Homogeneous Multi-Instance Learning with Arbitrary Dependence

Sivan Sabato and Naftali Tishby
School of Comp. Sci. & Eng., The Hebrew University, Jerusalem 91904, Israel
{sivan_sabato,tishby}@cs.huji.ac.il

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 a Boolean OR. The learner observes the bag labels but not the instance labels that generated them. MIL has numerous applications, and many heuristic algorithms have been used successfully on this problem. However, no guarantees on the result or generalization bounds have been shown for these algorithms. At the same time, theoretical analysis has shown MIL to be either trivial or too hard, depending on the assumptions. In this work we formally define a new setting which is more relevant for MIL applications than previous theoretic assumptions. The sample complexity of this setting is shown to be only logarithmically dependent on the size of the bag, and for the case of Boolean OR, an algorithm with proven guarantees is provided. We further extend the sample complexity results to a real-valued generalization of MIL.

1 Introduction

We consider the learning problem termed Multiple-Instance Learning (MIL), first introduced in [DLLP97]. MIL is a generalization of the classical supervised learning problem with binary labels. As in the classical setting, 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 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 the original MIL setting the label of a bag is determined using a simple rule: it is the Boolean OR of the labels of the instances it contains. Various generalizations to MIL have been proposed [Rae98, WFP03]. We consider the generalization obtained by replacing OR with an arbitrary Boolean function. To differentiate the two settings, we shall refer to the original setting as OR-MIL and reserve the term MIL for the generalized problem. Note that it is possible in principle to treat a MIL problem as a regular classification task, viewing a bag as a single example, where the instances in a bag are treated as a mere part of its internal representation. Such treatment, however, cannot take advantage of the special structure of a MIL problem. In particular, it does not allow the possibility of employing known techniques for classifying instances to help classify bags.

There are numerous applications for OR-MIL. In [DLLP97], the drug design application motivates this setting: in this problem, 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. Other applications include image classification [MR98], web index page recommendation [ZJL05], and text categorization [And07].

Several heuristic algorithms have been proposed for OR-MIL, usually in the context of a specific application. For instance, Dietterich et. al [DLLP97] propose algorithms for finding an Axis-Parallel Rectangle (APR) that predicts the label of an instance and of a bag. Diverse Density [MLP98], EM-DD [ZG01] use assumptions on a specific structure of the bags of instances. DP-Boost [AH03], mi-SVM and MI-SVM [ATH02] are heuristic approaches for learning OR-MIL using margins. Many of the proposed algorithms work well in practice on relevant data sets, however none of the works above provide generaliza-

tion guarantees for any of the algorithms.

Previous theoretical analysis of MIL has focused on OR-MIL. Two different settings of OR-MIL have been investigated. In the first setting, it is assumed that each of the instances in each bag may be classified using a different classification rule. We refer to such a bag classification rule as *heterogeneous*. In addition, it is assumed that the bags are drawn according to an arbitrary distribution over bags, so that the instances in a bag may be statistically dependent. We thus term this setting *heterogeneous-dependent*. In [ALS98] the problem of learning APRs in this setting is investigated. It is shown that PAC-learning a disjunction of r APRs in $\mathbb{R}^n$ using a sample of labeled bags is as hard as learning DNF formulas. Thus a PAC algorithm which is polynomial in r and n exists only if $\mathcal{RP} = \mathcal{NP}$ [PV86].

In the second setting that has been theoretically investigated [ALS98, BK98, LT98] it is assumed that all the instances in a bag are classified according to the same classification rule, so that the bag classification rule is *homogeneous*. In addition, it is assumed that the instances in all bags are drawn i.i.d from a single arbitrary distribution over instances, so that the instances in a bag are statistically independent. We thus term this setting *homogeneous-independent*. In [BK98] the following theorem is proven for homogeneous-independent OR-MIL:

Theorem 1 (Blum and Kalai, in [BK98]) *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 OR-MIL, assuming that the instances in all bags are drawn i.i.d from a single arbitrary distribution.*

This theorem is based on the fact that in an independent setting the sample of bags may be used to build an i.i.d. sample of instances with one-sided noise. The algorithms proposed in [BK98] are polynomial in both the sample size and in the number of instances in a bag.

Examining the above-mentioned theoretical results, it can be seen that on the one hand, homogeneous-independent OR-MIL is provably easy. However, its bearing to actual application domains where OR-MIL is used is questionable, since in almost all applications it is not possible to assume that the instances in a bag are even approximately independent. On the other hand, heterogeneous--dependent OR-MIL is applicable to an extensive set of problems, but is provably hard in the worst case. In addition, this setting includes the implicit assumption that instances in a bag are ordered, whereas in many applications this ordering is nonexistent or not informative.

In contrast, most (if not all) heuristic algorithms that have been researched and used in practice, including the algorithms listed above, learn a classification rule that is identical for all instances in a bag, as in a homogeneous setting. In addition, these algorithms operate on problems in which the bag is drawn from an arbitrary distribution, as in a dependent setting. In this work we thus define a new MIL setting, which is *homogeneous-dependent*. Our sample complexity analysis shows that the sample size required for MIL depends on the number of instances in a bag only logarithmically. For the computational aspect, we show an algorithm with proven guarantees for OR-MIL. Our analysis thus reduces the gap between existing theory, where even OR-MIL is either trivial or too hard, and practical algorithms, which exemplify good results in many application domains.

In Section 2 the problem setting is defined and notations are provided. In Section 3 the sample complexity of heterogeneous-dependent MIL and homogeneous-dependent MIL with binary hypotheses are compared, showing that homogeneous-dependent is strictly easier. Section 4 provides an algorithm with result guarantees for learning in the homogeneous-dependent OR-MIL setting. In Section 5 we analyze the sample complexity of the generalized MIL setting with real-valued functions. We conclude with a discussion in Section 6. Appendix A provides proofs that have been skipped in the text. .

2 Problem Setting and Notation

Let r be some positive natural number. We assume throughout this work that r is the number of instances in a bag. Let X be the domain of instances. We assume that bags are ordered sequences of r instances from X, thus the domain of bags is X^r.[1] It is assumed that an unknown classification rule $h : X \to \{-1, +1\}$ labels instances in X, and that the label of a bag is determined by the labels of the instances it contains. In classical MIL, which we term OR-MIL, the label of a bag is the Boolean OR of the labels of its instances. In generalized MIL the label of a bag is some r-ary Boolean function of the labels of its instances. Importantly, this function is known to the learner a-priori.

The learner receives a sample of labeled bags. The labels of instances in the bags remain unobserved. In dependent settings, the sample of bags is drawn i.i.d from an unknown and arbitrary distribution over X^r. The goal of the learner is to find a classification rule that would classify new *bags* drawn from the same distribution with low error. Note that in a dependent setting it is not possible in the general case to find a low-error classification rule for instances. As a simple counter example assume that in OR-MIL every bag includes both a positive instance and a negative instance. In this case all bags are labeled as positive, and it is not

[1] Note that our assumption that instances are ordered within a bag is merely a technicality if the bag classification rule is symmetric.

possible to distinguish the two types of instances by observing only bag labels.

We now turn to define notational conventions. Vectors are marked in boldface and elements in a vector are denoted by superscripts, so that $\mathbf{x} = (x^1, \ldots, x^k)$ for a k-element vector. Bags are considered vectors of instances and are marked accordingly. We also use the vector notation to denote vectors of functions. The following list summarizes the notations for functions and vectors, where $\mathbf{a}$ is some vector, f is a scalar function, $\mathbf{f} \triangleq (f^1, \ldots, f^k)$ is a vector of scalar functions, and $\hat{\mathbf{f}} \triangleq (\hat{f}^1, \ldots, \hat{f}^k)$ is a vector of functions from vectors to scalars:

$$\begin{aligned} f(\mathbf{a}) &\triangleq (f(a^1), \ldots, f(a^k)), \\ \mathbf{f}(a) &\triangleq (f^1(a), \ldots, f^k(a)), \\ \mathbf{f}(\mathbf{a}) &\triangleq (f^1(a^1), \ldots, f^k(a^k)), \\ \hat{\mathbf{f}}(\mathbf{a}) &\triangleq (\hat{f}^1(\mathbf{a}), \ldots, \hat{f}^k(\mathbf{a})). \end{aligned}$$

Unless otherwise mentioned, vectors have r elements. $\mathbf{x} \cdot \mathbf{y}$ denotes the dot product of two real vectors $\mathbf{x}$ and $\mathbf{y}$. The notations $\|.\|_\infty$ and $\|.\|_1$ denote the infinity norm and the L_1 norm respectively . 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.

