Weak to Strong Learning from Aggregate Labels

Yukti Makhija[1] **Rishi Saket**[1]

[1]Google DeepMind. {yuktimakhija,rishisaket}@google.com

Abstract

In learning from aggregate labels, the training data consists of sets or "bags" of feature-vectors (instances) along with an aggregate label for each bag derived from the (usually $\{0,1\}$-valued) labels of its constituent instances. In *learning from label proportions* (LLP), the aggregate label of a bag is the average of the instance labels, whereas in *multiple instance learning* (MIL) it is the OR i.e., disjunction. The goal is to train an instance-level predictor that maximizes the accuracy which is the fraction of *satisfied* bags i.e., those on which the model's induced labels are consistent with the target aggregate label. A weak learner in this context is one which has at a constant accuracy < 1 on the training bags, while a strong learner's accuracy can be arbitrarily close to 1. We study the problem of using a weak learner on such training bags with aggregate labels to obtain a strong learner. In a novel result, our work proves the impossibility of boosting in the LLP setting using weak learners of any accuracy < 1 by constructing a collection of bags for which such weak learners (for any weight assignment) exist, while not admitting any strong learner. A variant of this construction also rules out boosting in MIL for a non-trivial range of weak learner accuracy. In the LLP setting however, we show that a weak learner (with small accuracy) on large enough bags can in fact be used to obtain a strong learner for small bags, in polynomial time. We also provide more efficient, sampling based variant of our procedure with probabilistic guarantees which are empirically validated on three real and two synthetic datasets.

1 INTRODUCTION

In traditional, fully supervised learning, the training data consists of a collection of labeled feature-vectors (i.e., training examples) $\{(\mathbf{x}_i \in \boldsymbol{\mathcal{X}}, y_i = y(\mathbf{x}_i))\}_{i=1}^n$, for some domain $\boldsymbol{\mathcal{X}}$ where the mapping y provides the feature-vector labels. In this paper we will consider the binary setting i.e., the labels are $\{0,1\}$-valued. The usual training goal is to find a good classifier $f : \boldsymbol{\mathcal{X}} \to \{0,1\}$ which maximizes the training accuracy $|\{i : f(\mathbf{x}_i) = y_i\}|/n$. In recent times however, due to privacy [Rueping, 2010] or feasibility [Chen et al., 2004] constraints, in many applications the training label for each training example is not available. Instead, the training data consists of sets or *bags* of feature-vectors along with only the *average* or equivalently *sum* of the labels for each bag since bag size is known. This is called *learning from label proportions* (LLP) in which the training set consists of labeled bags $\{(B_j, \overline{y}_j\}_{j=1}^m$ where $B_j \subseteq \boldsymbol{\mathcal{X}}$ and $\overline{y}_j = \sum_{\mathbf{x} \in B_j} y(\mathbf{x})$. The training goal is to fit a good classifier $f : \boldsymbol{\mathcal{X}} \to \{0,1\}$ on this bag-level training data. A related problem is *multiple instance learning* (MIL) in which the label for each bag is the OR i.e., the boolean disjunction of the labels of its constituent feature vectors, while the goal of fitting a good feature-vector classifier remains the same. A natural metric for the goodness of fit in the LLP setting is to maximize the bag-level accuracy i.e., the fraction of *satisfied* training bags, where a bag $(B, \overline{y})$ is satisfied if $\overline{y} = \left(\sum_{\mathbf{x} \in B} f(\mathbf{x})\right)$. An analogous notion of accuracy for MIL is if $\overline{y} = \left(\bigvee_{\mathbf{x} \in B} f(\mathbf{x})\right)$. Recent works [Saket, 2021, 2022] have studied the computational learning aspect of LLP and MIL, and in particular showed that the problem of finding classifiers (even in the realizable case) of high bag-level accuracy can be NP-hard.

In supervised classification, *boosting* (see [Freund and Schapire, 1995, Schapire and Freund, 2012]) is a well known meta-technique which, given a training dataset uses an ensemble (typically a majority) of *weak* classifiers (on reweighed data) to output a hypothesis which has accuracy arbitrarily close to 1 i.e., a *strong* classifier. In the $\{0,1\}$-

labels case a weak classifier has accuracy at least $(1/2 + \varepsilon)$ for some $\varepsilon > 0$, while that for a strong classifier is $(1 - \nu)$ where ν can be made arbitrarily small. Thus, while a strong classifier is always a weak classifier (by making ν small enough), a weak classifier with accuracy $(1/2 + \varepsilon)$ is not strong unless ε can be made arbitrarily close to $1/2$ (see Sections 2.3.1 and 2.3.2 of [Schapire and Freund, 2012]). Note that the threshold of $1/2$ for weak classification is the expected accuracy of random prediction on the training set. In the rest of the paper, the notion of accuracy shall be used for bag-level accuracy in LLP or MIL, unless otherwise specified.

To address the algorithmic learning problems in LLP and MIL, one could hope to apply boosting techniques to LLP and MIL settings as well. Here, we can define a weak classifier having some constant accuracy on the bags, while the notion of a strong classifier remains the same: that with an arbitrarily high accuracy. For LLP, recent works [Saket, 2021, 2022] have given halfspace learning algorithms achieving accuracy $(2/5)$ and $(1/12)$ on satisfiable collections of 2-sized and 3-sized bags respectively. These algorithms are obtained by rounding a semi-definite programming relaxation, which is a standard algorithmic tool. It is plausible that weak classifiers can exist for larger bag sizes as well, possibly for special cases of feature-vector distributions or function classes other than halfspaces. Therefore, we ask:

is there a way to do boosting using weak-classifiers to obtain a strong classifier in learning from aggregate labels?

In this work we show that the above is *impossible* even on 2-sized bags for (i) LLP using weak classifiers of any accuracy < 1, and (ii) for MIL using weak classifiers of any accuracy $< 2/3$. Specifically, we construct a collection of bags such that any probability distribution over the bags admits a weak classifier of the desired accuracy, while the original collection does not admit *any* strong classifier i.e., any labeling to the underlying feature vectors satisfies at most some constant < 1 fraction of the bags. We note that on bags of size 2, for both LLP and MIL the worst-case accuracy obtained by using the random or any constant-valued classifier (all 0s or all 1s), is $1/2$. So, even for MIL we rule out boosting using weak classifiers with non-trivial accuracy in $[1/2, 2/3)$. Our impossibility of boosting stands in contrast to previous work (e.g. [Auer and Ortner, 2004, Qi et al., 2018]) which empirically evaluate boosting heuristics for LLP and MIL – our results are the first to show that such algorithms cannot provably yield a strong classifier.

While the above impossibility results are applicable to the boosting framework, one can ask:

is there some other way to derive a strong classifier from weak classifiers?

Our next result answers this question in the affirmative for LLP: a weak classifier (of any constant accuracy $\gamma > 0$) on large bags can be used to derive a strong classifier on a training set of (smaller) *original* bags. These large or *composite* bags are each a union of t training bags, where t depends only on γ and the desired accuracy of the strong classifier. While on m training bags, the number of ($\approx m^t$) unions are polynomial-time for constant t, we also provide a significantly more efficient sampling version of this approach which provides the same guarantees with high probability. These are to the best of our knowledge the first methods obtaining strong classifiers from weak classifiers for LLP. For MIL on the other hand the question of such weak to strong learning remains open.

1.1 PREVIOUS RELATED WORK

Multiple Instance Learning (MIL). The study by Dietterich et al. [1997] introduced MIL for drug activity detection, where the bag label is modeled as an OR of its (unknown) instance labels, all labels are $\{0, 1\}$-valued. The goal, given such a dataset of bags, is to train a classifier for instance labels. Theoretically, Blum and Kalai [1998] proved that noise tolerant PAC learnability implies MIL PAC learnability for iid bags, and generalization bounds for the classification error on bags were provided by Sabato and Tishby [2012]. Methods including logistic regression, maximum likelihood and boosting with differentiable approximations to the OR function [Ray and Craven, 2005, Ramon and De Raedt, 2000, Zhang et al., 2005] have been proposed. Diverse-density (DD) method [Maron and Lozano-Pérez, 1997] and its EM-based variant, EM-DD [Zhang and Goldman, 2001] are specialised MIL techniques. Over the years this approach has found many applications in numerous areas, including drug discovery [Maron and Lozano-Pérez, 1997], analysis of videos [Sikka et al., 2013], medical images [Wu et al., 2015], time series [Maron, 1998] and information retrieval [Lozano-Pérez and Yang, 2000].

Learning from Label Proportions (LLP). A variety of specialized LLP methods have been introduced till date: de Freitas and Kück [2005] and Hernández-González et al. [2013] developed MCMC techniques, Musicant et al. [2007] adapted traditional supervised learning techniques like k-NN and SVM, while clustering based methods were proposed by Chen et al. [2009] and Stolpe and Morik [2011]. Further, Quadrianto et al. [2009] and Patrini et al. [2014] devised specialized learning algorithms using bag-label mean estimates, and Yu et al. [2013] developed an SVM approach with bag-level constraints. Newer methods involve deep learning [Kotzias et al., 2015, Dulac-Arnold et al., 2019, Liu et al., 2019, Nandy et al., 2022] and others leverage characteristics of the distribution of bags [Saket et al., 2022, Zhang et al., 2022, Chen et al., 2023, Busa-Fekete et al., 2023]. The theoretical foundations of LLP were investigated by Yu et al. [2014], who defined the problem within the PAC framework and established bounds on the general-

ization error for the label proportion regression task. Recent work by Saket [2021], Saket [2022] and Brahmbhatt et al. [2023] addressed bag-classification using linear classifiers, providing algorithmic and hardness bounds. Applications of LLP include privacy in online advertising [O'Brien et al., 2022], high energy physics [Dery et al., 2017] and IVF predictions [Hernández-González et al., 2018].

