Journal of Machine Learning Research 17 (2016) 1-50 Submitted 4/15; Revised 3/16; Published 8/16

Multiple-Instance Learning from Distributions

Gary Doran GARY.DORAN@CASE.EDU

Soumya Ray SRAY@CASE.EDU

Department of Electrical Engineering and Computer Science
Case Western Reserve University
10900 Euclid Ave, Glennan 320
Cleveland, OH 44106, USA

Editor: Luc De Raedt

Abstract

We propose a new theoretical framework for analyzing the multiple-instance learning (MIL) setting. In MIL, training examples are provided to a learning algorithm in the form of labeled sets, or "bags," of instances. Applications of MIL include 3-D quantitative structure–activity relationship prediction for drug discovery and content-based image retrieval for web search. The goal of an algorithm is to learn a function that correctly labels new bags or a function that correctly labels new instances. We propose that bags should be treated as latent distributions from which samples are observed. We show that it is possible to learn accurate instance- and bag-labeling functions in this setting as well as functions that correctly rank bags or instances under weak assumptions. Additionally, our theoretical results suggest that it is possible to learn to rank efficiently using traditional, well-studied "supervised" learning approaches. We perform an extensive empirical evaluation that supports the theoretical predictions entailed by the new framework. The proposed theoretical framework leads to a better understanding of the relationship between the MI and standard supervised learning settings, and it provides new methods for learning from MI data that are more accurate, more efficient, and have better understood theoretical properties than existing MI-specific algorithms.

Keywords: multiple-instance learning, learning theory, ranking, classification

1. Introduction

The standard supervised learning setting, in which labeled training examples are represented with individual feature vectors, is well-studied with numerous applications. However, there remain many compelling real-world problems that require learning from more structured data. For example, in text categorization, a document might contain a set of passages or paragraphs. A typical approach is to simply ignore the internal structure within the document by treating it as a "bag of words," then to represent the document with a single feature vector based on word frequencies. However, because such an approach destroys the internal structure of the document, it becomes challenging to determine *which* passages or paragraphs correspond to the category of interest. Similarly, for the content-based image retrieval (CBIR) domain, the goal is to learn to retrieve images that contain some object of interest to the user. One approach is to use a flat feature vector representation of each image given pixel color values. Again, using such an approach, it is not clear how one might identify *which* object in or region of the image was of interest to a user.

For such problems, the multiple-instance (MI) setting offers a richer representation for structure objects as sets, or "bags," of feature vectors, each of which is called an "instance" (Dietterich et al., 1997). In the text categorization example above, a document is a bag of passages or paragraphs, which are the instances. For CBIR, an image is a bag of segments or objects. The MI setting further assumes that labels exist at both the level of instances and bags, where a bag's label is the logical conjunction of Boolean instance labels. That is, a bag is positive if *at least one* instance in the bag is positive and negative if *all* of the instances in the bag are negative. This logical relationship corresponds to the fact that a document or an image is of the class of interest if and only if at least one of the passages or objects it contains is of the class of interest.

In the standard supervised setting, there is typically only one target concept of interest. For MI learning, one might be interested in learning either a bag or an instance concept from MI data. For example, in the 3-Dimensional Quantitative Structure–Activity Relationship (3D-QSAR) domain, the goal is to learn to predict whether a molecule will bind to a given target receptor (Dietterich et al., 1997). Because a molecule has flexible bonds, it exists in multiple shapes, or *conformations*, in solution. Thus, mapping this problem into the MI setting, conformations are instances and molecules are bags. A *bag-labeling* function can be used to predict whether a given molecule will bind to a target receptor. On the other hand, an *instance-labeling* function can be used to predict which specific conformations will bind to a receptor, providing useful, difficult-to-measure information about the receptor's physical structure.