We define two operators that map a classification rule over instances into a classification rule over bags: one for the heterogeneous setting and one for the homogeneous setting.

Definition 2 *Let Y be some domain, and let $f : Y^r \to Y$.* The heterogeneous bag-labeling operator*, denoted by ψ_r^f, is a function mapping r hypotheses over instances to a hypothesis over bags, defined as follows:*

$$\begin{aligned} &\psi_r^f : \left(Y^X\right)^r \to Y^{X^r}, \text{ and} \\ &\forall \mathbf{h} = (h^1, \ldots, h^r) \in \left(Y^X\right)^r, \mathbf{x} \in X^r, \qquad (1) \\ &\quad \psi_r^f(\mathbf{h})(\mathbf{x}) \triangleq f(\mathbf{h}(\mathbf{x})) \equiv f(h^1(x^1), \ldots, h^r(x^r)). \end{aligned}$$

Definition 3 *Let Y be some domain, and let $f : Y^r \to Y$.* The homogeneous bag-labeling operator*, denoted by ϕ_r^f, is a function mapping a hypothesis over instances to a hypothesis over bags, defined as follows:*

$$\begin{aligned} &\phi_r^f : Y^X \to Y^{X^r}, \text{ and} \\ &\forall h \in Y^X, \mathbf{x} \in X^r, \qquad (2) \\ &\quad \phi_r^f(h)(\mathbf{x}) \triangleq f(h(\mathbf{x})) \equiv f(h(x^1), \ldots, h(x^r)). \end{aligned}$$

Setting $f \triangleq$ OR in ψ_r^f and in ϕ_r^f we have the OR-MIL problem in the heterogeneous setting and in the homogeneous setting respectively.

We use $\mathcal{H}$ to denote a set of hypotheses on instances. Hypotheses may be binary, so that $\mathcal{H} \subseteq \{-1, +1\}^X$, or they may be real-valued, so that $\mathcal{H} \subseteq [-1, +1]^X$. The assumptions on $\mathcal{H}$ will be specified in context.

We denote by $\psi_r^f(\mathcal{H})$ the set of hypotheses over bags that are generated from $\mathcal{H}$ by ψ_r^f :

$$\psi_r^f(\mathcal{H}) = \{\psi_r^f(\mathbf{h}) \mid \mathbf{h} \in \mathcal{H}^r\}.$$

Similarly, $\phi_r^f(\mathcal{H})$ is the set of hypotheses over bags that are generated from $\mathcal{H}$ by ϕ_r^f :

$$\phi_r^f(\mathcal{H}) = \{\phi_r^f(h) \mid h \in \mathcal{H}\}.$$

3 Sample Complexity for Binary Hypotheses

It is shown in [ALS98] that heterogeneous-dependent OR-MIL is computationally hard for APRs, however for other hypothesis classes this setting may be computationally feasible. In this section we show that notwithstanding the computational question, in terms of sample complexity heterogeneous-dependent MIL is strictly harder than homogeneous-dependent MIL. We show that for any non-trivial Boolean function, the VC-dimension of heterogeneous-dependent MIL is at least linear in r, the number of instances in a bag, while the VC-dimension of homogeneous-dependent MIL is at most logarithmic in r. We start with a lower bound on the VC-dimension in heterogeneous-dependent MIL.

Theorem 4 *Let $r \in \mathbb{N}^+$. Let $f : \{-1, +1\}^r \to \{-1, +1\}$ be an r-ary Boolean function that is not constant in any of its operands. Let ψ_r^f be the heterogeneous bag-labeling operator defined as in Eq. (1). Let $\mathcal{H} \subseteq \{-1, +1\}^X$ be a hypothesis class with a finite VC-dimension d_I, and denote the VC-dimension of $\psi_r^f(\mathcal{H})$ by d_B. Then*

$$d_B \geq r(d_I - 2).$$

Proof: Let $S_I = \{a_1, \ldots, a_{d_I}\} \subseteq X$ be a set of instances of size d_I that can be shattered by $\mathcal{H}$. Assume that $d_I > 2$, otherwise the bound trivially holds. Let $\mathbf{e}_{(y,j)} = (1, \ldots, 1, y, 1, \ldots, 1)$ be a vector with r elements, where element j equals y. For every $j \in [r]$, let $\mathbf{c}_j$ be a vector of r Boolean values such that

$$\forall y \in \{-1, +1\}, \quad f(\mathbf{c}_j \cdot \mathbf{e}_{(y,j)}) = y. \qquad (3)$$

Since f is not constant in any of its operands, such a vector exists for all $j \in [r]$.

We define a set of bags $S_B \triangleq \{\mathbf{x}_i\} \subseteq X^r$ of size $r(d_I - 2)$ and show that it can be shattered by $\psi_r^f(\mathcal{H})$. Letting $J(i) \triangleq \lfloor \frac{i-1}{d_I - 2} \rfloor + 1$, define bag $\mathbf{x}_i$ as follows:

$$x_i^j \triangleq \begin{cases} a_{[(i-1) \bmod (d_I-2)]+1} & \text{if } J(i) = j; \\ a_{d_I - 1} & \text{if } J(i) \neq j \ \& \\ & \quad c^j_{J(i)} = -1; \\ a_{d_I} & \text{if } J(i) \neq j \ \& \\ & \quad c^j_{J(i)} = +1. \end{cases}$$

We now show that for any labeling $Y_B \triangleq \{y_i\}_{i\in[r(d_I-2)]}$ there exists a hypothesis in $\psi_r^f(\mathcal{H})$ that assigns Y_B to S_B. For a given Y_B define d_I Boolean vectors $\mathbf{t}_k$ as follows:

$$t_k^j \triangleq \begin{cases} c_j^j \cdot y_{(j-1)(d_I-2)+k} & \text{if } k \le d_I-2; \\ -1 & \text{if } k = d_I-1; \\ +1 & \text{if } k = d_I. \end{cases}$$

Since the VC-dimension of $\mathcal{H}$ is d_I, we have that for all $j \in [r]$ there exists a vector of r hypotheses $\mathbf{h} \triangleq (h^1, \ldots, h^r) \in \mathcal{H}^r$ such that $\mathbf{h}(a_k) = \mathbf{t}_k$ for all $k \in [d_I]$. Set $\hat{h} \triangleq \psi_r^f(\mathbf{h}) \in \psi_r^f(\mathcal{H})$. From the definition of $\mathbf{x}_i$ we have

$$h^j(x_i^j) = \begin{cases} t^j_{[(i-1) \bmod (d_I-2)]+1} & \text{if } J(i) = j; \\ t^j_{d_I-1} & \text{if } J(i) \neq j \;\&\; c^j_{J(i)} = -1; \\ t^j_{d_I} & \text{if } J(i) \neq j \;\&\; c^j_{J(i)} = +1. \end{cases}$$

Using the definition of t_k^j above, we find that for $j \neq J(i)$, $h^j(x_i^j) = c^j_{J(i)}$, and that for $j = J(i)$,

$$\begin{aligned} h^{J(i)}(x_i^{J(i)}) &= t^{J(i)}_{[(i-1) \bmod (d_I-2)]+1} = \\ &= c^{J(i)}_{J(i)} \cdot y_{(J(i)-1)(d_I-2)+[(i-1) \bmod (d_I-2)]+1} \\ &= c^{J(i)}_{J(i)} \cdot y_{\lfloor \frac{i-1}{d_I-2} \rfloor (d_I-2)+[(i-1) \bmod (d_I-2)]+1} \\ &= c^{J(i)}_{J(i)} \cdot y_i. \end{aligned}$$

Therefore, $\mathbf{h}(\mathbf{x}_i) = \mathbf{c}_{J(i)} \cdot \mathbf{e}_{(y_i, J(i))}$. By the definition of $\hat{h}$ and Eq. (3),

$$\begin{aligned} &\forall i \in [r(d_I-2)], \\ &\hat{h}(\mathbf{x}_i) = \psi_r^f(h)(\mathbf{x}_i) = f(\mathbf{h}(\mathbf{x}_i)) = \\ &= f(\mathbf{c}_{J(i)} \cdot \mathbf{e}_{(y_i, J(i))}) = y_i. \end{aligned}$$

We showed a set S_B of size $r(d_I-2)$ that can be shattered by $\psi_r^f(\mathcal{H})$, therefore $d_B \ge r(d_I-2)$. ■

Having shown a lower bound for the sample complexity of the heterogeneous-dependent setting that is linear in r, we now show that the sample complexity for homogeneous-dependent MIL is no more than logarithmic in r.