Boosting. The first boosting algorithm was given by Schapire [1989] which was followed by a more efficient algorithm by Freund [1990] and subsequently the famous AdaBoost algorithm [Freund and Schapire, 1995]. Further work [Chen and Guestrin, 2016, Warmuth et al., 2008, Freund, 2001] resulted in the development of several boosting techniques, while Mason et al. [1999] showed that several boosting algorithms (including AdaBoost [Freund and Schapire, 1995] and LogitBoost [Friedman et al., 2000]) implicitly perform gradient descent in the functional space and fall into the AnyBoost framework. Related techniques include ensemble methods such as bootstrapping aggregation (bagging) and stacking [Mienye and Sun, 2022].

If we consider bags themselves as examples, one can directly apply existing boosting frameworks to obtain strong bag-level classifiers (see for e.g. [Lai et al., 2023]). However, our goal is to obtain feature-vector level strong classifiers with high accuracy on bags. Previous works have adapted a subset of the above mentioned boosting approaches to LLP [Viola et al., 2005, Auer and Ortner, 2004, Qi et al., 2018] – however they are empirically evaluated heuristics and not guaranteed to output strong classifiers. For MIL, Sabato and Tishby [2012] show that an accurate instance-level PAC-learner can be used as an oracle in a boosting subroutine to obtain an MIL PAC-learner. Our results on the other hand rule out boosting weak MIL learners to strong MIL learners, and are complementary to the algorithmic results of Sabato and Tishby [2012].

1.2 PROBLEM DEFINITION AND OUR RESULTS

Let $\boldsymbol{\mathcal{X}} \subseteq \mathbb{R}^d$ for some $d \in \mathbb{Z}^+$ be the space of feature-vectors, while a *bag* B is a finite subset of $\boldsymbol{\mathcal{X}}$. Let $\mathcal{Y} \subseteq \mathbb{R}$ be the space of feature-vector labels, and $\overline{\mathcal{Y}} \subseteq \mathbb{R}$ be the space of bag-level aggregate labels with some aggregation function Agg mapping finite $\mathcal{Y}$-valued tuples to $\overline{\mathcal{Y}}$. We say that a bag $B = (\mathbf{x}_1, \dots, \mathbf{x}_q)$ with aggregate label σ is *satisfied* by a classifier $f : \boldsymbol{\mathcal{X}} \to \mathcal{Y}$ if $\mathsf{Agg}(f(\mathbf{x}_1), \dots, f(\mathbf{x}_q)) = \sigma$. For convenience we will use bag to refer to a bag and its aggregate-label. We illustrate this in Figure 1.

An m-sized *training set* $\mathcal{B}$ is a collection $\{(B_j, \sigma_j) \in 2^{\boldsymbol{\mathcal{X}}} \times \overline{\mathcal{Y}}\}_{j=1}^m$ of m bags and their aggregate-labels along with weights $w_j \geq 0$ for bag B_j ($j = 1, \dots, m$) such that $\sum_{j=1}^m w_j = 1$. The *accuracy* of a classifier on $\mathcal{B}$ is the weighted fraction of bags satisfied by it. For bags without weights i.e., the unweighted case, each bag is assumed to have the same weight $(1/|\mathcal{B}|)$.

We define a *weak* classifier to be one with constant accuracy $\gamma > 0$, and a *ν-strong* classifier to have an accuracy $(1 - \nu)$. For ease of exposition we call the latter a strong classifier when ν can be taken to be an arbitrarily small positive constant.

For this study, the underlying feature-vector level task is binary classification, so $\mathcal{Y} = \{0, 1\}$. For multiple instance learning (MIL) the aggregation function is OR and therefore $\overline{\mathcal{Y}} = \{0, 1\}$. On the other hand, in learning from label proportions (LLP) we take the aggregation function to be SUM i.e., the real sum of labels, and therefore $\overline{\mathcal{Y}} = \{0, 1, 2, \dots\}$. Note that for LLP we could have equivalently taken average as the aggregation (since the size of any bag is known), however for convenience we use SUM.

We also define the $\mathsf{Trv}_{\mathsf{LLP}}(\mathcal{B})$ for a collection of LLP bags, to denote the trivial accuracy threshold on $\mathcal{B}$. Specifically, it is the minimum weighted accuracy given by the best among the random classifier and the two constant valued classifiers (all 0s and all 1s classifiers), over all possible weight assignments to the bags $\mathcal{B}$. For a collection of MIL bags $\mathcal{B}$, $\mathsf{Trv}_{\mathsf{MIL}}(\mathcal{B})$ is defined analogously.

We shall also use the *halfspace* classifier whose value at point $\mathbf{x} \in \mathbb{R}^d$ is given by $\mathsf{pos}(\langle \mathbf{r}, \mathbf{x} \rangle + c)$ for some $\mathbf{r} \in \mathbb{R}^d$, $c \in \mathbb{R}$ where $\mathsf{pos}(a) = 1$ if $a > 0$ and 0 otherwise. We say that the halfspace passes through the origin i.e., is *homogeneous* if $c = 0$. Next we state this paper's results.

1.2.1 Our Results

We begin with the impossibility results for boosting in the LLP (Theorem 1.1) and MIL (Theorem 1.2) settings. These theorems coupled with the definition of the boosting meta algorithm (Section 2.1) imply our impossibility results.

Theorem 1.1 (Impossibility of boosting in LLP). *Let $\alpha \in [1/2, 1)$ be any constant. Then, for any arbitrarily small constant $\varepsilon > 0$ there exist positive integers d, m and a collection of bags $\mathcal{B} = \{B_j \subseteq \mathbb{R}^d\}_{j=1}^m$ where $|B_j| = 2$ and the aggregate label (i.e. sum of labels in LLP setting) of B_j is 1 ($j = 1, \dots, m$) and the following properties are satisfied:*

(Existence of weak halfspace classifiers): *For any assignment of weights w_j to B_j ($j = 1, \dots, m$) such that $\sum_{j=1}^m w_j = 1$, for the weighted collection of bags there is a halfspace classifier with accuracy α.*

(No Strong Classifier): *For the unweighted set of bags $\{B_j \subseteq \mathbb{R}^d\}_{j=1}^m$ there is no classifier $f : \cup_{j=1}^m B_j \to \{0, 1\}$ having accuracy greater than $\alpha + \varepsilon$.*

The above theorem, proved in Section 3, is optimal from multiple perspectives: firstly the bags are of size at most 2

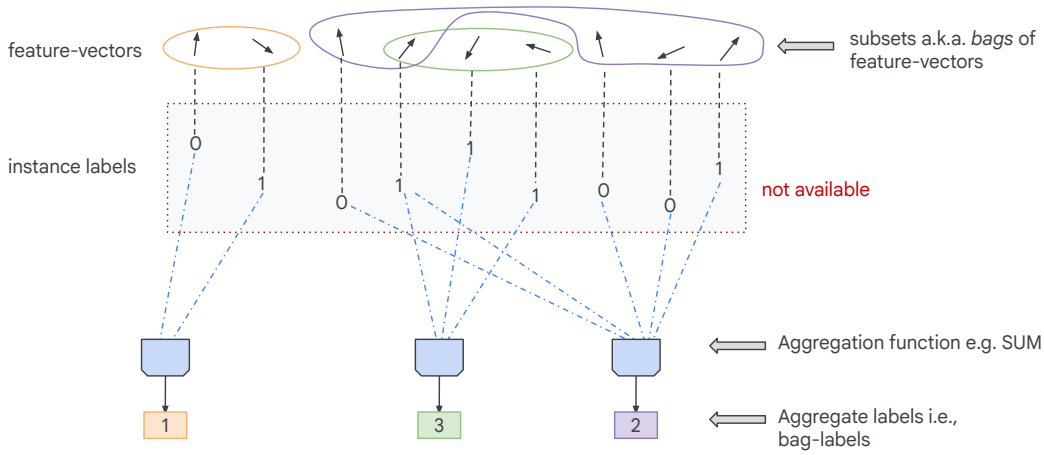

Figure 1: Aggregate Labels

whereas when bags are all of size 1 (i.e., supervised learning) boosting is indeed possible, showing that as soon as we transition from the fully supervised to the LLP setting in terms of bag size, boosting becomes impossible. Secondly, the result shows that even if weak learners of *any* constant accuracy in $[1/2, 1)$ exist, there is no classifier with even a slightly greater accuracy, by applying the theorem taking α as the accuracy and ε the slight increment in the accuracy to be ruled out. This rules out any non-trivial advantage of boosting, let alone the possibility of obtaining a strong classifier. In Appendix A we give a simple argument showing that $\mathsf{Trv}_{\mathsf{LLP}}(\mathcal{B}) = 1/2$ for the bags $\mathcal{B}$ constructed in the above theorem. We now state our result on the impossibility of boosting in the MIL setting.

Theorem 1.2 (Impossibility of boosting in MIL)**.** *For any arbitrarily small constant $\varepsilon > 0$ there exists positive integer m and a collection of bags $\mathcal{B} = \{B_j \subseteq \mathbb{R}^d\}_{j=1}^m$ along with the aggregate labels σ_j for B_j where $|B_j| = 2$ ($j = 1, \ldots, m$) and the following properties are satisfied:*

(Existence of weak halfspace classifiers): *For any assignment of weights w_j to B_j ($j = 1, \ldots, m$) such that $\sum_{j=1}^m w_j = 1$, for the weighted collection of bags there is a halfspace classifier with accuracy $2/3 - \varepsilon$.*

(No Strong Classifier): *For the unweighted set of bags $\{B_j \subseteq \mathbb{R}^d\}_{j=1}^m$ there is no classifier $f : \cup_{j=1}^m B_j \to \{0, 1\}$ having accuracy greater than $3/4$.*

The above theorem, whose proof is deferred to Appendix B, shows that in the MIL setting, weak classifiers with any accuracy $< 2/3$ cannot be boosted to a strong classifier with accuracy $> 3/4$. As shown in Appendix A, $\mathsf{Trv}_{\mathsf{MIL}}(\mathcal{B}) = 1/2$ for the bags $\mathcal{B}$ of the above theorem, and therefore our result applies to non-trivial weak classifier accuracy in $(1/2, 2/3)$.