Despite the importance of these two learning tasks, only the bag-labeling task has received much attention in recent prior work characterizing the learnability of MI concepts (Sabato and Tishby, 2012). When learnability of instance concepts has been addressed, it has been under the strict, unrealistic assumption that instances across all bags are independent and identically distributed (IID) samples *from the same underlying distribution* (Blum and Kalai, 1998). However, as far as we know, the result of Blum and Kalai (1998) has remained the only positive result on instance concept learnability in the MI setting for over a decade. In this paper, we describe new *positive* results for both instance- and bag-concept learnability. Our contributions are summarized as follows:

1. We describe a new generative model for MI data and show that it subsumes some previously proposed generative models for MI learning (MIL).

2. We provide novel results for learning accurate bag-level concepts from MI data.

3. We describe the first positive instance concept learnability results since those of Blum and Kalai (1998).[1]

4. We prove the first results, to our knowledge, that formally describe the ability to rank both instances and bags in the MI setting.

5. We empirically evaluate a surprising implication of our theoretical results: that *standard supervised approaches* can effectively rank both instances and bags in the MI

[1] These results were also presented in Doran and Ray (2014).

setting. Our evaluation uses 55 data sets from a wide variety of domains, and supports both our theoretical results as well as the assumptions made by our generative model.

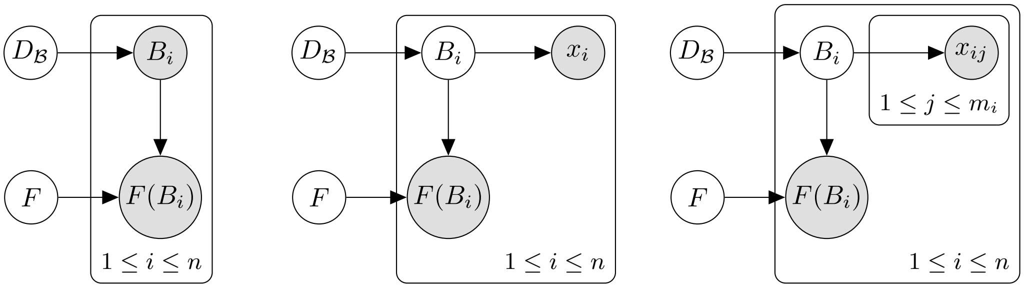
2. Bags as Distributions

In this section, we describe a generative model for MI data in which bags are viewed as *distributions over instances* rather than as sets of instances. We show that the proposed generative model actually encompasses previous, standard models of MI learning in which bags are sets or tuples. The choice of framing a problem within a particular theoretical model has significant practical consequences for designing or selecting an algorithm to solve the problem. This section provides a theoretical framework in which the MI classification problem can be analyzed. The model allows us to derive positive instance- and bag-concept learnability results for the MI setting as described in Section 3. Furthermore, as Section 4 shows, the generative model leads to a surprising yet testable hypothesis that standard supervised algorithms can learn from MI data. This hypothesis is evaluated experimentally, supporting the assumptions made by the model.

2.1 The Generative Model

At the heart of this work is the claim that bags are best viewed as distributions rather than as finite sets of instances. Below, we formally define what we mean by this statement. But first, the example domain of drug activity prediction provides an intuitive justification for this claim. As described in Section 1, in the drug activity prediction domain, the goal is to predict the ability of molecules to activate, or bind to, a receptor. To cast the problem as binary classification, we select some threshold so that each molecule's activity level either corresponds to an "active" or "inactive" label. In this case, we can think of each molecule (bag) as being drawn from a distribution $D_{\mathcal{B}}$ over molecules. Ignoring for the moment that each molecule has numerous conformations, this molecule either activates the receptor or not, so in nature the labeling function is defined at the level of bags. Prior models represent each molecule as a set or multiset of conformations, so they implicitly assume that each molecule exists in only a finite number of conformations. In reality, a molecule can transform continuously from conformation to conformation, producing an infinite set of conformations. In particular, each molecule exists in a state of dynamic equilibrium in which the amount of time it spends in each conformation is distributed according to Gibbs free energy such that low-energy conformations are preferred. Hence, the molecule (bag) corresponds to a *distribution over instances.* Constructing a bag from low-energy conformations, the common procedure for constructing bags in the drug activity domain, can be thought of as sampling instances from this distribution. Note that each molecule will have a *unique* distribution over conformations; thus, prior generative models for MIL that assume all instances are drawn from the *same* distribution are not applicable (Blum and Kalai, 1998). In Section 2.5, we describe how this view of the MI generative process can be applied to other problem domains.