Theorem 5 *Let $f : \{-1,+1\}^r \to \{-1,+1\}$ be an r-ary Boolean function. Let ϕ_r^f be the homogeneous bag-labeling operator defined as in Eq. (2). Let $\mathcal{H} \subseteq \{-1,+1\}^X$ be a hypothesis class with a finite VC-dimension d_I. Denote the VC-dimension of the hypothesis class $\phi_r^f(\mathcal{H})$ by d_B. Then*

$$d_B \le \max\{2d_I(\log r - \log d_I + \log e), 4{d_I}^2, 16\}.$$

Proof: In the following proof the notation $\mathcal{X}_{|_A}$, for a set of hypotheses $\mathcal{X}$ and a set of examples A, denotes the restriction of the hypotheses in $\mathcal{X}$ to their values on the set A, so that

$$\mathcal{X}_A \triangleq \{h_{|_A} \mid h \in \mathcal{X}\}.$$

By the definition of VC-dimension, there exists a set of bags $S = \{\mathbf{x}_i\}_{i\in[d_B]} \subseteq X^r$ that is shattered by $\phi_r^f(\mathcal{H})$. There are 2^{d_B} different ways to label the bags in S, thus $|\phi_r^f(\mathcal{H})_{|_S}| = 2^{d_B}$. Let $S^\cup = \{x_i^j\}_{i\in[m],j\in[r]}$ be the set of instances in bags in S. Lemma 20, proven in Appendix A, states that $|\phi_r^f(\mathcal{H})_{|_S}| \le |\mathcal{H}_{|_{S^\cup}}|$. Therefore $2^{d_B} \le |\mathcal{H}_{|_{S^\cup}}|$. We have $|S^\cup| \le r|S| = rd_B$. Therefore, by Sauer's lemma [Sau72, VC71], $|\mathcal{H}_{|_{S^\cup}}|$ is bounded as follows:

$$2^{d_B} \le |\mathcal{H}_{|_{S^\cup}}| \le \left(\frac{e|S^\cup|}{d_I}\right)^{d_I} \le \left(\frac{erd_B}{d_I}\right)^{d_I},$$

Where e is the base of the natural logarithm. It follows that

$$d_B \le d_I(\log r - \log d_I + \log e) + d_I \log d_B.$$

If $d_I \log d_B \le \frac{1}{2} d_B$,

$$d_B \le 2d_I(\log r - \log d_I + \log e).$$

Otherwise, $d_I \log d_B \ge \frac{1}{2} d_B$, hence,

$$\frac{d_B}{2 \log d_B} \le d_I. \tag{4}$$

If $d_B > 16$ then $\log d_B \le \sqrt{d_B}$, therefore from Eq. (4) it follows that $\frac{1}{2}\sqrt{d_B} \le d_I$, that is $d_B \le 4{d_I}^2$. Combining all the cases, the bound on d_B follows. ■

To complement Theorem 5, we show in the following theorem that for the class of separating hyperplanes the VC-dimension of homogeneous-dependent OR-MIL is at least logarithmic in r, hence the logarithmic dependence in r cannot be removed in the general case.

Theorem 6 *Let ϕ_r^{OR} be the homogeneous bag-labeling operator for OR-MIL, and let $\mathcal{H}$ be the class of separating hyperplanes in $\mathbb{R}^k$ for $k \ge 2$. Denote by d_B the VC-dimension of $\phi_r^{\mathrm{OR}}(\mathcal{H})$. Then*

$$d_B \ge \lfloor \log r \rfloor + 1.$$

Proof: Denote $D \triangleq \lfloor \log r \rfloor + 1$. We show that there exists a set $S = (\mathbf{x}_1, \ldots, \mathbf{x}_D) \subseteq (\mathbb{R}^2)^r$ of D bags with r instances that can be shattered by the class of separating hyperplanes in two dimensions. It follows that the same is true for $\mathbb{R}^k$ with $k \ge 2$. The following construction is illustrated in Figure 1.

For $k \in [Dr]$ define instances in $\mathbb{R}^2$ as follows:

$$a_k = (\cos(\frac{2\pi k}{Dr}), \sin(\frac{2\pi k}{Dr})),$$

so that the instances are equidistant on the unit circle. Let $N = (n_1, \ldots, n_{Dr})$ be a sequence of indexes from $[D]$ which is a concatenation of all the subsets of $[D]$ in some arbitrary order. Each index in $[D]$ appears in half of the subsets of $[D]$, therefore for all $i \in [D]$, there are $2^{D-1} \le r$ indexes k

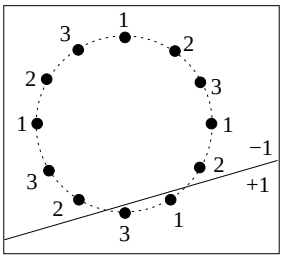

Figure 1: An illustration of the construction in Theorem 6. In this example $r = 4$ and $D = \log 4 + 1 = 3$. Each dot corresponds to an instance. The numbers next to the instances denote the bag to which an instance belongs, and match the sequence N. In this illustration bags 1 and 3 are labeled as positive by the OR rule.

such that $n_k = i$. For $i \in [D]$, let $\mathbf{x}_i$ be a vector of the instances a_k such that $n_k = i$, in some arbitrary order. If $2^{D-1} < r$, duplicate one of the instances in each bag to achieve a bag of size exactly r.

Let $(y_1, \ldots, y_D)$ be an arbitrary binary labeling of D bags. We show that there exists a separating hyperplane $\mathbf{w} \cdot x - d = 0$ that induces this labeling on S: Let $T = \{i \mid y_i = +1\}$. By the definition of N, there exists a sub-sequence $n_{j_1}, \ldots, n_{j_2}$ in N such that $\{n_k \mid j_1 \leq k \leq j_2\} = T$. Let $\mathbf{w} = (\cos(\frac{\pi(j_1+j_2)}{Dr}), \sin(\frac{\pi(j_1+j_2)}{Dr}))$ and let $d = \cos(\frac{\pi(j_2-j_1)}{Dr})$. We have that $\mathbf{w} \cdot a_k - d \geq 0$ if and only if $j_1 \leq k \leq j_2$. Therefore a bag $\mathbf{x}_i$ is assigned $+1$ by the OR function exactly if there exists a $k \in \{j_1, \ldots, j_2\}$ such that $n_k = i$, that is exactly if $i \in T$.

We have shown that S can be shattered by the class of separating hyperplanes, therefore $d_B \geq |S| = \lfloor \log r \rfloor + 1$. ■

Since the bound on sample complexity is proportional to the VC-dimension of the problem [VC71], it follows from Theorem 5 that the sample size required for learning from bags is larger than the sample size required for learning from instances only by a logarithmic factor of r. In this section we have shown that homogeneous-dependent MIL is strictly easier than heterogeneous-dependent MIL in terms of sample complexity. In the following section we focus on the computational aspect of OR-MIL, and show that in homogeneous-dependent OR-MIL we can classify bags using an algorithm that classifies instances.

4 An Algorithm for Homogeneous-Dependent OR-MIL

In this section we present `MILearn`, a learning algorithm for the OR-MIL problem, and show that in homogeneous-dependent OR-MIL, given an algorithm $\mathcal{A}$ that minimizes the training error on a sample of instances, one can use `MILearn` to classify bags efficiently, with guarantees on the result. If $\mathcal{A}$ minimizes one-sided error, then it is possible to learn from separable samples and from samples with one-sided error. If $\mathcal{A}$ minimizes two-sided error, then it is also possible to learn from samples with small two-sided error. Comparing this to heterogeneous-dependent OR-MIL, note that the hardness result on learning APRs in the latter setting [ALS98] is achieved using a hypothesis class with only a finite number of APRs. Therefore, this problem is hard even when $\mathcal{A}$ exists.