Next we state our results (proved in Section 4) in the LLP setting for obtaining a strong classifier on a collection of original bags using a weak classifier on a derived collection of larger, composite bags. In this case we consider unweighted collection of bags, since a weighted collection of m bags can easily be converted into an unweighted collection of Tm bags while preserving the accuracy of any classifier up to an additive error of $O(1/T)$ (see Appendix C). To state our result we assume that there is an oracle $\mathcal{O}_{q,\alpha}(\overline{\mathcal{B}})$ which given weighted collection of bags $\overline{\mathcal{B}}$ along with their aggregate labels, where each bag has size at most q, outputs a classifier f with weighted accuracy α on $\overline{\mathcal{B}}$.

Theorem 1.3 (Weak to Strong LLP Learning)**.** *For parameters $\alpha, \varepsilon > 0$ there exists $t = O(1/(\varepsilon\alpha^2))$, and algorithms $\mathcal{A}_1$ and $\mathcal{A}_2$ s.t. given an unweighted collection of m bags $\mathcal{B}$, where $k = \max_{(B,\sigma)\in\mathcal{B}} |B|$ and $n := \left|\cup_{(B,\sigma)\in\mathcal{B}} B\right|$, and assuming that $\mathcal{O}_{kt,\alpha}$ exists,*

- *$\mathcal{A}_1$ creates a weighted collection $\overline{\mathcal{B}}_1$ of at most m^{t+1} bags each of size at most kt such that $\mathcal{O}_{kt,\alpha}(\overline{\mathcal{B}}_1)$ outputs a classifier which has accuracy $(1 - \varepsilon)$ on $\mathcal{B}$.*
- *for any $\delta > 0$, $\mathcal{A}_2$ creates a random collection $\overline{\mathcal{B}}_2$ of $s = O\left(\frac{1}{\alpha}\left(n + \log\left(\frac{1}{\delta}\right)\right)\right)$ each of size at most kt such that $\mathcal{O}_{kt,\alpha}(\overline{\mathcal{B}}_2)$ has accuracy $(1 - \varepsilon)$ on $\mathcal{B}$ with probability at least $(1 - \delta)$. If $\mathcal{O}_{kt,\alpha}$ is guaranteed to output a classifier of VC dimension r then $s = O\left(\frac{r}{\alpha}\log\left(\frac{n}{r}\right) + \log\left(\frac{1}{\delta}\right)\right)$ suffices.*

Theorem 1.3 presents algorithms that, when applied to collections of original bags in the LLP setting, yields high-accuracy classifiers by employing weak classifiers trained on a reasonably sized collections of composite bags. This can in particular be achieved by an efficient randomized algorithm $\mathcal{A}_2$. We also conduct experiments (see Section 5) – on both real and synthetic datasets – to demonstrate the effectiveness of $\mathcal{A}_2$. We use it to construct a limited collection of composite bags from a given collection of original bags and experimentally show that a weak classifier on the composite bags yields one with significantly higher accuracy on the constituent original bags.

1.3 OVERVIEW OF TECHNIQUES

Impossibility of Boosting in LLP (Theorem 1.1). Our construction follows from the well-known *semi-definite programming* (SDP) integrality gap of Feige and Schechtman [2002] for the Max-Cut problem. In this, for some arbitrarily small $\varepsilon > 0$, with d depending on ε, the vertices of the graph are given by points on the $(d-1)$-dimensional unit sphere $\mathbb{S}^{d-1}$. For any constant $\alpha \in [1/2, 1)$, each edge is between points that are at an angle of at least $\alpha\pi$. Using techniques related to spherical isoperimetry and concentration of measure in high dimensions, the authors prove that there is no cut in the graph separating more than $(\alpha+\varepsilon)$-fraction of the edges. By creating a 2-sized bag corresponding to each edge with latter's two end-points being the bag's two feature-vectors, we create a collection of bags, and for each one we assign an aggregate label 1 i.e., any bag is satisfied if exactly one of its feature-vectors is labeled 1 or equivalently the corresponding edge is separated. The cut upper bound of $(\alpha+\varepsilon)$ thus directly gives us the upper bound on the best possible accuracy of any classifier. On the other hand, since the angle between the feature-vectors of any edge is at least $\alpha\pi$, a random halfspace passing through the origin – given by $\mathsf{pos}\left(\mathbf{r}^{\mathsf{T}}\mathbf{x}\right)$ for a random unit vector $\mathbf{r}$ – has expected accuracy α for any weight assignment to the bags, and therefore there is some halfspace achieving accuracy α.

Impossibility of Boosting in MIL (Theorem 1.2). Since the aggregation function is OR the Max-Cut construction of Feige and Schechtman [2002] is not applicable. Instead we hand-craft the set of bags as follows. The set of feature-vectors is all points on the unit circle $\mathbb{S}^1$ and for some $\alpha \in (1/2, 1)$, we create a bag with two points if the angle between them is exactly $\alpha\pi$ and give an aggregate label 1 to all such two sized bags (let us call them 1-bags). We also construct 2-sized bags with aggregate label 0 when the angle between two points is exactly $(1-\alpha)\pi$ (called as 0-bags). If we consider any reweighted collection of these bags then a simple threshold based case-analysis yields weak classifier of accuracy $2/3 - (1-\alpha)/2$. To rule out any strong classifier, we consider a labeling where z-fraction of the points in $\mathbb{S}^1$ are labeled as 1. We show that the maximum accuracy possible is 3/4 which is achieved at $z = 1/2$. We choose $\alpha = 1 - \varepsilon$ while losing an additional error of $\varepsilon/2$ in the weak-classifier accuracy due to discretization to obtain the desired bounds.

Weak to Strong LLP Learning (Theorem 1.3). The main idea is, given a collection of original bags $\mathcal{B}$, to construct all possible composite bags which are unions of up to t bags from $\mathcal{B}$. Note that the aggregate label for the union is simply the sum of the aggregate labels of the constituent bags, and the error of a classifier w.r.t. the aggregate label on the union of bags is the sum of errors on the constituent bags. Let f be a classifier with accuracy $\gamma > 0$ on the composite bags, and assume for a contradiction that f has accuracy less than $(1-\varepsilon)$ on $\mathcal{B}$, for some $\varepsilon > 0$. Call the bags in $\mathcal{B}$ on which f has a non-zero error $\in \mathbb{Z} \setminus \{0\}$ w.r.t. the aggregate label, as the *error* bags. Now, if t is large enough then a random set of t bags from $\mathcal{B}$ has, with high probability $\approx \varepsilon t$ error bags. Using a sampling argument we show that the error on the union of t random bags from $\mathcal{B}$ is distributed like a random Bernoulli combination of the errors on $\approx 2\varepsilon t$ bags. We then apply the Littlewood-Offord-Erdős anti-concentration lemma to obtain that with probability at least $(1 - O(1/(\sqrt{\varepsilon t}))$, the union of the bags has non-zero error induced by f. By choosing t large enough we obtain a contradiction with the accuracy of α on the composite bags. Standard sampling techniques can be applied to obtain a more efficient procedure with high probability guarantees.

We also note here that the above algorithmic techniques are inapplicable to the MIL setting (see Appendix A.1). In Appendix A.2 we informally describe how are results and techniques can be applied to multi-class classification settings of LLP and MIL, in which the aggregate label of a bag is a histogram over the label-set.

2 PRELIMINARIES

Lemma 2.1 (Chernoff Bounds). *Let $X = \sum_{i=1}^{n} X_i$, where X_i = 1 with probability p_i and $X_i = 0$ with probability $1 - p_i$, and all X_i are independent. Let $\mu = \mathbb{E}(X) = \sum_{i=1}^{n} p_i$. Then (i) Lower Tail: $\Pr[X \leq (1-\eta)\mu] \leq e^{-\eta^2\mu/2}\ \forall\ 0 < \eta < 1$, and (ii) Upper Tail: $\Pr[X \leq (1+\eta)\mu] \leq e^{-\eta^2\mu/(2+\eta)}\ \forall\ 0 \leq \eta$.*

Lemma 2.2 (Littlewood-Offord-Erdős Lemma Erdös [1945]). *Let $X_1, X_2, \ldots, X_n$ be* i.i.d *$\{0,1\}$-Bernoulli random variables with $\Pr[1] = 1/2$, and let $a_1, a_2, ..., a_n \in \mathbb{R}$ s.t. $|a_i| \geq 1,\ \forall i \in [n]$. Then, there exists an absolute constant $C > 0$ such that*

$$\Pr_{X_1,\ldots,X_n}\left[\left|\sum_{i\in[n]} a_i X_i + \theta\right| \leq 1\right] \leq \frac{C}{\sqrt{n}}$$

for any constant θ.

Theorem 2.3 (Theorem 3.7 from Anthony and Bartlett [1999]). *For a $\{0,1\}$-valued class $\mathcal{H}$ of functions with VC-dimension $\text{VC-dim}(\mathcal{H}) = v$, let $\Pi_{\mathcal{H}}(n)$ denote the maximum number of possible $\{0,1\}$-labelings to any set of n points from the domain of $\mathcal{H}$. If $n \leq v$, $\Pi_H(n) \leq 2^n$ and for $n > v$, $(\frac{en}{v})^v$. Refer to Section 3.3 of Anthony and Bartlett [1999] for more details on VC Dimension.*

2.1 BOOSTING META ALGORITHM FOR AGGREGATE LABEL SETTING

Given a collection of bags and aggregate labels, a prototypical boosting algorithm (given in Figure 2) in the aggregate label setting, involves repeating certain steps over

some number of rounds: in each round the training data is reweighed, for which a weak classifier is computed. The final output is some function over the ensemble of computed weak classifiers.

Input: $\mathcal{B} = (B_i, \overline{y}_i)_{i=1}^m$: Collection of bags and aggregate labels, $D_1(i) = 1/m$: initial weight distribution associated with each bag, T: Number of steps of boosting.
1. for $t \in [T]$:
1.1 Train a weak classifier $h_t : \boldsymbol{\mathcal{X}} \to \{0, 1\}$ for the bag distribution D_t.
1.2 Using $\{h_r\}_{r=1}^t$, compute a new distribution D_{t+1} over $\mathcal{B}$.

