{4r-2}$ can be achieved using AdaBoost* with $N = 8(2r-1)^2 \ln m$ runs of $\mathcal{A}$. If we assume instead a one-sided PAC-learning algorithm $\mathcal{B}(\epsilon, \delta, S)$, we can achieve a similar result with a probability of at least $1 - \delta$, using Theorem 13 with the confidence parameter δ/N. Thus, if the complexity of $\mathcal{A}$ is bounded by a polynomial in $1/\epsilon$ and $1/\delta$, then the complexity of the MIL algorithm is polynomial in r, $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$.

The generalization bound for boosting in Eq. (7) depends on $1/M^2(\alpha, S)$, which is polynomial in r in the cases described above, and on d, the complexity of $\phi_r^{\max}(\mathcal{H})$. For binary hypothesis classes, d is the VC-dimension of $\phi_r^{\max}(\mathcal{H})$, which, by Theorem 3, grows logarithmically with r. For real-valued hypotheses, d is the pseudo-dimension of $\phi_r^{\max}(\mathcal{H})$. Similar generalization results for boosting can be derived for margin-learning as well, using covering-numbers arguments as discussed in Schapire et al. (1998). In Section 5 the sample complexity of MIL with real-valued hypotheses is analyzed, showing that the dependence of the class complexity on r is poly-logarithmic. Thus, in all cases, Eq. (7) implies that the required sample size to achieve learning with error ϵ and confidence $1 - \delta$ is polynomial in r, $\frac{1}{\epsilon}$ and $\ln(\frac{1}{\delta})$.

4.3 From Single-Instance Learning to Multi-Instance Learning

From the discussion in the previous section we can draw the following conclusion on the relationship between single-instance learning and efficient MIL for the realizable case.

Corollary 14 *If there exists a one-sided PAC-learning algorithm for $\mathcal{H}$ whose computational complexity is polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$, then there exists a PAC-learning algorithm for MIL on $\mathcal{H}$ which is polynomial in r,$\frac{1}{\epsilon}$ and $\frac{1}{\delta}$.*

Theorem 1 and Cor. 14 are similar in structure: Both state that if the single-instance problem is solvable with one-sided error, then the realizable MIL problem is solvable. Theorem 1 applies only to bags with statistically independent instances, while Cor. 14 applies to bags drawn from an arbitrary distribution. The assumption of Theorem 1 is weaker, though, as it only requires that the single-instance PAC-learning algorithm handle random one-sided noise, while Cor. 14 requires that the single-instance algorithm handle arbitrary one-sided noise.

Of course, Cor. 14 does not contradict the hardness result provided for APRs in Auer et al. (1998). Indeed, this hardness result states that if there exists a MIL algorithm for d-dimensional APRs which is polynomial in both r and d, then $\mathcal{RP} = \mathcal{NP}$. Our result does not imply that such an algorithm exists, since there is no known agnostic or one-sided PAC-learning algorithm for APRs which is polynomial in d.

Nontheless, MILearn and Cor. 14 provide us with a simple and general way, independent of hypothesis class, to create a PAC-learning algorithm for MIL from a non-MIL one-sided learning algorithm. Whenever an appropriate polynomial algorithm exists for the non-MIL learning problem, the resulting MIL algorithm will also be polynomial in r. To illustrate, consider Shalev-Shwartz et al. (2010), in which an algorithm $\mathcal{B}$ is described for agnostic PAC-learning of fuzzy kernelized half-spaces with an L-Lipschitz transfer function, where L is a constant. The proposed $\mathcal{B}$ has time-complexity and sample-complexity at most $\mathrm{poly}((\frac{L}{\epsilon})^L \cdot \ln(\frac{1}{\delta}))$. Since this complexity bound is polynomial in $1/\epsilon$ and in $1/\delta$, Cor. 14 applies and we can generate an algorithm for PAC-learning MIL with complexity which depends directly on the complexity $\mathcal{B}$, and is polynomial in r, $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$. More generally, using the construction we proposed here, any advancement in the development of algorithms for agnostic or one-sided learning of any hypothesis class translates immediately to an algorithm for PAC-learning MIL with the same hypothesis class, and with corresponding complexity guarantees.