More abstractly, our generative process needs several components. First, *bags are distributions over instances*, and we will assume that these bags are sampled from a *distribution over bags* and labeled according to a *bag-labeling function.* In addition to the bag-labeling



(a) Generative Process for Bags

(b) Generative Process for Bag-labeled Instances

(c) Generative Process for Bag Samples

Figure 1: A comparison of the generative processes for bags, individual bag-labeled instances, and bag samples.

function, in the MI setting there is also an *instance-labeling* function with the standard *MI assumption* relating the bag- and instance-labeling functions. We describe the *instance distribution* that is consistent with the generative process, which will be useful for discussing instance concept learnability. In addition to these essential components, we introduce two *additional weak assumptions* that make efficient learning in this setting possible.

Bags as Distributions. To formalize the intuition above, suppose we have an instance space $\mathcal{X}$. Typically, the space of bags is some subset of $\mathcal{X}^*$, the set of all finite subsets of $\mathcal{X}$. However, here, we let the space of bags be $\mathcal{B} = \mathscr{P}(\mathcal{X})$, the set of probability distributions on the input space. Hence, each bag $B \in \mathcal{B}$ is a probability distribution over instances, denoted $\mathrm{P}(x \mid B)$.

Bag Distribution. We propose that, at the level of bags, the MI generative process is similar to that for supervised learning. In particular, bags are sampled from some fixed distribution $D_{\mathcal{B}}$, which is a distribution over instance distributions ($D_{\mathcal{B}} \in \mathscr{P}(\mathscr{P}(\mathcal{X}))$). From this distribution $D_{\mathcal{B}}$, we sample some set of bags $\{B_i\}_{i=1}^n$, as illustrated by the plate model in Figure 1(a).

Bag-Labeling Function. As in supervised learning, we assume that there exists some labeling function $F : \mathcal{B} \to \{0, 1\}$ that labels bags. Thus, a supervised data set $\{(B_i, F(B_i))\}_{i=1}^n$ could be produced by sampling bags IID from $D_{\mathcal{B}}$ and applying the labeling function F.

Instance-Labeling Function. In the MI setting, we assume that in addition to the bag-labeling function F, there also exists an *instance-labeling* function $f : \mathcal{X} \to \{0, 1\}$. A key component of the MI setting is not only the existence of both bag and instance labeling functions, but the relationship between the two as well. Traditionally, the MI assumption is stated with respect to particular sets of instances so that a bag label $F(B_i)$ is the logical `OR` (for boolean labels), or maximum (for numerical labels), of its instances' labels: $F(B_i) = \max_j f(x_{ij})$. However, in the proposed generative model, bags are distributions with *a priori* labels regardless of the instances sampled from them. Therefore, our generative model requires a more nuanced description of the relationship between bag and instance labels.

The MI Assumption. We state the relationship between F and f at the level of the generative model. Accordingly, a bag is negative ($F(B) = 0$) if and only if probability of sampling a positive instance within the bag is zero: $\mathrm{P}_{x \sim B}\left[f(x) = 1\right] = 0$. In measure theoretic terms, instances sampled within negative bags are almost surely negative, which implies that positive instances are almost surely sampled only within positive bags. This condition corresponds to the standard MI assumption that negative bags contain only negative instances.