2. For some g, output $h^* = g(h_1, \ldots, h_T)$ as a (presumably) strong classifier for $\mathcal{B}$.

Figure 2: Boosting for aggregate label setting

Note that bootstrapping aggregation (bagging) ensemble method [Mienye and Sun, 2022] can also be framed as a boosting algorithm. This is because the weak learners in bagging are trained in parallel using independent samples from the training data. This fits the iterative framework of boosting, where each iteration can be made independent of the rest, using an independent random sample of the training data which is a special reweighting of the dataset. Stacking ensemble methods [Mienye and Sun, 2022], which are more general than bagging as they allow heterogeneous parallel weak learners, also align with the boosting meta-algorithm. Therefore, the impossibility results applies to bagging and stacking as well.

3 IMPOSSIBILITY OF BOOSTING IN LLP

The Max-Cut problem is: given an undirected graph $G(V, E)$ find a cut given by the assignment $g : V \to \{0, 1\}$ which separates the maximum number of edges in E i.e., maximizes $|\{e = \{u, v\} \in E | g(u) \neq g(v)\}|$. We shall use the following construction of graph $G_{\text{FS}}(V_{\text{FS}}, E_{\text{FS}})$ given in Sec. 3.1 of Feige and Schechtman [2002]:

Construction. Let $\alpha\pi = \theta \in [\pi/2, \pi)$ and $\varepsilon > 0$ be an arbitrarily small parameter such that $\theta + \varepsilon\pi < \pi$. Let $d = O(1/\varepsilon \log(1/\varepsilon))$ and $\gamma = \varepsilon^2/(2d)$. Divide the $(d-1)$-dimensional unit sphere $\mathbb{S}^{d-1}$ into $\left(\frac{O(1)}{\gamma}\right)^d$ equal sized cells of diameter at most γ each (this is shown to be possible in Lemma 21 of Feige and Schechtman [2002]). From each cell pick an arbitrary point $\mathbf{v}$ and add it to V_{FS}. Add an edge $\{\mathbf{u}, \mathbf{v}\}$ to E_{FS} for each pair of points $\mathbf{u}, \mathbf{v} \in V_{\text{FS}}$ whose angle is between θ and $\theta + \varepsilon$.

Section 3.1 of Feige and Schechtman [2002] shows[1] that

$$\Pr_{\{\mathbf{u},\mathbf{v}\} \in E_{\text{FS}}} [g(\mathbf{u}) \neq g(\mathbf{v})] \leq \theta/\pi + O(\varepsilon^2) = \alpha + O(\varepsilon^2) \quad (1)$$

for any $g : V_{\text{FS}} \to \{0, 1\}$.

3.1 PROOF OF THEOREM 1.1

Let $G_{\text{FS}}(V_{\text{FS}}, E_{\text{FS}})$ be the graph constructed above using $\theta = \alpha\pi \in [\pi/2, \pi)$ and let ε taken to be the same as that in the statement of Theorem 1.1. Taking V_{FS} to be the underlying set feature-vectors, let the set of bags $\mathcal{B}$ be E_{FS} i.e., each edge $\{\mathbf{y}, \mathbf{v}\}$ is a bag. All aggregate labels are 1, so that any bag is satisfied by $g : V_{\text{FS}} \to \{0, 1\}$ iff the corresponding edge is separated by g.

Now, for any bag $\{\mathbf{u}, \mathbf{v}\}$ in $\mathcal{B}$, from the construction of $G_{\text{FS}}(V_{\text{FS}}, E_{\text{FS}})$, the angle between $\mathbf{u}$ and $\mathbf{v}$ is at least θ. Thus, a random homogeneous halfspace (given by $\mathsf{pos}\left(\mathbf{r}^\mathsf{T}\mathbf{x}\right)$ for $\mathbf{r}$ chosen uniformly at random from $\mathbb{S}^{d-1}$) satisfies the bag with probability at least $\theta/\pi = \alpha$.

Thus for any assignment of weights w_B for bags $B \in \mathcal{B}$, the expected weight of bags satisfied by a random homogeneous halfspace is $\sum_B w_B \Pr_{\mathbf{r} \leftarrow \mathbb{S}^{d-1}} \left[B \text{ is satisfied by } \mathsf{pos}\left(\mathbf{r}^\mathsf{T}\mathbf{x}\right)\right] = \alpha \sum_B w_B$ by linearity of expectation. Therefore, there is one classifier with weighted accuracy α.

The upper bound on the accuracy of *any* classifier on $\mathcal{B}$ follows directly from (1) and small enough $\varepsilon > 0$.

4 WEAK TO STRONG CLASSIFICATION IN LLP

Given $\alpha, \varepsilon > 0$ we set t to be $\frac{32}{\varepsilon}\left(\frac{C_0}{\alpha}\right)^2$ where $C_0 > 0$ is an absolute constant to be decided. We begin by defining in Fig. 3 a distribution $\overline{D}$ over bags $(\overline{B}, \overline{\sigma})$ where $\overline{B}$ is the union of at most t bags from $\mathcal{B}$ and $\overline{\sigma}$ is the sum of their aggregate labels.

To aid our subsequent analysis we shall use the following straightforward lemma.

Lemma 4.1. *For $\kappa \in [0, 1]$ and any subset $\mathcal{S} \subseteq \mathcal{B}$ s.t. $|\mathcal{S}| \geq \kappa|\mathcal{B}|$, in Step 1. of Fig. 3, $\Pr\left[|\{i \mid (B_i, \sigma_i) \in \mathcal{S})\}| < \kappa t/2\right] \leq \exp(-\kappa t/8)$.*

Proof. Since each (B_i, σ_i) independently belongs to $\mathcal{S}$ w.p. κ, $\Pr[(B_i, \sigma_i) \in \mathcal{S}] \geq \kappa$ and therefore $\mu := \mathbb{E}\left[|\{i \mid (B_i, \sigma_i) \in \mathcal{S})\}|\right] \geq \kappa t$. Thus, $\Pr\left[|\{i \mid (B_i, \sigma_i) \in \mathcal{S})\}| < \kappa t/2\right] \leq \Pr\left[|\{i \mid (B_i, \sigma_i) \in \mathcal{S})\}| < \mu/2\right] \leq \exp(-\mu/8) \leq \exp(-\kappa t/8)$, where we use the Chernoff Tail Bound

[1] While Feige and Schechtman [2002] state the proof of (1) for a specific value of θ, the proof applies to all values of $\theta \in [\pi/2, \pi)$.

Input: : Bags $\mathcal{B}$, t.
Steps:

1. Independently for $i = 1, \ldots, t$, let $\mathcal{P}_i = (B_i, \sigma_i)$ where (B_i, σ_i) is sampled u.a.r. from $\mathcal{B}$.
2. Independently for $i = 1, \ldots, t$: set $\mathcal{Q}_i = \mathcal{P}_i$ w.p. $1/2$ and set $\mathcal{Q}_i = \star$ w.p. $1/2$.
3. Output $(\overline{B}, \overline{\sigma})$ where

$$\overline{B} = \bigcup_{\{i \mid \mathcal{Q}_i = (B_i, \sigma_i) \neq \star\}} B_i, \quad \overline{\sigma} = \sum_{\{i \mid, \mathcal{Q}_i = (B_i, \sigma_i) \neq \star\}} \sigma_i \tag{2}$$

Figure 3: Distribution $\overline{D}$.

(Lemma 2.1) using $\eta = 1/2$ and the lower bound of κt for μ. □

4.1 ANALYSIS FOR A FIXED CLASSIFIER h

We prove the following lemma.

Lemma 4.2. *Let $h : \boldsymbol{\mathcal{X}} \to \{0, 1\}$ be a classifier such that h has accuracy $< (1 - \zeta)$ on $\mathcal{B}$. Then,*

$$\Pr_{(\overline{B}, \overline{\sigma}) \leftarrow \overline{D}} \left[\sum_{\mathbf{x} \in \overline{B}} h(\mathbf{x}) = \overline{\sigma} \right] \leq C_0 / \sqrt{\zeta t} + \exp(-\zeta t / 8)$$

for some absolute constant $C_0 > 0$.

Proof. Let $\mathcal{B}_{\text{err}}$ be the *error* bags $(B, \sigma) \in \mathcal{B}$ on which $\sum_{\mathbf{x} \in B} h(\mathbf{x}) \neq \sigma$, so that $|\mathcal{B}_{\text{err}}| \geq \zeta |\mathcal{B}|$. For convenience, we shall abuse the notation $h(B)$ to denote $\sum_{\mathbf{x} \in B} h(\mathbf{x})$, and therefore, for an error bag B, $|h(B) - \sigma| \geq 1$. Depending on the choices in Step 1. of Fig. 3, define the set $I := \{i \mid (B_i, \sigma_i) \in \mathcal{B}_{\text{err}})\}$ and let E_0 be the event that the following occurs: $\{|I| \geq \zeta t / 2\}$. Further, let E_1 be the event that the LHS of the following equivalence occurs:

$$h(\overline{B}) = \overline{\sigma} \Leftrightarrow \sum_{\{i \mid, \mathcal{Q}_i = (B_i, \sigma_i) \neq \star\}} (h(B_i) - \sigma_i) = 0 \tag{3}$$

where $(\overline{B}, \overline{\sigma})$ is the output in Step 3. Now,

$$\begin{aligned} \Pr[E_1] = & \Pr[E_1 | E_0] \Pr[E_0] + \Pr[E_1 | \neg E_0] \Pr[\neg E_0] \\ \leq & \Pr[E_1 | E_0] + \Pr[\neg E_0] \end{aligned}$$

Since $|\mathcal{B}_{\text{err}}| \geq \zeta |\mathcal{B}|$, Lemma 4.1 yields that $\Pr[\neg E_0] \leq \exp(-\zeta t / 8)$. On the other hand, fix the set I and bags $\{(B_i, \sigma_i)\}_{i \in I}$ and let $a_i := h(B_i) - \sigma_i$ $(i = 1, \ldots t)$. Defining $\{X_i \mid i \in I\}$ to be i.i.d $\{0, 1\}$-valued Bernoulli random variables which are 1 w.p. $1/2$, we obtain that $\Pr[E_1] = \Pr[\sum_{i \in I} a_i X_i = 0] \leq C / \sqrt{|I|}$ by applying Lemma 2.2. Therefore, $\Pr[E_1 | E_0] \leq C / \sqrt{(\zeta/2) t}$ and using the above bounds, $\Pr[E_1] \leq C / \sqrt{(\zeta/2) t} + \exp(-\zeta t / 8)$. □

4.2 DETERMINISTIC ALGORITHM $\mathcal{A}_1$

Input: : Bags $\mathcal{B}$, $k = \max_{(B, \sigma) \in \mathcal{B}} |B|$, $\alpha > 0$, t, oracle $\mathcal{O}_{kt, \alpha}$.
Steps:

1. Let $\mathsf{supp}(\overline{D})$ be the support of $\overline{\mathcal{D}}$ (Fig. 3), and for each $(\overline{B}, \overline{\sigma}) \in \mathsf{supp}(\overline{D})$ let its weight $w_{(\overline{B}, \overline{\sigma})}$ be its probability under $\overline{D}$. Let $\overline{\mathcal{B}}$ be $\mathsf{supp}(\overline{D})$ with weights $w_{(\overline{B}, \overline{\sigma})}$.
2. Output the classifier h^* given by $\mathcal{O}_{kt, \alpha}(\overline{\mathcal{B}})$.