5. MIL with Real-Valued Functions

We now return to the issue of sample complexity, and extend our analysis to hypotheses that range over real values and to large-margin learning. For classes of thresholded functions, we show in Section 5.1 that if the bag classification rule is an extension of a monotone Boolean function, then the sample complexity of MIL depends logarithmically on r, as was shown in Section 3 for binary hypotheses. For margin learning, a poly-logarithmic bound on sample complexity is shown Section 5.2. This bound holds for all Lipschitz bag-labeling functions, including extensions of monotone Boolean functions.

5.1 Thresholded Functions

Monotone Boolean functions map Boolean vectors from $\{-1, +1\}^n$ into $\{-1, +1\}$, such that the map is monotone-increasing in every operand. The set of monotone Boolean functions is exactly the set of functions that can be represented by some composition of AND and OR functions. A natural extension of monotone Boolean functions to real functions from $[-1, +1]^n$ into $[-1, +1]$ is achieved by replacing OR with $\max$ and AND with $\min$. Formally, the real functions that extend monotone Boolean functions are defined as follows:

Definition 15 *A function from* $[-1, +1]^r$ *into* $[-1, +1]$ *is* an extension of an r-ary monotone Boolean function *if it belongs to the set* $\mathcal{M}_r$ *defined inductively as follows, where the input to a function is* $\mathbf{y} \in [-1, +1]^r$*:*

$$\begin{array}{ll} (1)\ \forall j \in [n], & \mathbf{y} \mapsto y[j] \in \mathcal{M}_r; \\ (2)\ \forall k \in \mathbb{N}^+, & f_1, \ldots, f_k \in \mathcal{M}_r \Longrightarrow \mathbf{y} \mapsto \max_{j\in[k]}\{f_j(\mathbf{y})\} \in \mathcal{M}_r; \\ (3)\ \forall k \in \mathbb{N}^+, & f_1, \ldots, f_k \in \mathcal{M}_r \Longrightarrow \mathbf{y} \mapsto \min_{j\in[k]}\{f_j(\mathbf{y})\} \in \mathcal{M}_r. \end{array} \tag{8}$$

In the following theorem we bound the pseudo-dimension (see e.g. Anthony and Bartlett (1999) for definitions) of the generalized MIL problem, where the bag-labeling operator is an extension of a monotone Boolean function. The result has the same form as Theorem 3, which applied to binary hypotheses and Boolean bag-labeling functions.

Theorem 16 *Let* $\mathcal{H} \subseteq [-1, +1]^{\mathcal{X}}$ *be a set of instance hypotheses with pseudo-dimension* d *. Let* $f : [-1, +1]^r \rightarrow [-1, +1]$ *be an extension of a monotone Boolean function, and let* d_r *be the pseudo-dimension of* $\phi^f_r(\mathcal{H})$*. Then*

$$d_r \leq \max\{2d\log(2er), 16\}.$$

Proof For a function h from some domain into $[-1, +1]$ and a scalar $z \in \mathbb{R}$, let h_z be a function from the same domain into $\{-1, +1\}$, defined by $h_z(y) = \operatorname{sign}(h(y) - z)$, where $\operatorname{sign}(x) = +1$ if $x \geq 0$, and $\operatorname{sign}(x) = -1$ otherwise. For a set of functions H, define the set $B_H \triangleq \{h_z \mid h \in H, z \in \mathbb{R}\}$. The pseudo-dimension of H is equal to the VC-dimension of B_H (Anthony and Bartlett, 1999).

Denote $\mathbf{1} = (1, \ldots, 1)$. Using Def. 15, it is easy to verify by induction that for $f \in \mathcal{M}_r$

$$\operatorname{sign}(f(\mathbf{y}) - z) \equiv \operatorname{sign}(f(\mathbf{y} - z\mathbf{1})) \equiv f(\operatorname{sign}(\mathbf{y} - z\mathbf{1})).$$

Consider the thresholded function $\phi^f_r(h)_z$ for $h \in \mathcal{H}$ and $z \in \mathbb{R}$. For all $\bar{\mathbf{x}} \in \mathcal{X}^r$,

$$\begin{aligned} \phi^f_r(h)_z(\bar{\mathbf{x}}) &= \operatorname{sign}(\phi^f_r(h)(\bar{\mathbf{x}}) - z) = \operatorname{sign}(f(h(\bar{\mathbf{x}})) - z) \\ &= f(\operatorname{sign}(h(\bar{\mathbf{x}}) - z\mathbf{1})) = f(h_z(\bar{\mathbf{x}})) = \phi^f_r(h_z)(\bar{\mathbf{x}}). \end{aligned}$$