Before presenting the algorithm, we define some notation. A labeled and weighted sample of instances is a set $S \subseteq \mathbb{R}^+ \times X \times \{-1, +1\}$ of triplets, where in a triplet (w, x, y) 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 $S \subseteq \mathbb{R}^+ \times X^r \times \{-1, +1\}$ of triplets $(w, \mathbf{x}, y)$ defined in a similar fashion.

In the following we consider real-valued hypotheses, that is $\mathcal{H} \subseteq [-1, +1]^X$. The OR-MIL homogeneous bag-labeling operator is accordingly defined as $\phi_r^{\max}$ instead of ϕ_r^{OR}.

For an instance hypothesis $h : X \to [-1, +1]$ and a weighted and labeled instance sample $S = \{(w_i, x_i, y_i)\}_{i \in [m]}$, the edge of h on S, denoted by $\Gamma(h, S)$, is defined as follows.

$$\Gamma(h, S) \triangleq \sum_{i \in |S|} w_i y_i h(x_i) / \sum_{i \in |S|} w_i.$$

If h is a bag hypothesis and S is a bag sample, $\Gamma(h, S)$ is defined identically except that the sum is over bags $\mathbf{x}_i$ instead of instances x_i. Note that for binary hypotheses, the edge of h on S equals $1 - 2\epsilon$ where ϵ is the error of h on S.

`MILearn` accepts as input a bag sample, S_B, and an algorithm, $\mathcal{A}$. $\mathcal{A}$ receives a labeled and weighted instance sample and returns an instance hypothesis $\mathcal{A}(S) \in \mathcal{H}$. In `MILearn`, h_{pos} denotes the constant positive hypothesis: $\forall x \in X, \quad h_{\mathrm{pos}}(x) = +1$. It is assumed that $h_{\mathrm{pos}} \in \mathcal{H}$. The output of the algorithm is a bag hypothesis in $\phi_r^{\max}(\mathcal{H})$ that classifies S_B. The edge of the returned hypothesis depends on the best achievable edge for S_B, as we presently show.

`MILearn`, listed as Algorithm 1 below, is a very short algorithm. It constructs a sample of instances S_I from the instances that make up bags in S_B, labeling each instances in S_I with the same label of the bag it came from. The weight of an instance with a positive label is times $\frac{1}{r}$ of the weight of the bag it came from, while the weight of an instance with a negative label is 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 $\phi_r^{\max}(h_{\mathrm{pos}})$, whichever provides the better edge on S_B.

This simple algorithm provides guarantees for the edge of the resulting hypothesis, as we show in the following theorem. The theorem is composed of two parts – the first part refers to the best achievable one-sided edge, and the second part relates to the overall best achievable edge.

Algorithm 1: `MILearn`

Assumptions: $h_{\text{pos}} \in \mathcal{H}$.
Input:

- $S_B \triangleq \{(w_i, \mathbf{x}_i, y_i)\}_{i\in[m]}$ – a labeled and weighted sample of bags;
- $\mathcal{A}$ – an algorithm that receives an instance sample and returns a hypothesis in $\mathcal{H}$.

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

1 $\alpha(+1) \leftarrow \frac{1}{r}$, $\alpha(-1) \leftarrow 1$
2 Create an instance sample S_I with rm instances as follows:

$$S_I \leftarrow \{(\alpha(y_i)w_i, x_i^j, y_i)\}_{i\in[m],j\in[r]}.$$

3 $h_I \leftarrow \mathcal{A}(S_I)$
4 **if** $\Gamma(\phi_r^{\max}(h_I), S_B) \geq \Gamma(\phi_r^{\max}(h_{\text{pos}}), S_B)$ **then**
5 $\quad h_M \leftarrow \phi_r^{\max}(h_I)$
6 **else**
7 $\quad h_M \leftarrow \phi_r^{\max}(h_{\text{pos}})$.

Before stating the theorem, some auxiliary notation is required. We first define the set of hypotheses that classify a sample S with one-sided error. Since in OR-MIL positive and negative labels are not interchangeable, we specifically require that such hypotheses err only on *positive* examples in S.

Definition 7 *The set of hypotheses with one-sided error on S is denoted by $\Omega(S)$. If $S = \{(w_i, x_i, y_i)\}_{i\in[m]}$ is an instance sample then $\Omega(S)$ is defined as follows:*

$$\Omega(S) \triangleq \{h \in [-1,+1]^X \mid \forall i \in [m], h(x_i) \neq y_i \Rightarrow y_i = +1\}.$$

If $S = \{(w_i, \mathbf{x}_i, y_i)\}_{i\in[m]}$ is a bag sample,

$$\Omega(S) \triangleq \{h \in [-1,+1]^X \mid \forall i \in [m], \phi_r^{\max}(h)(\mathbf{x}_i) \neq y_i \Rightarrow y_i = +1\}.$$

Two short-hand notations are used for the best achievable edge for S_B, the input sample to `MILearn`:

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

That is, γ^* is the best achievable edge on S_B with hypotheses in $\phi_r^{\max}(\mathcal{H})$, and γ_+^* is the best achievable edge on S_B with hypotheses in $\phi_r^{\max}(\mathcal{H})$ that err only on positive bags.

Theorem 8 *Let $\mathcal{H} \subseteq [-1,+1]^X$ 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

$$\Gamma(\mathcal{A}(S), S) \geq \max_{h\in\mathcal{H}\cap\Omega(S)} \Gamma(h, S), \quad (5)$$

that is, $\mathcal{A}$ minimizes one-sided error on S, then

$$\gamma \geq \frac{\gamma_+^*}{2r-1}. \quad (6)$$

(b) If the following conditions hold:
i. for any instance sample S

$$\Gamma(\mathcal{A}(S), S) \geq \max_{h\in\mathcal{H}} \Gamma(h, S), \quad (7)$$

ii. $\gamma^* \geq 1 - \frac{1}{r^2}$,
then

$$\gamma \geq \frac{r^2(\gamma^* - 1) + 1}{2r-1} \geq 0. \quad (8)$$

Proof:[of Theorem 8(a)] Denote the total weight of examples in a sample S by $W(S)$. For $\{w_i\}$ the weights of bags in S_B, let

$$W_+ \triangleq \sum_{i:y_i=+1} w_i, \text{ and } W_- \triangleq \sum_{i:y_i=-1} w_i.$$

We assume w.l.o.g. that $W(S_B) = W_+ + W_- = 1$. In the proof we refer to S_I and h_I as defined in steps 2 and 3 of `MILearn`.

The proof employs the following three lemmas, whose 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 *Define*

$$h_+^* \triangleq \operatorname*{argmax}_{h\in\mathcal{H}\cap\Omega(S_B)} \Gamma(\phi_r^{\max}(h), S_B).$$

If Eq. (5) holds, then $\Gamma(h_I, S_I) \geq \Gamma(h_+^*, S_I)$.