Figure 4: Algorithm $\mathcal{A}_1$.

Figure 4 describes algorithm $\mathcal{A}_1$ using[2] the distribution $\overline{D}$ defined in Figure 3. Suppose for a contradiction that the output h^* of $\mathcal{A}_1$ has accuracy $< (1 - \varepsilon)$ on $\mathcal{B}$. Then, from Lemma 4.2 we obtain that the probability that $(\overline{B}, \overline{\sigma})$ sampled from $\mathcal{D}$ is satisfied by h^* is at most $C_0 / \sqrt{\varepsilon t} + \exp(-\varepsilon t / 8)$ which – upon plugging in the value of t – is at most $\alpha / 2$ which contradicts the accuracy of h^* on $\overline{\mathcal{B}}$.

We next describe a more efficient, albeit randomized, variant of the algorithm.

4.3 RANDOMIZED ALGORITHM $\mathcal{A}_2$

Figure 6 provides the algorithm $\mathcal{A}_2$. Fix any h that has accuracy $< (1 - \varepsilon)$ on $\mathcal{B}$. Then, by Lemma 4.2, and our setting of t we obtain that $\Pr_{(\hat{B}, \hat{\sigma}) \leftarrow \overline{D}}[(\hat{B}, \hat{\sigma}) \text{ satisfied by } h] \leq \alpha / 2$. Therefore, in Step 1 of $\mathcal{A}_2$ it is easy to see by monotonicity that

$$\Pr \left[\left| \{ j \in [s] \mid (\hat{B}_j, \hat{\sigma}_j) \text{ satisfied by h} \} \right| \geq \alpha s \right] \leq \mathsf{P} \left[\sum_{\ell = 1}^{s} X_\ell \geq \alpha s \right] \tag{4}$$

where each X_ℓ $(\ell = 1, \ldots, s)$ is an independent $\{0, 1\}$-valued Bernoulli random variable taking value 1 with probability $\alpha / 2$. Therefore, using Chernoff Upper Tail bound from Lemma 2.1 we can upper bound the LHS of (4) by $\exp(-\alpha s / 6)$ which is the upper bound on the probability that h has accuracy $\geq \alpha$ on $\hat{\mathcal{B}}$.

Let $\mathcal{C}$ be the classifier class to which the output of $\mathcal{O}_{kt, \alpha}$ is guaranteed to belong. With n being the total number of distinct feature-vectors in the bags $\mathcal{B}$, $\Pi_{\mathcal{C}}(n)$ (as given in Theorem 2.3) is the number of possible $\{0, 1\}$-assignments to n points induced by classifiers in $\mathcal{C}$. Taking a union-bound over all of them, we obtain that with probability at most $\Pi_{\mathcal{C}}(n) \exp(-\alpha s / 6)$ the output of $\mathcal{A}_2$ has accuracy at least $(1 - \varepsilon)$ on $\mathcal{B}$.

[2] We include in Appendix D an explanation on computing the probabilities under $\overline{D}$.

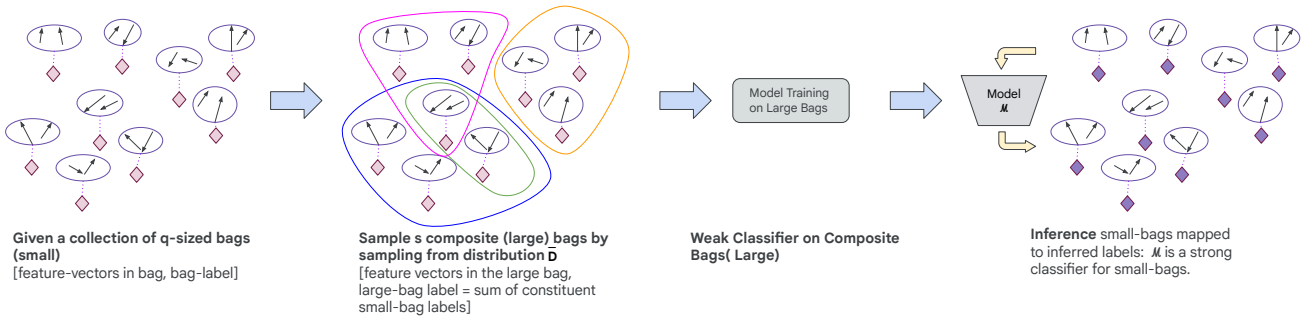

Figure 5: Overview of our proposed randomized algorithm for obtaining strong classifiers on original bags using a weak classifier on composite bags.

Input: : Bags $\mathcal{B}$, $k = \max_{(B,\sigma)\in\mathcal{B}} |B|$, $\alpha > 0$, t, oracle $\mathcal{O}_{kt,\alpha}$, $s \in \mathbb{Z}^+$.
Steps:

1. Let $\hat{\mathcal{B}} = \{(\hat{B}_j, \hat{\sigma}_j)\}_{j=1}^{s}$ be s i.i.d. samples from $\overline{D}$ (Fig. 3).
2. Output the classifier $\tilde{h}$ given by $\mathcal{O}_{kt,\alpha}(\hat{\mathcal{B}})$.

Figure 6: Algorithm $\mathcal{A}_2$.

When $\mathcal{C}$ is unrestricted then $\Pi_{\mathcal{C}}(n) \leq 2^n$ and therefore $\Pi_{\mathcal{C}}(n)\exp(-\alpha s/6) \leq \delta$ is ensured by taking $s = O\left((n + \log(1/\delta))/\alpha\right)$. On the other hand if the VC dimension of $\mathcal{C}$ is at most r, then $\Pi_{\mathcal{C}}(n) \leq (en/r)^r$ (from Theorem 2.3) , and therefore taking $s = O\left(\frac{r}{\alpha}\log\left(\frac{n}{r}\right) + \log\left(\frac{1}{\delta}\right)\right)$ suffices.

We include Figure 5 illustrating how our algorithm trains a strong classifier for original (small) bags using a weak classifier trained on composite (large) bags.

5 EXPERIMENTS

In our experiments, we generate a collection of original q-sized bags as training data using fully supervised datasets. We use a fixed value of $q \in \{5, 15\}$.

Synthetic Datasets. In this case we experiment in the realizable setting for which we select a random linear classifier f^* passing though the origin to provide $\{0, 1\}$-labels to the feature-vectors. For a given bag-size $q \in \{5, 15\}$, we generate two types of bag collections as follows:

1. *Random*: In this case each q-sized bag is created by randomly sampling points uniformly from the unit sphere as its constituent feature vectors.
2. *Hard Bags*: For these bags we first randomly construct pairs of points on the unit-sphere which are either (i) very close but have different labels under f^*, or (ii) nearly antipodal but have the same label. Each bag consists of several such randomly constructed pairs and one random point (since q is odd).

In both the above cases, the aggregate label of a bag is the sum of the labels of its feature-vectors given by f^*.We also have a test-set of labeled feature-vectors whose distribution is given by sampling each u.a.r. from a random training bag.

Real Datasets. We use the following supervised UCI datasets: *Heart* (303 instances, [Janosi and Detrano, 1988]), *Australian* (690 instances, [Quinlan]) and *Adult* (48842 instances, [Becker and Kohavi, 1996]) which have previously been used by Patrini et al. [2014] to evaluate LLP methods. The feature-vector labels are available and the bags are created by partitioning the training-set into -sized bags. The test-set is given by a random subset of 15% of the dataset.

Applying Algorithm $\mathcal{A}_2$. For each collection of training bags, and an appropriate choice of t and s (see Figure 6) we create a collection of s composite bags by sampling each iid from the distribution $\overline{D}$ given in Figure 3.

Model Training. We train a linear model $g(\mathbf{x})$ with a sigmoid activation function on the composite bags using bag-level MSE loss between the aggregate label of a bag and its aggregate prediction. In particular, for a composite bag $\overline{B}$ and aggregate label $\overline{\sigma}$ the contribution to the loss is $\left(\overline{\sigma} - \sum_{\mathbf{x}\in\overline{B}} g(\mathbf{x})\right)^2$. and the total loss is the sum over the composite bags in collection. The optimization is done using a mini-batch training with 512 bags in each mini-batch. The learning rate is 1e-2 with SGD optimizer for all experiments, and the model is trained till it reaches convergence on the instance-level test set.