Therefore, $B_{\phi^f_r(\mathcal{H})} = \phi^f_r(B_{\mathcal{H}})$. Hence d_r is the VC dimension of $\phi^f_r(B_{\mathcal{H}})$. Now, d is the VC-dimension of $B_{\mathcal{H}}$, $B_{\mathcal{H}}$ is a set of hypotheses into $\{-1, +1\}$, and f restricted to $\{-1, +1\}^r$ is also binary. Therefore Theorem 3 applies and the desired bound follows. ■

5.2 Learning with a Margin

To complete the picture for real-valued hypotheses, we address the sample complexity of large-margin classification for MIL, used, for instance, in MI-SVM (Andrews et al., 2002). MI-SVM attempts to optimize an adaptation of the soft-margin SVM objective to MIL, in which the margin of a bag is the maximal margin achieved by any of its instances. It has not been shown, however, whether minimizing the objective function of MI-SVM or analogous margin formulations for MIL allows learning with a reasonable sample size. We fill in this gap in Theorem 17 below, which bounds the γ-Fat-shattering dimension (see e.g. Anthony and Bartlett 1999) of MIL. The objective of MI-SVM amounts to replacing the hypothesis class $\mathcal{H}$ of separating hyperplanes with the class of bag-hypotheses $\phi_r^{\max}(\mathcal{H})$. Since $\max$ is the real-valued extension of OR, this objective function is natural in our formulation. Our result applies more generally to any bounded c-Lipschitz bag-labeling function, and to any hypothesis class over instances.

A function $f : \mathbb{R}^r \rightarrow \mathbb{R}$ is c-Lipschitz with respect to the infinity norm if

$$\forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^r, |f(\mathbf{a}) - f(\mathbf{b})| \leq c\|\mathbf{a} - \mathbf{b}\|_\infty.$$

It is easy to verify using induction on $\mathcal{M}_r$ that extensions of monotone Boolean functions are 1-Lipschitz with respect to the infinity norm. For $\gamma > 0$, denote by $\mathcal{F}_1(\gamma)$ the γ-fat-shattering dimension of $\mathcal{H}$, and by $\mathcal{F}_r(\gamma)$ the γ-fat-shattering dimension of $\phi_r^f(\mathcal{H})$.

Theorem 17 *Let $B, c > 0$. Let $\mathcal{H} \subseteq [0, B]^\mathcal{X}$ be a real-valued hypothesis class and let $f : [0, B]^r \rightarrow [0, cB]$ be a function which is c-lipschitz with respect to the infinity norm. Then*

$$\mathcal{F}_r(\gamma) \leq 6\mathcal{F}_1(\frac{\gamma}{64c}) \log^2(2048\frac{B^2c^4}{\gamma^2} r\mathcal{F}_r(\gamma)). \tag{9}$$

Before turning to the proof of Theorem 17, we note that the bound in Eq. (9) is in implicit form, since $\mathcal{F}_r(\gamma)$ appears on both sides of the bound. To better understand its meaning, we restate the bound as a function of r. Fixing γ and $\mathcal{F}_1(\gamma/(64c))$ and setting $\beta = 6\mathcal{F}_1(\gamma/(64c))$ and $\eta = 2048B^2c^4/\gamma^2$, we have

$$\sqrt{\mathcal{F}_r} - \sqrt{\beta}\log\mathcal{F}_r \leq \sqrt{\beta}\log(\eta r). \tag{10}$$

Therefore Eq. (9) indicates a poly-logarithmic bound on $\mathcal{F}_r$.