Lemma 11 *For any instance hypothesis h,*

$$\begin{aligned} &W(S_I)\Gamma(h, S_I) \geq \\ &\geq (\tfrac{1}{r} - 1)W_+ + \tfrac{1}{r}\Gamma(\phi_r^{\max}(h), S_B) \\ &\quad + (r - \tfrac{1}{r})(W_- - \sum_{y_i=-1} w_i \|h(\mathbf{x}_i) + 1\|_\infty). \end{aligned}$$

From these three lemmas we have that if Eq. (5) holds then

$$\begin{aligned} &\Gamma(\phi_r^{\max}(h_I), S_B) \\ &\geq W(S_I)\Gamma(h_I, S_I) + (1-r)W_- \\ &\geq W(S_I)\Gamma(h_+^*, S_I) + (1-r)W_- \\ &\geq (\tfrac{1}{r} - 1)W_+ + \tfrac{1}{r}\gamma_+^* + (1-r)W_- \\ &\quad + (r - \tfrac{1}{r})(W_- - \sum_{y_i=-1} w_i \|h_+^*(\mathbf{x}_i) + 1\|_\infty) \\ &= (\tfrac{1}{r} - 1)W_+ + (1 - \tfrac{1}{r})W_- + \tfrac{1}{r}\gamma_+^* \\ &\quad - (r - \tfrac{1}{r}) \sum_{y_i=-1} w_i \|h_+^*(\mathbf{x}_i) + 1\|_\infty \\ &= (1 - \tfrac{1}{r})(1 - 2W_+) + \tfrac{1}{r}\gamma_+^* \\ &\quad - (r - \tfrac{1}{r}) \sum_{y_i=-1} w_i \|h_+^*(\mathbf{x}_i) + 1\|_\infty, \end{aligned} \quad (9)$$

where the last equality follows from the assumption that $W_+ + W_- = 1$. Now, we have that $h^*_+ \in \Omega(S_B)$, therefore for all $i \in [m]$ such that $y_i = -1$, $h^*_+(\mathbf{x}_i) = -1$. It follows that $h^*_+(x^j_i) = -1$ for all $i \in [m], j \in [r]$. Therefore $\sum_{y_i=-1} w_i \|h^*_+(\mathbf{x}_i) + 1\|_\infty = 0$. Eq. (9) is thus reduced to

$$\Gamma(\phi^{\max}_r(h_I), S_B) \geq (1 - \tfrac{1}{r})(1 - 2W_+) + \tfrac{1}{r}\gamma^*_+.$$

From step 4 of the algorithm we have

$$\gamma \geq \max\{\Gamma(\phi^{\max}_r(h_I), S_B), \Gamma(\phi^{\max}_r(h_{\text{pos}}), S_B)\}.$$

Therefore

$$\gamma \geq \max\left\{(1 - \tfrac{1}{r})(1 - 2W_+) + \tfrac{1}{r}\gamma^*_+,\ 2W_+ - 1\right\}.$$

Since the first option in the maximization is decreasing in W_+ and the second option is increasing in W_+, we have that $\gamma \geq 2W^*_+ - 1$, where W^*_+ satisfies

$$(1 - \tfrac{1}{r})(1 - 2W^*_+) + \tfrac{1}{r}\gamma^*_+ = 2W^*_+ - 1.$$

It is easy to verify that $W^*_+ = \frac{1}{2} + \frac{\gamma^*_+}{4r-2}$, therefore

$$\gamma \geq 2W^*_+ - 1 = \frac{\gamma^*_+}{2r-1}.$$

■

The proof of Theorem 8(b) follows similar lines and is provided in Appendix A. Theorem 8 guarantees that under certain conditions `MILearn` achieves an approximation to the optimal edge of a hypothesis in $\phi^{\max}_r(\mathcal{H})$ on the input sample. Given this guarantee, one can also guarantee an approximation of the L_1 margin of a linear combination of hypotheses from $\phi^{\max}_r(\mathcal{H})$, by using a Boosting scheme such as AdaBoost or one of its variants ([SFBL98], and see [SSS08] for an elegant analysis). This is because if there exists an L_1 margin of γ^* over $\phi^{\max}_r(\mathcal{H})$ then `MILearn` is a *weak learner* which provides hypotheses with an edge of at least γ as stated in Eq. (8), for any re-weighing of the input bag sample. For the separable case, using `MILearn` as a weak learner in a Boosting algorithm guarantees separability with a margin of $\frac{1}{2r-1}$, by a linear combination of hypotheses from $\phi^{\max}_r(\mathcal{H})$. We mention that [GFKS02] presents a method for conserving linear separability by defining a kernel adaptation to OR-MIL. However, the resulting MIL margin bound goes to zero as the sample size grows. Further research may allow combining the two approaches to achieve a kernel with guaranteed margin bounds.

The generalization bound for the use of AdaBoost with `MILearn` and binary hypotheses depends on the achieved L_1 margin and the VC-dimension of $\phi^{\max}_r(\mathcal{H})$ [SFBL98]. The VC-dimension was bounded in Theorem 5 above, thus bounding generalization error. For real-valued hypotheses, however, a bound on the pseudo-dimension of $\phi^{\max}_r(\mathcal{H})$ is required [SS99]. In the following section the case of real-valued hypotheses is analyzed for generalized MIL and the necessary bound is provided.

Before we proceed, it is instructive to compare the computational result we achieved for the separable case with `MILearn` to Theorem 1 [BK98], which also addresses the separable case. From the above discussion on Boosting we have:

Corollary 12 *If it is possible to minimize the training error on a sample with one-sided error by a hypothesis class $\mathcal{H}$ in polynomial time, then it is possible to PAC-learn a bag sample in OR-MIL in polynomial time, where the examples are drawn from an arbitrary distribution over bags.*

Theorem 1 and Cor. 12 are similar in structure: Both state that if the single-instance problem is solvable with one-sided error, then the MIL problem is solvable if it is separable. Theorem 1 applies only to the case of statistically independent instances, while Cor. 12 applies to bags drawn from an arbitrary distribution. It should be noted though, that the conditions required in Cor. 12 are stronger, requiring training error minimization with handling of arbitrary one-sided error, while Theorem 1 requires PAC-learnability with handling of one-sided random noise.

5 MIL with Real-Valued Functions

We now extend the discussion to hypotheses and bag classification rules that range over real values. We show that if the bag classification rule is an extension of a monotone Boolean function, then here too the sample complexity of MIL depends logarithmically on r. For margin learning our results hold for the larger class of Lipschitz functions.

Monotone Boolean functions (also called *positive Boolean functions*) map Boolean vectors in $\{-1,+1\}^n$ 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 13 *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 denoted by $\mathbf{x} \in [-1,+1]^r$:*

$$\begin{aligned}
&(1)\ \forall j \in [n], \quad \mathbf{x} \mapsto x^j \in \mathcal{M}_r;\\
&(2)\ \forall k \in \mathbb{N}^+, \quad f^1, \ldots, f^k \in \mathcal{M}_r \Longrightarrow \\
&\qquad\qquad \mathbf{x} \mapsto \max(\mathbf{f}(\mathbf{x})) \in \mathcal{M}_r; \qquad (10)\\
&(3)\ \forall k \in \mathbb{N}^+, \quad f^1, \ldots, f^k \in \mathcal{M}_r \Longrightarrow \\
&\qquad\qquad \mathbf{x} \mapsto \min(\mathbf{f}(\mathbf{x})) \in \mathcal{M}_r,
\end{aligned}$$

where $\mathbf{f} \triangleq (f^1, \ldots, f^k)$.

5.1 Thresholded Functions

In the following we bound the pseudo-dimension (see e.g. [AB99] for definitions) of the generalized MIL problem with any extension of a monotone Boolean function, showing that here too as in Theorem 5, the sample complexity of MIL is larger than that of the single-instance problem by at most a logarithmic factor of r. This also shows that using Boosting with `MILearn` generalizes when $\mathcal{H}$ ranges over real-valued hypotheses.

Theorem 14 *Let $\mathcal{H} \subseteq [-1,+1]^X$ be a set of instance hypotheses with pseudo-dimension d_I . Let $f : [-1,+1]^r \to [-1,+1]$ be an extension of a monotone Boolean function, and let d_B be the pseudo-dimension of $\phi_r^f(\mathcal{H})$. Then*

$$d_B \leq \max\{2d_I(\log r - \log d_I + 1), 4{d_I}^2, 16\}.$$

Proof: We use the equivalence between the pseudo-dimension of a class of real-valued functions and the VC-dimension of the class of binary functions generated by thresholding the real-valued functions, following [AB99]. For a function h from some domain into $[-1,+1]$ and a scalar $y \in \mathbb{R}$, let h_y be a function from the same domain into $\{-1,+1\}$, defined by $h_y(x) = \mathrm{sign}(h(x)-y)$, where $\mathrm{sign}(x) = +1$ if $x \geq 0$, and $\mathrm{sign}(x) = -1$ otherwise. For a set of functions H, define the set $B_H \triangleq \{h_y \mid h \in H, y \in \mathbb{R}\}$. The pseudo-dimension of H is equal to the VC-dimension of B_H.