Results. Tables 1 and 2 have the experimental results for the synthetic, Heart, Australian and Adult datasets respectively. For each setting of q, t and s, we report the mean accuracy and standard deviation on the training set for both composite bags and their constituent original bags, along with the accuracy on test instances, averaged over 15 runs. The main takeaways from the experimental results are:

1. In all experiments, even with low accuracy on composite bags we obtain classifiers with high accuracy on the constituent original bags and even higher accuracy on the

Table 1: Results on the Synthetic Datasets.

q	t	s	Random Bags Composite	Random Bags Original	Random Bags Test Instance	Hard Bags Composite	Hard Bags Original	Hard Bags Test Instance
5	10	5000	52.891 ± 5.196	85.357 ± 3.085	96.067 ± 1.218	32.629 ± 3.439	68.374 ± 4.428	91.120 ± 1.978
		15000	72.295 ± 5.275	93.089 ± 2.057	97.840 ± 0.829	47.276 ± 5.241	81.802 ± 3.789	95.160 ± 1.365
	50	5000	21.330 ± 3.110	85.513 ± 3.434	96.453 ± 0.780	12.789 ± 2.192	68.463 ± 5.828	91.427 ± 1.785
		15000	32.890 ± 5.032	93.076 ± 1.466	97.867 ± 0.626	18.311 ± 2.544	82.562 ± 3.637	95.560 ± 1.299
15	10	5000	21.792 ± 3.189	50.133 ± 7.520	93.037 ± 1.674	14.731 ± 2.337	31.600 ± 5.138	86.855 ± 2.638
		15000	32.259 ± 3.444	68.733 ± 4.334	96.566 ± 0.890	17.115 ± 1.501	40.067 ± 5.189	89.939 ± 1.921
	50	5000	8.674 ± 1.537	52.400 ± 7.079	93.778 ± 2.060	5.252 ± 1.715	34.000 ± 5.438	85.657 ± 3.132
		15000	11.106 ± 3.042	67.467 ± 4.389	96.067 ± 1.412	6.409 ± 1.457	40.800 ± 6.753	91.677 ± 2.336

Table 2: Results on the Real Datasets.

q	t	s	Composite Bags	Original Bags	Test Instance
			Heart		
5	10	2500	24.207 ± 4.418	55.407 ± 8.419	79.911 ± 4.349
		10000	31.337 ± 5.363	65.333 ± 8.516	77.956 ± 3.767
	50	2500	5.356 ± 2.715	47.407 ± 8.172	78.400 ± 3.676
		10000	9.128 ± 3.192	59.556 ± 8.021	77.689 ± 5.622
15	10	2500	12.950 ± 7.030	35.111 ± 15.006	71.378 ± 7.870
		10000	20.539 ± 8.041	49.778 ± 16.498	69.156 ± 7.089
	50	2500	0.803 ± 1.521	26.222 ± 16.226	73.867 ± 5.829
		10000	1.946 ± 2.143	30.667 ± 10.328	72.178 ± 6.852
			Australian		
5	10	3500	24.956 ± 3.709	55.962 ± 5.783	84.275 ± 2.626
		10000	29.774 ± 2.600	62.692 ± 4.319	84.039 ± 1.999
	50	3500	5.454 ± 4.127	53.846 ± 9.449	82.039 ± 3.015
		10000	9.303 ± 2.806	58.141 ± 6.510	82.431 ± 2.837
15	10	3500	10.396 ± 4.906	28.190 ± 8.072	75.313 ± 6.233
		10000	15.746 ± 4.950	37.524 ± 10.792	78.222 ± 5.824
	50	3500	0.257 ± 0.596	24.190 ± 7.233	74.707 ± 5.073
		10000	1.342 ± 1.910	30.095 ± 8.215	77.657 ± 4.000
			Adult		
5	10	10000	11.169 ± 1.156	41.418 ± 2.684	80.234 ± 2.526
		80000	17.055 ± 0.591	47.873 ± 0.716	83.802 ± 0.243
	50	10000	0.168 ± 0.148	34.396 ± 2.668	75.651 ± 3.222
		80000	2.161 ± 0.306	46.835 ± 1.060	83.111 ± 0.831
15	10	10000	1.515 ± 0.853	13.000 ± 1.970	76.005 ± 3.249
		80000	5.801 ± 0.760	22.878 ± 1.316	83.461 ± 0.822
	50	10000	0.001 ± 0.003	8.797 ± 5.715	75.077 ± 2.638
		80000	0.044 ± 0.036	21.498 ± 0.667	82.185 ± 0.908

instance-level test set. For example, on synthetic random bags with $q = 5, t = 50$ and $s = 5000$, an accuracy of just 21.3% on composite bags yields an accuracy of 85.5% on original bags and 96.4% on the test set. On the Adult dataset, with $q = 15, t = 50$ and $s = 80000$, with accuracy of just 0.044% on composite bags, we obtain a classifier with accuracy of 21.5% on original bags and 82.2% on the test set.

2. For a given q and t, increasing the number of composite bags s improves performance across the board, consistent with our theoretical bounds.
3. The bag-level performance scores are noticeably lower on the hard bags case as compared to the random bags case, even though both are from the realizable setting.
4. Accuracy scores on composite bags decrease with increasing q or t. This is understandable since this results in increased size of composite bags, making them more difficult to satisfy.

The above observations, especially points 1 and 2, demonstrate that Algorithm $\mathcal{A}_2$ does indeed provide a way to use weak classifiers on composite bags to obtain strong classifiers on original bags, which in turn are strong classifiers at the instance-level. The scalability of our techniques is also validated by the experiments on the substantially sized Adult dataset. Each of these experiments on a standard GPU/CPU took less than 12 hrs, and most completed within an hour[3]. For each dataset, the original bags were fixed, and composite bags were sampled for each repeated run of the experiment. For the synthetic, Heart, and Australian datasets, the model was trained for 160 epochs, while for the Adult dataset, it was trained for 60 epochs. Each experiment was run on a single NVIDIA A100 40GB GPU and 2x Intel Broadwell 22 cores 44 threads CPU. In Appendix E, we include additional experiments for training on the original bags.

6 CONCLUSION

In conclusion, our study is the first to demonstrate the impossibility of boosting weak classifiers to a strong classifier in the LLP and MIL settings. For LLP our work rules out boosting using weak classifiers of any accuracy < 1, while for MIL the possibility of boosting weak classifiers with accuracy $< 2/3$ is eliminated. Complementing these findings in the LLP context, we propose an algorithm that converts a weak classifier for composite bags into a strong classifier for an input collection of original bags. The algorithm constructs unions of constantly many original bags to achieve error amplification. A more efficient sampling based version of the same provides high probability guarantees, which we also experimentally validate on three real and two synthetic datasets. Future work includes ruling out boosting for MIL using weak classifiers with accuracy in $[2/3, 1)$. A related question remains on how to effectively obtain a strong classifier in the MIL setting on using weak classifiers on composite bags.

[3]The experimental code for the paper is available at `https://github.com/google-deepmind/wtos_agglabels_uai25`

References

M. Anthony and P.L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999. ISBN 9780521573535. URL https://books.google.co.in/books?id=UH6XRoEQ4h8C.

Peter Auer and Ronald Ortner. A boosting approach to multiple instance learning. In *Machine Learning: ECML 2004*, pages 63–74, 2004.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Avrim Blum and Adam Kalai. A note on learning from multiple-instance examples. *Machine learning*, 30:23–29, 1998.

Anand Paresh Brahmbhatt, Rishi Saket, and Aravindan Raghuveer. PAC learning linear thresholds from label proportions. In *Proc. NeurIPS*, 2023. URL https://openreview.net/forum?id=5Gw9YkJkFF.

Robert Istvan Busa-Fekete, Heejin Choi, Travis Dick, Claudio Gentile, and Andres Munoz medina. Easy learning from label proportions. *arXiv*, 2023. URL https://arxiv.org/abs/2302.03115.

L. Chen, Z. Huang, and R. Ramakrishnan. Cost-based labeling of groups of mass spectra. In *Proc. ACM SIGMOD International Conference on Management of Data*, pages 167–178, 2004.

Lin Chen, Thomas Fu, Amin Karbasi, and Vahab Mirrokni. Learning from aggregated data: Curated bags versus random bags. *arXiv*, 2023. URL https://arxiv.org/abs/2305.09557.

S. Chen, B. Liu, M. Qian, and C. Zhang. Kernel k-means based framework for aggregate outputs classification. In *Proc. ICDM*, pages 356–361, 2009.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

N. de Freitas and H. Kück. Learning about individuals from group statistics. In *Proc. UAI*, pages 332–339, 2005.

L. M. Dery, B. Nachman, F. Rubbo, and A. Schwartzman. Weakly supervised classification in high energy physics. *Journal of High Energy Physics*, 2017(5):1–11, 2017.

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

G. Dulac-Arnold, N. Zeghidour, M. Cuturi, L. Beyer, and J. P. Vert. Deep multi-class learning from label proportions. *CoRR*, abs/1905.12909, 2019. URL http://arxiv.org/abs/1905.12909.

Paul Erdös. On a lemma of littlewood and offord. *Bulletin of the American Mathematical Society*, 51:898–902, 1945. URL https://api.semanticscholar.org/CorpusID:122046405.

Uriel Feige and Gideon Schechtman. On the optimality of the random hyperplane rounding technique for max cut. *Random Structures & Algorithms*, 20 (3):403–440, 2002. doi: https://doi.org/10.1002/rsa.10036. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/rsa.10036.

Yoav Freund. Boosting a weak learning algorithm by majority. In *Proc. COLT*, pages 202–216, 1990.

Yoav Freund. An adaptive version of the boost by majority algorithm. *Machine Learning*, 43(3):293–318, Jun 2001. ISSN 1573-0565. doi: 10.1023/A:1010852229904. URL https://doi.org/10.1023/A:1010852229904.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Proc. EuroCOLT*, volume 904 of *Lecture Notes in Computer Science*, pages 23–37. Springer, 1995. URL https://doi.org/10.1007/3-540-59119-2_166.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28(2):337 – 407, 2000.

J. Hernández-González, I. Inza, and J. A. Lozano. Learning bayesian network classifiers from label proportions. *Pattern Recognit.*, 46(12):3425–3440, 2013.

J. Hernández-González, I. Inza, L. Crisol-Ortíz, M. A. Guembe, M. J. Iñarra, and J. A. Lozano. Fitting the data from embryo implantation prediction: Learning from label proportions. *Statistical methods in medical research*, 27(4):1056–1066, 2018.