To prove Theorem 17, we use the *covering number* of the single-instance and multi-instance hypothesis classes. For $\mathcal{H} \subseteq [0, B]^\mathcal{X}$, $\gamma > 0$ and $S \subseteq \mathcal{X}$, the set of γ-*covers* of S by $\mathcal{H}$ is

$$\mathrm{cov}_\gamma(\mathcal{H}, S) \triangleq \{C \subseteq \mathcal{H} \mid \forall h \in \mathcal{H}, \exists \hat{h} \in C, \max_{s \in S} |h(s) - \hat{h}(s)| \leq \gamma\}.$$

The γ-*covering number* of a hypothesis class $\mathcal{H} \subseteq [0, B]^\mathcal{X}$ and a number $m \in \mathbb{N}$ is

$$\mathcal{N}_\infty(\gamma, \mathcal{H}, m) \triangleq \max_{S \subseteq \mathcal{X}:|S|=m} \min_{C \in \mathrm{cov}_\gamma(\mathcal{H},S)} |C|.$$

The following two theorems link the covering number of a function class with its fat-shattering dimension. For a function class F, denote its γ-Fat shattering dimension by $\mathrm{Fat}_F(\gamma)$.

Theorem 18 (Anthony and Bartlett, 1999, Theorem 12.10) *Let F be a set of real functions and let $\gamma > 0$. For $m \geq \mathrm{Fat}_F(16\gamma)$,*

$$e^{\mathrm{Fat}_F(16\gamma)/8} \leq \mathcal{N}_\infty(\gamma, F, m). \tag{11}$$

Theorem 19 (Anthony and Bartlett, 1999, Theorem 12.8) *Let F be a set of real functions from a domain $\mathcal{X}$ to the bounded interval $[0, B]$. Let $\gamma > 0$. Let $d = \mathrm{Fat}_F(\frac{\gamma}{4})$. For all $m \geq d$,*

$$\mathcal{N}_\infty(\gamma, F, m) < 2\left(\frac{4B^2m}{\gamma^2}\right)^{d\log\frac{4eBm}{d\gamma}}. \tag{12}$$

To prove Theorem 17, we bound the covering number of a bag hypothesis class by the covering number of the single-instance hypothesis class:

Lemma 20 *Let $f : [0, B]^r \to [0, cB]$ be c-Lipschitz with respect to the infinity norm for some $c > 0$. For any natural $m, r > 0$, and real $\gamma > 0$, and for any hypothesis class $\mathcal{H} \subseteq [0, B]^{\mathcal{X}}$,*

$$\mathcal{N}_\infty(c\gamma, \phi_r^f(\mathcal{H}), m) \le \mathcal{N}_\infty(\gamma, \mathcal{H}, rm). \tag{13}$$

Proof Let $S = \{\bar{\mathbf{x}}_i\}_{i\in[m]} \subseteq \mathcal{X}^r$ be a set of m bags. Let $S^\cup = \{x_i[j]\}_{i\in[m],j\in[r]}$ be the set of instances in bags of S. Let $C \in \mathrm{cov}_\gamma(\mathcal{H}, S^\cup)$ be a γ-cover of $S^\cup$. For all $h \in \mathcal{H}$ there exists an $\hat{h} \in C$ such that $\max_{i\in[m]} \|h(\bar{\mathbf{x}}_i) - \hat{h}(\bar{\mathbf{x}}_i)\|_\infty \le \gamma$. From the Lipschitz condition on f we have

$$|\phi_r^f(h)(\bar{\mathbf{x}}_i) - \phi_r^f(\hat{h})(\bar{\mathbf{x}}_i)| \equiv \|f(h(\mathbf{x}_i)) - f(\hat{h}(\mathbf{x}_i))\|_\infty \le c\|h(\mathbf{x}_i) - \hat{h}(\mathbf{x}_i)\|_\infty \le c\gamma.$$

Since $\phi_r^f(\hat{h}) \in \phi_r^f(C)$, it follows that $\phi_r^f(C) \in \mathrm{cov}_{c\gamma}(\phi_r^f(\mathcal{H}), S)$. Since this is true for all $C \in \mathrm{cov}_\gamma(\mathcal{H}, S^\cup)$, we have $\phi_r^f(\mathrm{cov}_\gamma(\mathcal{H}, S^\cup)) \subseteq \mathrm{cov}_{c\gamma}(\phi_r^f(\mathcal{H}), S)$. Therefore,

$$\begin{aligned}
\mathcal{N}_\infty(c\gamma, \phi_r^f(\mathcal{H}), m) &\equiv \max_{S\subseteq\mathcal{X}^r:|S|=m} \; \min_{\phi_r^f(C)\in\mathrm{cov}_{c\gamma}(\phi_r^f(\mathcal{H}),S)} |\phi_r^f(C)| \\
&\le \max_{S\subseteq\mathcal{X}^r:|S|=m} \; \min_{\phi_r^f(C)\in\phi_r^f(\mathrm{cov}_\gamma(\mathcal{H},S^\cup))} |\phi_r^f(C)| \\
&= \max_{S\subseteq\mathcal{X}^r:|S|=m} \; \min_{C\in\mathrm{cov}_\gamma(\mathcal{H},S^\cup)} |C| \\
&= \max_{S\subseteq\mathcal{X}:|S|=rm} \; \min_{C\in\mathrm{cov}_\gamma(\mathcal{H},S)} |C| = \mathcal{N}_\infty(\gamma, \mathcal{H}, rm).
\end{aligned}$$