Using Def. 13, it is easy to verify that for f which is an extension of a monotone Boolean function, the following holds, where $\mathbf{1} = (1,\ldots,1)$:

$$\begin{aligned}&\mathrm{sign}(f(\mathbf{x}) - y) \equiv \mathrm{sign}(f(\mathbf{x} - y\mathbf{1}))\\ &\equiv f(\mathrm{sign}(\mathbf{x} - y\mathbf{1})),\end{aligned}$$

Now let us examine the thresholded function $\phi_r^f(h)_y$ for $h \in \mathcal{H}$ and $y \in \mathbb{R}$. For all $\mathbf{x} \in X^r$,

$$\begin{aligned}&\phi_r^f(h)_y(\mathbf{x}) = \mathrm{sign}(\phi_r^f(h)(\mathbf{x}) - y)\\ &= \mathrm{sign}(f(h(\mathbf{x})) - y) = f(\mathrm{sign}(h(\mathbf{x}) - y\mathbf{1}))\\ &= f(h_y(\mathbf{x})) = \phi_r^f(h_y)(\mathbf{x}).\end{aligned}$$

Therefore, $B_{\phi_r^f(\mathcal{H})} = \phi_r^f(B_{\mathcal{H}})$. We have that d_I is the VC-dimension of $B_{\mathcal{H}}$ and d_B is the VC dimension of $B_{\phi_r^f(\mathcal{H})} = \phi_r^f(B_{\mathcal{H}})$. Note that $B_{\mathcal{H}}$ is a set of hypotheses into $\{-1,+1\}$, and that f when restricted to $\{-1,+1\}^r$ is also binary, therefore we may apply Theorem 5, substituting $\mathcal{H}$ with $B_{\mathcal{H}}$. 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. MI-SVM [ATH02] is a practical algorithm for learning OR-MIL with a margin. This algorithm attempts to optimize an adaptation of the soft-margin SVM objective, in which the margin of a bag is the maximal margin achieved by any of its instances. This amounts to replacing the hypothesis class $\mathcal{H}$ of separating hyperplanes by $\phi_r^{\max}(\mathcal{H})$. Since $\max$ is the extension of OR, this objective function is natural in our formulation. It has not been shown, however, that minimizing the objective function of MI-SVM and analogous margin formulations for MIL allows learning. This is provided by Theorem 15 below, which bounds the γ-Fat shattering dimension (see e.g. [AB99]) of MIL. This theorem applies to any bag classification rule which is a Lipschitz function, where the Lipschitz condition is formally defined as follows: A function $f : X^r \to \mathbb{R}$ is c-Lipschitz with respect to the infinity norm if

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

The bound in Theorem 15 shows that these MIL problems are indeed learnable with a small penalty on the sample size, if the single-instance problem is learnable.

The following theorem assumes that the real-valued hypotheses are bounded. Also assume w.l.o.g. that the hypotheses and the bag classification rule are non-negative. For a function class F, let $\mathrm{Fat}_F(\gamma)$ be its γ-Fat shattering dimension.

Theorem 15 *Let $B, c > 0$. Let $\mathcal{H} \subseteq [0,B]^X$ be a hypothesis class and let $f : [0,B]^r \to [0,cB]$ be c-lipschitz with respect to the infinity norm. For $\gamma > 0$, denote $\mathcal{F}_1(\gamma) \triangleq \mathrm{Fat}_{\mathcal{H}}(\gamma)$, and $\mathcal{F}_r(\gamma) \triangleq \mathrm{Fat}_{\phi_r^f(\mathcal{H})}(\gamma)$. Then*

$$\mathcal{F}_r(\gamma) \leq 6\mathcal{F}_1(\frac{\gamma}{4c})\log^2(8\frac{B^2c^4}{\gamma^2}r\mathcal{F}_r(16\gamma)). \quad (11)$$

The bound in Eq. (11) 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 $\mathcal{F}_1(\gamma/4c)$ and setting $\beta = 6\mathcal{F}_1(\gamma/4c)$ and $\eta = 8B^2c^4/\gamma^2$, we have

$$\sqrt{\mathcal{F}_r} - \sqrt{\beta}\log\mathcal{F}_r \leq \sqrt{\beta}\log(\eta r). \quad (12)$$

Therefore the bound on $\mathcal{F}_r$ is asymptotically polylogarithmic in r. This bound applies to extensions of monotone Boolean functions as well, since they are Lipschitz functions, as the following lemma shows.

Lemma 16 *Extensions of monotone Boolean functions are 1-Lipschitz with respect to the infinity norm.*

This lemma is proven inductively using Def. 13. The full proof is provided in Appendix A.

To prove Theorem 15, we first bound the covering number of MIL by the covering number of the single-instance problem. For $\mathcal{H} \subseteq [0,B]^X$, $\gamma > 0$ and $S \subseteq X$, the set of γ-covers of S by $\mathcal{H}$ is

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

The γ-covering number of a hypothesis class $\mathcal{H} \subseteq [0,B]^X$ and a number $m > 0$ is defined by

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

The following lemma provides the necessary bound.

Lemma 17 *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 $\mathcal{H} \subseteq [0,B]^X$,*

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

Proof: Let $S = \{\mathbf{x}_i\}_{i \in [m]} \subseteq X^r$ 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 $\max_{i \in [m]} \|h(\mathbf{x}_i) - \hat{h}(\mathbf{x}_i)\|_\infty \leq \gamma$. From the Lipschitz condition on f we have

$$\begin{aligned} &|\phi_r^f(h)(\mathbf{x}_i) - \phi_r^f(\hat{h})(\mathbf{x}_i)| \equiv \\ &\equiv \|f(h(\mathbf{x}_i)) - f(\hat{h}(\mathbf{x}_i))\|_\infty \\ &\leq c\|h(\mathbf{x}_i) - \hat{h}(\mathbf{x}_i)\|_\infty \leq c\gamma. \end{aligned}$$

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

$$\phi_r^f(\mathrm{cov}_\gamma(\mathcal{H}, S^\cup)) \subseteq \mathrm{cov}_{c\gamma}(\phi_r^f(\mathcal{H}), S).$$

Therefore,

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

■

Lastly, in the proof of Theorem 15 we use the following two theorems.

Theorem 18 ([BKP97]) *Let F be a set of real functions and let $\gamma > 0$. For $m \geq \mathrm{Fat}_F(16\gamma)$,*

$$e^{\mathrm{Fat}_F(16\gamma)/8} \leq \mathcal{N}_\infty(\gamma, F, m). \quad (14)$$

Theorem 19 (Theorem 12.8 in [AB99]) *Let F be a set of real functions from a domain X to the bounded interval $[0,B]$. Let $\gamma > 0$. Let $d = \mathrm{Fat}_F(\frac{\gamma}{4})$. For all $m \geq d$,*

$$\mathcal{N}_\infty(\gamma, F, m) < 2\left(\frac{4B^2m}{\gamma^2}\right)^{d \log \frac{4eBm}{d\gamma}}. \quad (15)$$

We are now ready to prove the fat shattering bound.

Proof:[of Theorem 15] From Theorem 18 and Lemma 17 it follows that for $m \geq \mathcal{F}_r(16\gamma)$,

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

This expression can be bounded from above using Theorem 19: Rearranging Eq. (15) we have that if $m \geq d = \mathrm{Fat}_F(\frac{\gamma}{4}) \geq 1$ and F is into $[0,B]$ then, for $\gamma \leq B/e$,

$$\begin{aligned} &\log \mathcal{N}_\infty(\gamma, \mathcal{H}, m) < \\ &< d \log(\frac{4eBm}{d\gamma}) \log\left(\frac{4B^2m}{\gamma^2}\right) + 1 \\ &\leq d \log(\frac{4eBm}{\gamma}) \log\left(\frac{4B^2m}{\gamma^2}\right) + 1 \\ &\leq d \log^2(\frac{4B^2m}{\gamma^2}) \\ &= \mathrm{Fat}_F(\frac{\gamma}{4}) \log^2(\frac{4B^2m}{\gamma^2}). \end{aligned}$$

Combining this with Eq. (16) and substituting B with cB it follows that if $m \geq \mathcal{F}_r(16\gamma)$ and $rm \geq \mathcal{F}_1(\frac{\gamma}{4c}) \geq 1$, then

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

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

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

Substituting 16γ with γ, we have that the bound in Eq. (11) holds for $\gamma/16 \leq B/e$, which always holds since $\gamma \leq B$. ■

6 Discussion

In this work we have analyzed Multiple Instance Learning in a new theoretical setting. The assumptions in this setting are closer to the ones made in practice, and unlike previously investigated settings, do not reduce MIL to a very hard problem nor to a trivial one. We have shown that the dependence of the sample complexity of MIL on the number of instances in a bag is no more than logarithmic. This result extends to any Boolean function, on top of the Boolean OR used in classical MIL. It would be of interest to compare this tradeoff to similar phenomena in other settings with partial information on labels, such as Active Learning and Semi Supervised Learning. We would

also like to investigate whether under certain conditions, such as a high cost of labels, it may be preferable to use bag learning instead of instance learning.