Steinbrunn-William Pfisterer Matthias Janosi, Andras and Robert Detrano. Heart Disease. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C52P4X.

D. Kotzias, M. Denil, N. de Freitas, and P. Smyth. From group to individual labels using deep features. In *Proc. SIGKDD*, pages 597–606, 2015.

Jiantao Lai, Yanshan Xiao, and Bo Liu. Boost two-view learning-based method for label proportions problem. *Applied Intelligence*, 53(19):21984–22001, Oct 2023. ISSN 1573-7497. doi: 10.1007/s10489-023-04643-z. URL https://doi.org/10.1007/s10489-023-04643-z.

J. Liu, B. Wang, Z. Qi, Y. Tian, and Y. Shi. Learning from label proportions with generative adversarial networks. In *Proc. NeurIPS*, pages 7167–7177, 2019.

T. Lozano-Pérez and C. Yang. Image database retrieval with multiple-instance learning techniques. In *Proc. ICDE*, page 233, 2000.

O. Maron. *Learning from ambiguity*. PhD thesis, Massassachusetts Institute of Technology, 1998.

Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. NIPS'97, page 570–576, 1997.

Llew Mason, Jonathan Baxter, Peter L. Bartlett, and Marcus R. Frean. Boosting algorithms as gradient descent. In *Proc. NIPS*, pages 512–518, 1999.

Ibomoiye Domor Mienye and Yanxia Sun. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access*, 10:99129–99149, 2022.

D. R. Musicant, J. M. Christensen, and J. F. Olson. Supervised learning by training on aggregate outputs. In *Proc. ICDM*, pages 252–261. IEEE Computer Society, 2007.

J. Nandy, R. Saket, P. Jain, J. Chauhan, B. Ravindran, and A. Raghuveer. Domain-agnostic contrastive representations for learning from label proportions. In *Proc. CIKM*, pages 1542–1551, 2022.

Conor O'Brien, Arvind Thiagarajan, Sourav Das, Rafael Barreto, Chetan Verma, Tim Hsu, James Neufeld, and Jonathan J. Hunt. Challenges and approaches to privacy preserving post-click conversion prediction. *CoRR*, abs/2201.12666, 2022. URL https://arxiv.org/abs/2201.12666.

G. Patrini, R. Nock, T. S. Caetano, and P. Rivera. (almost) no label no cry. In *Proc. Advances in Neural Information Processing Systems*, pages 190–198, 2014.

Zhiquan Qi, Fan Meng, Yingjie Tian, Lingfeng Niu, Yong Shi, and Peng Zhang. Adaboost-LLP: A boosting method for learning with label proportions. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8):3548–3559, 2018.

N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le. Estimating labels from label proportions. *J. Mach. Learn. Res.*, 10:2349–2374, 2009.

Ross Quinlan. Statlog (Australian Credit Approval). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C59012.

Jan Ramon and Luc De Raedt. Multi instance neural networks. In *Proceedings of the ICML-2000 workshop on attribute-value and relational learning*, pages 53–60, 2000.

Soumya Ray and Mark Craven. Supervised versus multiple instance learning: an empirical comparison. In *Proc. ICML*, page 697–704, 2005.

S. Rueping. SVM classifier estimation from group probabilities. In *Proc. ICML*, pages 911–918, 2010.

Sivan Sabato and Naftali Tishby. Multi-instance learning with any hypothesis class. *Journal of Machine Learning Research*, 13(97):2999–3039, 2012. URL http://jmlr.org/papers/v13/sabato12a.html.

R. Saket. Learnability of linear thresholds from label proportions. In *Proc. NeurIPS*, 2021. URL https://openreview.net/forum?id=5BnaKeEwuYk.

R. Saket. Algorithms and hardness for learning linear thresholds from label proportions. In *Proc. NeurIPS*, 2022. URL https://openreview.net/forum?id=4LZo68TuF-4.

Rishi Saket, Aravindan Raghuveer, and Balaraman Ravindran. On combining bags to better learn from label proportions. In *AISTATS*, volume 151 of *Proceedings of Machine Learning Research*, pages 5913–5927. PMLR, 2022. URL https://proceedings.mlr.press/v151/saket22a.html.

Robert E. Schapire. The strength of weak learnability (extended abstract). In *Proc. FOCS*, pages 28–33, 1989.

Robert E. Schapire and Yoav Freund. *Boosting: Foundations and Algorithms*. The MIT Press, 2012. ISBN 0262017180.

Karan Sikka, Abhinav Dhall, and Marian Bartlett. Weakly supervised pain localization using multiple instance learning. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2013.

M. Stolpe and K. Morik. Learning from label proportions by optimizing cluster model selection. In *ECML PKDD Proceedings, Part III*, volume 6913, pages 349–364. Springer, 2011.

Paul Viola, John C. Platt, and Cha Zhang. Multiple instance boosting for object detection. In *NIPS*, page 1417–1424, 2005.

Manfred K. Warmuth, Karen A. Glocer, and S. V. N. Vishwanathan. Entropy regularized lpboost. In Yoav Freund, László Györfi, György Turán, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, pages 256–271, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-87987-9.

J. Wu, Yinan Yu, Chang Huang, and Kai Yu. Deep multiple instance learning for image classification and auto-annotation. In *Proc. CVPR*, pages 3460–3469, 2015.

F. X. Yu, D. Liu, S. Kumar, T. Jebara, and S. F. Chang. $\propto$SVM for learning with label proportions. In *Proc. ICML*, volume 28, pages 504–512, 2013.

F. X. Yu, K. Choromanski, S. Kumar, T. Jebara, and S. F. Chang. On learning from label proportions. *CoRR*, abs/1402.5902, 2014. URL `http://arxiv.org/abs/1402.5902`.

Cha Zhang, John Platt, and Paul Viola. Multiple instance boosting for object detection. In *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005.

J. Zhang, Y. Wang, and C. Scott. Learning from label proportions by learning with label noise. In *Proc. NeurIPS*, 2022.

Qi Zhang and Sally Goldman. Em-dd: An improved multiple-instance learning technique. *Advances in neural information processing systems*, 14, 2001.

A TRIVIAL ACCURACY IN THE LLP AND MIL

First consider the LLP bags $\mathcal{B}$ from Theorem 1.1, each bag is of size 2 with aggregate label 1 i.e., it is satisfied if exactly one of its feature-vectors is labeled 1. Now consider just one bag from $\mathcal{B}$. This bag is not satisfied by the constant 0 or constant 1 classifier. On the other hand the expected accuracy of random labeling is $1/2$, and therefore $\mathsf{Trv}_{\mathsf{LLP}}(\mathcal{B}) = 1/2$.

Next, let $\mathcal{B}$ be the MIL bags from Theorem 1.2. These are bags of size 2 each and some of them have aggregate label 0 and some have aggregate label 1. Consider just two bags one with aggregate label 0 and the other with aggregate label 1. Now, the constant a labeling satisfies the bag with aggregate label a and does not satisfy the bag with aggregate label $(1-a)$, for $a \in \{0, 1\}$. On the other hand the random labeling satisfies the 0 aggregate label bag with probability $1/4$ and the bag with aggregate label 1 with probability $3/4$. Thus, the expected number of bags satisfied is the random labeling is 1. Therefore, $\mathsf{Trv}_{\mathsf{MIL}}(\mathcal{B}) = 1/2$.

A.1 INAPPLICABILITY OF BAG COMPOSITION FOR WEAK TO STRONG LEARNING IN MIL

Suppose we are given a classifier on the original small bags with accuracy bounded away from 1. In LLP, when we form composite bags, each as a union of several randomly chosen original bags, we obtain an erroneous prediction on most of the composite bags thus formed. Our LLP algorithm uses this error-gap amplification. However, in MIL, the union of several randomly chosen original bags will give a bag-label of 1 even if one of the constituent original bags has bag-label 1. This would happen with high probability when a significant number of original bags have bag-label 1. Thus, taking large unions would result in most bags having bag-label 1 and therefore the constant predictor will have a high accuracy, even if it has low accuracy on the original bags.

A.2 EXTENSION OF OUR RESULTS TO MULTI-CLASS CLASSIFICATION

Our results on the impossibility of boosting (Theorems 1.1, 1.2) rule out boosting in LLP and MIL for binary classification and since this is a special case of the multi-class setting, they also rule out boosting in the multi-class setting. Our algorithmic results (Theorem 1.3) are also for binary classification. However, they can be extended to multi-class classification. For this we can define LLP in the multi-class setting, where the bag-label is a histogram over label-set, and a bag is satisfied if the predicted label histogram matches its bag-label. The algorithms will be the same, up to change in parametric dependencies on the label-set size. The application of Lemma 2.2 can be done separately for each label along with a union bound over the error probability.

B IMPOSSIBILITY OF BOOSTING IN MIL

Along similar lines as the previous section, we provide a geometric construction of MIL on 2-sized bags. We begin with a continuous set of points which we analyze and subsequently discretize while preserving its key properties. We fix a parameter $\alpha \in (1/2, 1)$.

Construction. Let $\boldsymbol{\mathcal{X}}_c$ be set of all points on the unit circle $\mathbb{S}^1$. For any two points that subtend an angle of exactly $\alpha\pi$ we create a 2-sized bag with aggregate label 1 (we call it a 1-bag) containing those points. Similarly, bags with aggregate label 0 (which we call 0-bags) are formed by pairs of points at an angle of $(1-\alpha)\pi$. By mapping a 1-bag to the mid-point of the smaller arc subtended by the two points in the bag (end-points), and noting that all the 1-bags have unique mid-points, we obtain that the measure of the set of 1-bags is same as that of $\mathbb{S}^1$. Similarly, this holds true for the set of 0-bags. To construct a measure, define the following bag-sampling procedure: sample a uniform point on the unit circle and randomly output either the unique 1-bag corresponding to it with probability $1/2$ or the unique 0-bag corresponding to it with probability $1/2$. In particular, the set of 0-bags and the set of 1-bags are of equal measure. Let $\mathcal{B}_c$ be this infinite (continuous) collection of 1-bags and 0-bags.