■

Proof [of Theorem 17] From Theorem 18 and Lemma 20 it follows that for $m \ge \mathcal{F}_r(16\gamma)$,

$$\mathcal{F}_r(16\gamma) \le \frac{8}{\log e} \log \mathcal{N}_\infty(\gamma, \phi_r^f(\mathcal{H}), m) \le 6 \log \mathcal{N}_\infty(\gamma/c, \mathcal{H}, rm).$$

In addition, from Eq. (12) we have that if $m \ge d \triangleq \mathrm{Fat}_F(\frac{\gamma}{4}) \ge 1$ and F is into $[0, B]$ then, for $\gamma \le B/e$,

$$\log \mathcal{N}_\infty(\gamma, \mathcal{H}, m) < d \log^2(\frac{4B^2 m}{\gamma^2}) = \mathrm{Fat}_F(\frac{\gamma}{4}) \log^2(\frac{4B^2 m}{\gamma^2}).$$

Combining the two inequalities and substituting B with cB, we get that if $m \ge \mathcal{F}_r(16\gamma)$ and $rm \ge \mathcal{F}_1(\frac{\gamma}{4c}) \ge 1$, then

$$\mathcal{F}_r(16\gamma) \le 6\mathcal{F}_1(\frac{\gamma}{4c}) \log^2(\frac{4B^2c^4 rm}{\gamma^2}).$$

Setting $m = \lceil \mathcal{F}_r(16\gamma) \rceil \le \mathcal{F}_r(16\gamma) + 1$, it follows that if $\mathcal{F}_r(16\gamma) \ge 1$ and $\mathcal{F}_r(16\gamma) \ge \mathcal{F}_1(\frac{\gamma}{4c})/r \ge \frac{1}{r}$, then

$$\mathcal{F}_r(16\gamma) \le 6\mathcal{F}_1(\frac{\gamma}{4c}) \log^2(4\frac{B^2c^4}{\gamma^2} r(\mathcal{F}_r(16\gamma) + 1)) \le 6\mathcal{F}_1(\frac{\gamma}{4c}) \log^2(8\frac{B^2c^4}{\gamma^2} r\mathcal{F}_r(16\gamma)).$$

Substituting 16γ with γ, we have that the bound in Eq. (9) holds for $\gamma/16 \le B/e$, which trivially holds since $\gamma \le B$. ■

6. Conclusions

In this work we have provided a new theoretical analysis for Multiple Instance Learning with any underlying hypothesis class. We have shown that the dependence of the sample complexity of generalized MIL on the number of instances in a bag is only poly-logarithmic, thus implying that the performance of MIL is only mildly sensitive to the size of the bag. The analysis includes binary hypotheses, real-valued hypotheses, and margin learning, all of which are used in practice in MIL applications. For classical MIL, where the bag-labeling function is the Boolean OR, we have shown a new learning algorithm, that classifies bags by accessing a learning algorithm designed for single instances. This algorithm provably PAC-learns MIL. In both the sample complexity analysis and the computational analysis, we have shown tight connections between classical supervised learning and Multiple Instance Learning. This connection holds regardless of the underlying hypothesis class.

Acknowledgements

The authors thank Shai Shalev-Shwartz and Nati Srebro for helpful discussions. Enlightening comments from Phil Long and Sally Goldman on the preliminary version of this work have helped improve the presentation immensely. Sivan Sabato is supported by the Adams Fellowship Program of the Israel Academy of Sciences and Humanities.

References

S. Andrews. *Learning from ambiguous examples*. PhD thesis, Brown University, May 2007.

S. Andrews and T. Hofmann. Multiple-instance learning via disjunctive programming boosting. In *Advances in Neural Information Processing Systems 16*, 2003.