For the OR-MIL problem, we have provided a learning algorithm that classifies bags given an algorithm for minimizing training error over instances. This is the first OR-MIL algorithm with proven generalization performance that does not assume statistical independence of instances in a bag. Further research is required to generalize this reduction to other settings and to compare this strategy to other methods for generating weak learners. Lastly, we have generalized MIL further to handle real-valued hypotheses and bag classification rules, and have shown that here too the sample complexity is poly-logarithmic by the number of instances in a bag.

7 Acknowledgements

We thank Shai Shalev-Shwartz and Nati Srebro for helpful discussions.

References

[AB99] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[AH03] S. Andrews and T. Hofmann. Multiple-instance learning via disjunctive programming boosting. In *NIPS 16*, 2003.

[ALS98] P. Auer, P. M. Long, and A. Srinivasan. Approximating hyper-rectangles: learning and pseudorandom sets. *J. Comput. Syst. Sci.*, 57(3):376–388, 1998.

[And07] S. Andrews. *Learning from ambiguous examples*. PhD thesis, Brown University, May 2007.

[ATH02] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS 15*, pages 561–568, 2002.

[BK98] A. Blum and A. Kalai. A note on learning from multiple-instance examples. *Mach. Learn.*, 30(1):23–29, 1998.

[BKP97] P. L. Bartlett, S. R. Kulkarni, and S. E. Posner. Covering numbers for real-valued function classes. *IEEE Transactions on Information Theory*, 43(5):1721–1724, 1997.

[DLLP97] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, 89(1-2):31–71, 1997.

[GFKS02] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola. Multi-instance kernels. In *ICML '02*, pages 179–186, 2002.

[LT98] 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.

[MLP98] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *NIPS 10*, pages 570–576, 1998.

[MR98] O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In *ICML '98*, pages 341–349, 1998.

[PV86] L. Pitt and L. G. Valiant. Computational limitations on learning from examples. Technical report, Harvard University Aiken Computation Laboratory, July 1986.

[Rae98] L. De Raedt. Attribute-value learning versus inductive logic programming: The missing links (extended abstract). In *ILP '98*, pages 1–8, London, UK, 1998. Springer-Verlag.

[Sau72] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory Series A*, 13:145–147, 1972.

[SFBL98] 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.

[SS99] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):1–40, 1999.

[SSS08] S. Shalev-Shwartz and Y. Singer. On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms. In *COLT '08*, pages 311–322, 2008.

[VC71] 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.

[WFP03] N. Weidmann, E. Frank, and B. Pfahringer. A two-level learning method for generalized multi-instance problems, 2003.

[ZG01] Q. Zhang and S.A. Goldman. EM-DD: An improved multiple-instance learning technique. In *NIPS 14*, 2001.

[ZJL05] Z. Zhou, K. Jiang, and M. Li. Multi-instance learning based web mining. *Applied Intelligence*, 22(2):135–147, 2005.

A Technical Proofs

Lemma 20 *For any $f : \{-1,+1\}^X \to \{-1,+1\}$, hypothesis class $\mathcal{H}$, and a set of bags $S = \{\mathbf{x}_i\}_{i\in[d_I]} \subseteq X^r$, let $S^\cup = \{x_i^j\}_{i\in[m],j\in[r]} \subseteq X$. Then,*

$$\left|\phi_r^f(\mathcal{H})_{|_S}\right| \leq \left|\mathcal{H}_{|_{S^\cup}}\right|.$$

Proof: Let $h_B^1, h_B^2 \in \phi_r^f(\mathcal{H})$ be bag hypotheses. There exist instance hypotheses $h_I^1, h_I^2 \in \mathcal{H}$ such that $\phi_r^f(h_I^i) = h_B^i$ for $i = 1, 2$. Assume that $h_{B|_S}^1 \neq h_{B|_S}^2$. We show that $h_{I|_{S^\cup}}^1 \neq h_{I|_{S^\cup}}^2$, thus proving the lemma.

From the assumption it follows that $\phi_r^f(h_I^1)_{|_S} \neq \phi_r^f(h_I^2)_{|_S}$. There exists at least one bag $\mathbf{x}_i \in S$ such that $\phi_r^f(h_I^2)(\mathbf{x}_i) \neq \phi_r^f(h_I^2)(\mathbf{x}_i)$. Therefore

$$f(h_I^1(\mathbf{x}_i)) \neq f(h_I^2(\mathbf{x}_i)).$$

Hence there exists a $j \in [r]$ such that

$$h_I^1(x_i^j) \neq h_I^2(x_i^j).$$

By the definition of $S^\cup$, $x_i^j \in S^\cup$. Therefore

$$h_{I|_{S_I}}^1 \neq h_{I|_{S_I}}^2.$$

■

Proof:[of Lemma 9] Since we assume w.l.o.g. that $W(S_B) = 1$, we have

$$\begin{aligned}\Gamma(\phi_r^{\max}(h), S_B) &= \sum_{i\in[m]} w_i y_i \max_{j\in[r]}\{h(x_i^j)\} \\ &= \sum_{i\in[m]} w_i y_i (\|h(\mathbf{x}_i) + 1\|_\infty - 1).\end{aligned}$$

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

$$\begin{aligned}&W(S_I)\Gamma(h, S_I) = \\ &= \sum_{i\in[m],j\in[r]} \alpha(y_i) w_i y_i h(x_i^j) \\ &= \sum_{i\in[m]} \alpha(y_i) w_i y_i \sum_{j\in[r]} h(x_i^j) \\ &= \sum_{i\in[m]} \alpha(y_i) w_i y_i (\|h(\mathbf{x}_i) + 1\|_1 - r).\end{aligned}$$

From

$$\|h(\mathbf{x}_i) + 1\|_\infty \leq \|h(\mathbf{x}_i) + 1\|_1 \leq r\|h(\mathbf{x}_i) + 1\|_\infty$$

it follows that

$$y_i\alpha(y_i)\|h(\mathbf{x}_i) + 1\|_1 \leq y_i\|h(\mathbf{x}_i) + 1\|_\infty. \quad (17)$$

Therefore,

$$\begin{aligned}&W(S_I)\Gamma(h, S_I) \leq \\ &\leq \sum_{i\in[m]} w_i y_i (\|h(\mathbf{x}_i) + 1\|_\infty - \alpha(y_i) r) \\ &\leq \sum_{i\in[m]} w_i y_i (\|h(\mathbf{x}_i) + 1\|_\infty - 1) \\ &\quad + (r-1) \sum_{y_i=-1} w_i \\ &= \Gamma(\phi_r^{\max}(h), S_B) + (r-1)W_-.\end{aligned}$$

And the proof is completed. ■

Proof:[of Lemma 10] By Eq. (5) we have

$$\Gamma(h_I, S_I) = \Gamma(\mathcal{A}(S_I), S_I) \geq \max_{h\in\mathcal{H}\cap\Omega(S_I)} \Gamma(h, S_I).$$

Thus to prove that $\Gamma(h_I, S_I) \geq \Gamma(h_+^*, S_I)$ it suffices to show that $h_+^* \in \mathcal{H} \cap \Omega(S_I)$. By definition $h_+^* \in \mathcal{H}\cap\Omega(S_B)$, therefore it suffices to show that

$$\Omega(S_B) \subseteq \Omega(S_I),$$

that is, any hypothesis which errs only on positive instances on S_B, also errs only on positive instances on S_I. Let $h \in \Omega(S_B)$. From the definition of Ω it follows that

$$\forall i, \quad \phi_r^{\max}(h)(\mathbf{x}_i) \neq y_i \quad \Longrightarrow \quad y_i = +1.$$