Existence of Weak Classifier. Observe that the constant 0 classifier given by $\mathsf{pos}(-1)$ will satisfy all 0-bags and none of the 1-bags.

Now, consider a random homogeneous halfspace given by $\mathsf{pos}(\mathbf{r}^\mathsf{T}\mathbf{x})$ for $\mathbf{r}$ uniformly sampled from $\mathbb{S}^1$. The two points of a 0-bag will not be separated w.p. α and conditioned on this, with probability $1/2$ both will be assigned 0, implying that any 0-bag will be satisfied with probability $\alpha/2$. On the other hand, both the points of a 1-bag will be assigned 0 w.p. $(1-\alpha)/2$ implying that it will be satisfied w.p. $(1+\alpha)/2$.

Let there be any probability measure on $\mathcal{B}_c$ s.t. the measure of the 0-bags is p and that of the 1-bags is $(1-p)$. If $p \geq 2/3$ then the constant 0 classifier satisfies all the 0-bags yielding an accuracy of $p \geq 2/3$. If not, then the random homogeneous halfspace satisfies in expectation

$$\begin{aligned} p\alpha/2 + (1-p)(1+\alpha)/2 &= (1+\alpha)/2 - p/2 \\ &\geq 1/2 + \alpha/2 - 1/3 \\ &= 2/3 - (1-\alpha)/2 \end{aligned} \tag{5}$$

Therefore, there is always a weak classifier, for any reweighing of the bags, of accuracy $2/3 - (1-\alpha)/2$.

No Strong Classifier. Consider any $\{0,1\}$-labeling of $\mathbb{S}^1$, where the subset labeled 1 is measurable. Let $z \in [0,1]$ represent the fraction of points on $\mathbb{S}^1$ labeled as 1, with the remaining fraction $1-z$ labeled as 0. Sampling a 0-bag uniformly at random (u.a.r.) and randomly choosing one of its points yields the uniform distribution over $\mathbb{S}^1$. Thus, the probability that a random 0-bag is satisfied is $\leq 1-z$. Each point in $\mathbb{S}^1$ is an element of exactly two distinct 1-bags, so the probability that in a random 1-bag at least one of its points is labeled 1 is at most $\min\{2z, 1\}$.

Therefore, the probability that a random bag from $\mathcal{B}_c$ is satisfied by the labeling is at most

$$\frac{1 - z + \min\{2z, 1\}}{2} = \begin{cases} 1 - z/2 & \text{if } z \geq 1/2 \\ 1/2 + z/2 & \text{otherwise} \end{cases} \tag{6}$$

which attains a maximum of $3/4$ at $z = 1/2$. Thus, no classifier can have accuracy $> 3/4$ on $\mathcal{B}_c$

Discretization. Let T be a large positive integer, and divide $\mathbb{S}^1$ into $2T$ continuous, non-overlapping arcs $\{A_i\}_{i=1}^{2T}$ of length $\delta\pi$ each, where $\delta = 1/T$. We choose T large enough so that $2\delta < \min\{(2\alpha - 1), (1-\alpha)\}$, ensuring that:
(i) there is no segment that contains both endpoints of any bag in $\mathcal{B}_c$, and
(ii) for any pair of segments A_i and A_j, if there is a 0-bag in $\mathcal{B}_c$ with one point in A_i and another in A_j, then there is no such 1-bag, and similarly if there is a 1-bag in $\mathcal{B}_c$ with one point in A_i and another in A_j, then there is no such 0-bag.

Using property (ii) above, let us construct a discrete set of bags $\mathcal{B}_d$ as follows. If a pair of segments A_i and A_j are such that there is a 0-bag in $\mathcal{B}_c$ with one point in A_i and another in A_j, then add $\{A_i, A_j\}$ as 0-bag with weight as the measure of all the bags in $\mathcal{B}_c$ (which are necessarily 0-bags) with one point in A_i and another in A_j. Analogously, add pairs of segments as 1-bags. Note that from property (i), all bags in $\mathcal{B}_d$ have size 2.

No Strong Classifier. Let us first consider any $\{0,1\}$-labeling to $\{A_i\}_{i=1}^{2T}$. This directly corresponds to a $\{0,1\}$-labeling to $\mathbb{S}^1$ by assigning a point the label of the segment containing it. Further, from its construction, the weight of the bags $\mathcal{B}_c$ satisfied by the labeling to the segments equals the measure of the bags in $\mathcal{B}_c$ satisfied by the corresponding labeling to $\mathbb{S}^1$ which, as shown above, is at most $3/4$.

In particular, the above argument also shows that the measure of bags in $\mathcal{B}_d$ satisfied by the constant 0 labeling to $\{A_i\}_{i=1}^{2T}$ is the same as that in $\mathcal{B}_c$ satisfied by the constant 0 labeling to $\mathbb{S}^1$.

Weak Classifier. Lastly, we translate the labeling by a homogeneous halfspace on $\mathbb{S}^1$ to a labeling for $\{A_i\}_{i=1}^{2T}$ by assigning each A_i the label of its mid-point. Consider the *error* set of points in $\mathbb{S}^1$ whose label given by the homogeneous halfspace differs from the label of the segment containing it. For any homogeneous halfspace, the error set is entirely contained within the two diametrically opposite segments intersected by the halfspace. Similarly, the *error* bags in $\mathcal{B}_c$ are those whose aggregate label given by the homogeneous halfspace differs from the aggregate label of the corresponding bag in $\mathcal{B}_d$.

The *error* bags in $\mathcal{B}_c$ are a subset of those which have at least one end-point in the the error set of points. Given any bag in $\mathcal{B}_c$ the probability over a random homogeneous halfspace that it is an error bag is at most the probability that one of its endpoints is in a segment intersected by the halfspace. By symmetry, a segment is intersected with probability $1/T$. So the probability that any bag in $\mathcal{B}_c$ is an error bag is at most $2/T = 2\delta$.

Thus, from (5) we obtain that for any weighing of the bags in $\mathcal{B}_d$, there is a classifier of accuracy $2/3 - (1-\alpha)/2 - 2\delta$.

B.1 COMPLETING THE PROOF OF THEOREM 1.2.

For this, we can take ε to be small enough, say $\varepsilon \in (0, 0.1)$ and set $\alpha = 1 - \varepsilon$ along with $T = \lceil 4/\varepsilon \rceil$ so that $\delta \leq \varepsilon/4$ and $2\delta < \min\{(2\alpha - 1), (1-\alpha)\}$ and $2/3 - (1-\alpha)/2 - 2\delta \geq 2/3 - \varepsilon$.

C WEIGHTED BAGS TO UNWEIGHTED BAGS

Input: : Bags $\mathcal{B}_w = (B_i, w_i)_{i=1}^m, T$.

Steps:
1. Normalize the weight with a factor Z such that $\sum_{i=1}^m w_i = m$.
2. Define $\mathcal{B}$ to be the unweighted collection of bags and initialize it to $\emptyset$.
3. for $i \in [m]$:
  3.1 Define $n_i = \lceil w_i(T-1) \rceil$.
  3.2 Add n_i copies of B_i to $\mathcal{B}$.

Output: Output $\mathcal{B}$.

Figure 7: Weighted to unweighted collection of bags

The algorithm to convert a weighted collection of bags to an unweighted collection is given in Fig. 7. First, observe that $|\mathcal{B}| = \sum_{i=1}^m \lceil w_i(T-1) \rceil \leq \sum_{i=1}^m (w_i(T-1)+1) \leq (T-1)m + m = Tm$, where we use $\sum_{i=1}^m w_i = m$. On the other hand, $|\mathcal{B}| = \sum_{i=1}^m \lceil w_i(T-1) \rceil \geq (T-1)m$.

Now, to see that the error in accuracy is at most $O(1/T)$, observe that for any subset $I \subseteq [m]$, $\sum_{i \in I} w_i(T-1) \leq \sum_{i \in I} \lceil w_i(T-1) \rceil \leq \sum_{i \in I} w_i(T-1) + |I|$. Therefore, the normalized error in the weight corresponding to I is at most $|I|/((T-1)m) \leq m/((T-1)m) \leq 1/(T-1) = O(1/T)$ for $T > 1$.

D PROBABILITIES FOR THE SUPPORT OF $\overline{D}$

In Step 2 of Figure 3, the for a fixed configuration $\{\mathcal{Q}_i\}_{i=1}^t$ with $r : |\{i \in [t] \mid \mathcal{Q}_i = \star\}|$, its probability under $\overline{D}$ is $\frac{m^r}{m^t}\frac{1}{2^r}$, since the number of choices for the $\star$-coordinates is m^r, while the total number of choices is m^t. Further, with $(1/2)^t$ probability we have the specific choices of the r coordinates with $\star$ in Step 2. Iterating over all possible configurations $\{\mathcal{Q}_i\}_{i=1}^t$ and assigning theur probabilities to the resultant $(\overline{B}, \overline{\sigma})$ in Step 3, yields the support of $\overline{D}$ along with their probabilities.

E ADDITIONAL EXPERIMENTS

In Table 3, we report results obtained by training directly on the original small bags. When comparing these results with those in Tables 1 and 2, we find that training directly on original bags yields better accuracy on the test sets of original bags and individual instances for the Heart dataset and comparable performance on the Australian and Adult datasets. For both Synthetic datasets, however, the strong classifier obtained using our proposed algorithm achieves better performance on original bags compared to direct training.

q	Train Bags	Test Instances
	Heart	
5	46.370 ± 5.871	82.578 ± 4.767
15	31.111 ± 12.258	74.844 ± 5.960
	Australian	
5	55.064 ± 6.881	84.196 ± 3.815
15	26.190 ± 6.000	77.121 ± 5.776
	Adult	
5	47.368 ± 0.650	83.539 ± 0.588
15	12.899 ± 1.762	80.119 ± 2.172
	Synthetic Random	
5	81.783 ± 2.369	95.627 ± 0.736
15	42.400 ± 3.795	88.535 ± 2.357
	Synthetic Hard	
5	74.546 ± 6.029	92.170 ± 2.506
15	32.313 ± 4.112	82.592 ± 5.045

Table 3: Results after training directly on orginal (small) bags.