S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems 15*, pages 561–568, 2002.

M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

P. Auer, P. M. Long, and A. Srinivasan. Approximating hyper-rectangles: learning and pseudorandom sets. *Journal of Computer and System Sciences*, 57(3):376–388, 1998.

A. Blum and A. Kalai. A note on learning from multiple-instance examples. *Machine Learning*, 30(1): 23–29, 1998.

T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

Y. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, August 1997.

T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola. Multi-instance kernels. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 179–186, 2002.

P. M. Long and L. Tan. PAC learning axis-aligned rectangles with respect to product distributions from multiple-instance examples. *Machine Learning*, 30(1):7–21, 1998. ISSN 0885-6125.

O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *Advances in Neural Information Processing Systems 10*, pages 570–576, 1998.

O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 341–349, 1998.

L. Pitt and L. G. Valiant. Computational limitations on learning from examples. Technical report, Harvard University Aiken Computation Laboratory, July 1986.

L. De Raedt. Attribute-value learning versus inductive logic programming: The missing links (extended abstract). In *Proceedings of the 8th International Workshop on Inductive Logic Programming*, pages 1–8, London, UK, 1998. Springer-Verlag.

G. Rätsch and M. K. Warmuth. Efficient margin maximizing with boosting. *Journal of Machine Learning Research*, 6:2131–2152, December 2005.

S. Sabato and N. Tishby. Homogeneous multi-instance learning with arbitrary dependence. In *Proceedings of the Twenty Second Annual Conference on Computational Learning Theory*, 2009.

S. Sabato, N. Srebro, and N. Tishby. Reducing label complexity by learning from bags. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, pages 685–692, 2010.

N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory Series A*, 13:145–147, 1972.

R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):1–40, 1999.

R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, October 1998.

S. Shalev-Shwartz and Y. Singer. On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms. In *Proceedings of the Twenty First Annual Conference on Computational Learning Theory*, pages 311–322, 2008.

S. Shalev-Shwartz, O. Shamir, and K. Sridharan. Learning kernel-based halfspaces with the zero-one loss. In *Proceedings of the Twenty Third Annual Conference on Computational Learning Theory*, 2010.

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its applications*, XVI(2):264–280, 1971.

J. von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928.

M. Warmuth, K. Glocer, and G. Ratsch. Boosting algorithms for maximizing the soft margin. In *Advances in Neural Information Processing Systems 21*, 2007.

N. Weidmann, E. Frank, and B. Pfahringer. A two-level learning method for generalized multi-instance problems. In *Proceedings of the European conference on machine learning*, pages 468–479, 2003.

Q. Zhang and S.A. Goldman. EM-DD: An improved multiple-instance learning technique. In *Advances in Neural Information Processing Systems 14*, 2001.

Z. Zhou, K. Jiang, and M. Li. Multi-instance learning based web mining. *Applied Intelligence*, 22(2): 135–147, 2005.

Appendix A. Technical Proofs

Proof [of Lemma 9] Let α be defined as in `MILearn`, so that $\alpha(+1) = \frac{1}{r}$ and $\alpha(-1) = 1$. We have

$$W(S_I)\Gamma(h, S_I) = \sum_{i \in [m], j \in [r]} \alpha(y_i) w_i y_i h(x_i[j]) = \sum_{y_i=+1} \tfrac{1}{r} w_i \sum_{j \in [r]} h(x_i[j]) - \sum_{y_i=-1} w_i \sum_{j \in [r]} h(x_i[j])$$
$$\leq \sum_{y_i=+1} w_i \max_{j \in [r]} h(x_i[j]) - \sum_{y_i=-1} w_i (\max_{j \in [r]} h(x_i]j]) - r + 1) \leq \sum_{i \in [m]} w_i y_i \phi_r^{\max}(h)(\bar{\mathbf{x}}_i) + (r-1) \sum_{y_i=-1} w_i$$
$$= \Gamma(\phi_r^{\max}(h), S_B) + (r-1) W_-.$$

■

Proof [of Lemma 10]