Equivalently,

$$\forall i, \quad y_i = -1 \quad \Longrightarrow \quad \phi_r^{\max}(h)(\mathbf{x}_i) = -1.$$

Therefore,

$$\forall i, \quad y_i = -1 \quad \Longrightarrow \quad \forall j \in [r], h(x_i^j) = -1.$$

It follows that

$$\forall i \in [m], j \in [r], y_i = -1 \implies h(x_i^j) = -1.$$

Denoting $S_I = \{(\hat{w}_k, \hat{x}_k, \hat{y}_k)\}_{k\in[\hat{m}]}$, we have that $\hat{m} = rm$, $\hat{x}_k = x_j^i$ and $\hat{y}_k = y_i$ for some $i \in [m], j \in [r]$. Therefore

$$\forall k \in [\hat{m}], \hat{y}_k = -1 \implies h(\hat{x}_k) = -1.$$

Thus $h \in \Omega(S_I)$. Hence $\Omega(S_B) \subseteq \Omega(S_I)$, and the proof is concluded. ■

Proof:[of Lemma 11] The following chain of equalities provides the required result:

$$\begin{aligned}&W(S_I)\Gamma(h, S_I) = \\ &= \sum_{i\in[m]} \alpha(y_i) w_i y_i (\|h(\mathbf{x}_i) + 1\|_1 - r) \\ &= \sum_{y_i=+1} \tfrac{1}{r} w_i (\|h(\mathbf{x}_i) + 1\|_1 - r) \\ &\quad + \sum_{y_i=-1} w_i (r - \|h(\mathbf{x}_i) + 1\|_1) \\ &\geq \sum_{y_i=+1} \tfrac{1}{r} w_i (\|h(\mathbf{x}_i) + 1\|_\infty - r) \\ &\quad + \sum_{y_i=-1} w_i (r - r\|h(\mathbf{x}_i) + 1\|_\infty) \\ &= \sum_{y_i=+1} (\tfrac{1}{r} - 1) w_i \\ &\quad + \sum_{y_i=+1} \tfrac{1}{r} w_i (\|h(\mathbf{x}_i) + 1\|_\infty - 1) \\ &\quad + \sum_{y_i=-1} r w_i (1 - \|h(\mathbf{x}_i) + 1\|_\infty) \\ &= (\tfrac{1}{r} - 1) W_+ + \tfrac{1}{r}\Gamma(\phi_r^{\max}(h), S_B) \\ &\quad + (r - \tfrac{1}{r}) \sum_{y_i=-1} w_i (1 - \|h(\mathbf{x}_i) + 1\|_\infty). \\ &= (\tfrac{1}{r} - 1) W_+ + \tfrac{1}{r}\Gamma(\phi_r^{\max}(h), S_B) \\ &\quad + (r - \tfrac{1}{r})(W_- - \sum_{y_i=-1} \|h(\mathbf{x}_i) + 1\|_\infty).\end{aligned}$$

Thus the lemma is proven. ■

Proof:[of Theorem 8(b)] Denote

$$h^* \triangleq \operatorname*{argmax}_{h\in\mathcal{H}} \Gamma(\phi_r^{\max}(h), S_B).$$

We start with a similar inference to the one taken in the proof of Theorem 8(a). We use Lemma 9 and Lemma 11. Instead of Lemma 10 we use the trivial fact that if Eq. (7) holds, then $\Gamma(h_I, S_I) \geq \Gamma(h^*, S_I)$. Using these three facts, and replacing h_+^* with h^* and γ_+^* with γ^* in Eq. (9), we get that

$$\begin{aligned}\Gamma(\phi_r^{\max}(h_I), S_B) \geq \\ \geq (1-\tfrac{1}{r})(1-2W_+) + \tfrac{1}{r}\gamma^* \\ -(r-\tfrac{1}{r})\sum_{y_i=-1} w_i \|h^*(\mathbf{x}_i)+1\|_\infty.\end{aligned} \tag{18}$$

Now, assuming w.l.o.g that $W(S_B)=1$, we have

$$\begin{aligned}\gamma^* &= \sum_{i\in[m]} w_i y_i \max_{j\in[r]}\{h^*(x_i^j)\} \\ &= \sum_{i\in[m]} w_i y_i (\|h^*(\mathbf{x}_i)+1\|_\infty - 1) \\ &= \sum_{y_i=+1} w_i y_i (\|h^*(\mathbf{x}_i)+1\|_\infty - 1) \\ &\quad + \sum_{y_i=-1} w_i y_i (\|h^*(\mathbf{x}_i)+1\|_\infty - 1) \\ &\leq W_+ + \sum_{y_i=-1} w_i y_i (\|h^*(\mathbf{x}_i)+1\|_\infty - 1) \\ &= W_+ + W_- - \sum_{y_i=-1} w_i \|h^*(\mathbf{x}_i)+1\|_\infty \\ &= 1 - \sum_{y_i=-1} w_i \|h^*(\mathbf{x}_i)+1\|_\infty.\end{aligned}$$

Therefore

$$\sum_{y_i=-1} w_i \|h^*(\mathbf{x}_i)+1\|_\infty \leq 1-\gamma^*.$$

From Eq. (18) we thus have that

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

Therefore, similarly to the proof of Theorem 8(a), we have

$$\gamma \geq \max\left\{1 - r - 2(1-\tfrac{1}{r})W_+ + r\gamma^*, 2W_+ - 1\right\}.$$

Equating the two maximization options, we get

$$W_+^* = \frac{r}{4r-2}(2 - r + r\gamma^*).$$

Substituting W_+^* for W_+ in $2W_+ - 1$, we have

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

To guarantee $\gamma \geq 0$, we require $\gamma^* \geq 1 - \frac{1}{r^2}$. ■

Proof:[of Lemma 16] The claim is proven inductively using the conditions in Eq. (10). The case for $\mathbf{x} \mapsto x^j$ is trivial. For $\mathbf{x} \mapsto \max(\mathbf{f}(\mathbf{x}))$, let $\mathbf{f} = (f_1, \ldots, f_k)$, and assume that the Lipschitz condition holds for f_i, $1 \leq i \leq k$, that is

$$\|\mathbf{f}(\mathbf{a}) - \mathbf{f}(\mathbf{b})\|_\infty \leq \|\mathbf{a}-\mathbf{b}\|_\infty. \tag{19}$$

Note that $\max(\mathbf{f}(\mathbf{a})) = \|\mathbf{f}(\mathbf{a})+1\|_\infty - 1$. Therefore, by the triangle inequality and Eq. (19),

$$\begin{aligned}&|\max(\mathbf{f}(\mathbf{a})) - \max(\mathbf{f}(\mathbf{b}))| = \\ &= |\,\|\mathbf{f}(\mathbf{a})+1\|_\infty - \|\mathbf{f}(\mathbf{b})+1\|_\infty| \\ &\leq \|\mathbf{f}(\mathbf{a}) - \mathbf{f}(\mathbf{b})\|_\infty \\ &\leq \|\mathbf{a}-\mathbf{b}\|_\infty.\end{aligned}$$

For $\mathbf{x} \mapsto \min(\mathbf{f}(\mathbf{x}))$ again assume Eq. (19) holds. To prove the claim note that

$$\min(\mathbf{f}(\mathbf{a})) = -\max(-\mathbf{f}(\mathbf{a})) = 1 - \|-\mathbf{f}(\mathbf{a})+1\|_\infty,$$

Therefore using the triangle inequality we have,

$$\begin{aligned}&|\min(\mathbf{f}(\mathbf{a})\}) - \min(\mathbf{f}(\mathbf{b}))| = \\ &= |1 - \|-\mathbf{f}(\mathbf{a})+1\|_\infty - (1 - \|-\mathbf{f}(\mathbf{b})+1\|_\infty)| \\ &= |\,\|-\mathbf{f}(\mathbf{b})+1\|_\infty - \|-\mathbf{f}(\mathbf{a})+1\|_\infty| \\ &\leq \|\mathbf{f}(\mathbf{a}) - \mathbf{f}(\mathbf{b})\|_\infty \\ &\leq \|\mathbf{a}-\mathbf{b}\|_\infty.\end{aligned}$$

The claim has thus been proven for all three conditions. ■