Let $h \in \Omega(S_B)$. From the definition of Ω, for any i such that $y_i = -1$, $\phi_r^{\max}(h)(\bar{\mathbf{x}}_i) = -1$. Thus for all $j \in [r]$, $h(x_i[j]) = -1$. It follows that $\mathbb{P}_{S_I}[h(X) = -1 \mid Y = -1] = 1$, thus $h \in \Omega(S_I)$. Therefore $\Omega(S_B) \subseteq \Omega(S_I)$. By the condition on $\mathcal{A}$ in Theorem 8(a), $\Gamma(h_I, S_I) \geq \max_{h \in \mathcal{H} \cap \Omega(S_I)} \Gamma(h, S_I) - \eta$. In addition, since $h_+^* \in \mathcal{H} \cap \Omega(S_B)$, we have

$$\max_{h \in \mathcal{H} \cap \Omega(S_I)} \Gamma(h, S_I) \geq \max_{h \in \mathcal{H} \cap \Omega(S_B)} \Gamma(h, S_I) \geq \Gamma(h_+^*, S_I).$$

Therefore $\Gamma(h_I, S_I) \geq \Gamma(h_+^*, S_I) - \eta$. ■

Proof [of Lemma 11]

$$W(S_I)\Gamma(h, S_I) = \sum_{i \in [m]} \alpha(y_i) w_i y_i \sum_{j \in [r]} h(x_i[j])$$
$$= \sum_{y_i=+1} \tfrac{1}{r} w_i \sum_{j \in [r]} h(x_i[j]) - \sum_{y_i=-1} w_i \sum_{j \in [r]} h(x_i[j])$$
$$\geq \sum_{y_i=+1} \tfrac{1}{r} w_i (\max_{j \in [r]} h(x_i[j]) - r + 1) - r \sum_{y_i=-1} w_i \max_{j \in [r]} h(x_i[j])$$
$$= \tfrac{1}{r} \Gamma(\phi_r^{\max}(h), S_B) + (\tfrac{1}{r} - 1) W_+ - (r - \tfrac{1}{r}) \sum_{y_i=-1} w_i \phi_r^{\max}(h)(\bar{\mathbf{x}}_i)$$

■

Proof [of Theorem 8(b)] Denote $h^* \triangleq \operatorname{argmax}_{h \in \mathcal{H}} \Gamma(\phi_r^{\max}(h), S_B)$. From Lemma 9, Lemma 11 the assumptions in Theorem 8(b),

$$\Gamma(\phi_r^{\max}(h_I), S_B) \geq W(S_I)\Gamma(h_I, S_I) + (1-r)W_- \geq W(S_I)(\Gamma(h^*, S_I) - \eta) + (1-r)W_-$$
$$\geq \tfrac{1}{r}\gamma^* + (\tfrac{1}{r} - 1)W_+ + (1-r)W_- - (r - \tfrac{1}{r}) \sum_{y_i=-1} w_i \phi_r^{\max}(h^*)(\bar{\mathbf{x}}_i) - r\eta.$$

Now,

$$-\sum_{y_i=-1} w_i \phi_r^{\max}(h^*)(\bar{\mathbf{x}}_i) = \sum_{y_i=-1} w_i y_i \phi_r^{\max}(h^*)(\bar{\mathbf{x}}_i) = \gamma^* - \sum_{y_i=+1} w_i y_i \phi_r^{\max}(h^*)(\bar{\mathbf{x}}_i) \geq \gamma^* - W_+.$$

Therefore

$$\Gamma(\phi_r^{\max}(h_I), S_B) \geq \tfrac{1}{r}\gamma^* + (\tfrac{1}{r}-1)W_+ + (1-r)W_- + (r-\tfrac{1}{r})(\gamma^*-W_+) - r\eta = r\gamma^* + 1 - r - (2-\frac{2}{r})W_+ - r\eta,$$

where the last equality follows from $W_- + W_+ = 1$. By step 4 of `MILearn`,

$$\gamma = \max\{\Gamma(\phi_r^{\max}(h_I), S_B), \Gamma(\phi_r^{\max}(h_{\mathrm{pos}}), S_B)\} \geq \max\left\{r\gamma^* + 1 - r - (2 - \frac{2}{r})W_+ - r\eta,\ 2W_+ - 1\right\}.$$

It is easy to verify that for any $W_+ \in [0,1]$, $\gamma \geq \frac{r^2(\gamma^*-\eta-1)+1}{2r-1}$. To guarantee $\gamma \geq 0$, we require $\gamma^* - \eta \geq 1 - \frac{1}{r^2}$. ■