Instance Distribution. In order to talk about the learnability of f, we must define some instance distribution with respect to which we will measure risk. An instance distribution naturally arises from our generative model if we first sample a bag B randomly from $D_{\mathcal{B}}$, then sample an instance x randomly from the distribution corresponding to B. The instance distribution $D_{\mathcal{X}}$ resulting from this two-level sampling procedure is effectively the distribution that marginalizes out the individual bag distributions. That is, given a probability distribution $\mathrm{P}_{\mathcal{B}}$ over bags corresponding to $D_{\mathcal{B}}$, we can define a distribution $\mathrm{P}_{\mathcal{X}}$ corresponding to $D_{\mathcal{X}}$ as

$$\mathrm{P}_{\mathcal{X}}(x) = \int_{\mathcal{B}} \mathrm{P}(x \mid B)\, \mathrm{d}\mathrm{P}_{\mathcal{B}}(B). \tag{1}$$

Given that "x" is used to denote instances and "B" is used to denote bags, we subsequently drop subscripts from P when the sample space can be inferred from context. As we discuss in Section 3.4, the ability to marginalize out bag-specific distributions in our model plays a vital role in proving the learnability of instance- and bag-labeling functions. Given a bag distribution, the existence of such an instance distribution is guaranteed under relatively weak assumptions on the instance space $\mathcal{X}$ (Diestel and Uhl, 1977). Furthermore, note that while we can view instances in our generative model as being sampled IID from $D_{\mathcal{X}}$, this does not require the assumption that instances are IID across all bag distributions, as in prior generative models for MIL (Blum and Kalai, 1998). We discuss this point in detail in Section 3.6.

Additional Assumptions. As is the case in the standard MI framework, in our generative model, only bag labels are observed. Suppose we sample individual instances as illustrated in Figure 1(b) where we first sample a bag, record its label, and then sample an instance from the bag-specific distribution $\mathrm{P}(x \mid B)$ and assign the bag label to the instance. Then the resulting bag-labeled instances $\{(x_{i1}, F(B_i))\}_{i=1}^{n}$ are distributed according to $D_{\mathcal{X}}$, and will appear in positive bags some of the time and negative bags the remaining fraction of the time. Therefore, each instance will have some probability $c(x) \in [0, 1]$ of appearing with a positive label, which can be formally expressed as a probabilistic concept (p-concept) like the kind described by Kearns and Schapire (1994):

$$c(x) \triangleq \mathrm{P}\left[F(B) = 1 \mid x\right]. \tag{2}$$

That is, the probability of observing a positive label for instance x is the conditional probability that the bag B in the two-level sampling procedure was positive, given that x was observed within B. This conditional probability can be derived from the joint distribution over instances and bag labels corresponding to the generative process in Figure 1(b).

It follows from the previously-stated relationship between F and f that for any positive instance x_+, $c(x_+) = 1$, since each positive instance is observed almost surely (with probability 1) within a positive bag. In order to distinguish positive and negative instances, we make the following weak assumption: there exists some $\gamma > 0$ such that for every negative instance x_-, $c(x_-) \leq 1 - \gamma$. Intuitively, this corresponds to the assumption that every negative instance is observed with some nonzero probability in a negative bag.

To see why negative instances must appear in negative bags in order to learn a concept, consider trying to learn the instance concept "spoon" in the CBIR domain, as described in Section 1. To learn this concept, you are given a set of images containing spoons, and a set of images not containing spoons. However, suppose that in every image containing a spoon, there is also a fork nearby. Furthermore, forks never appear alone in images without spoons. In this unfortunate scenario, you have no means of determining which of the fork or spoon is the positive instance given only image-level labels. However, if there is a guarantee that eventually you will see a negative image containing a fork but not a spoon, you will be able to learn that the fork is not the positive instance. We discuss learnability further in Section 3 and Section 4.

Finally, for learning bag-level concepts, we show in Section 3.2 that we require one additional assumption that there is some minimum fraction π of positive instances in each positive bag. That is, for every positive bag B_+, $\mathrm{P}\left[f(x) = 1 \mid B_+\right] \geq \pi$. Without this assumption, there might be positive bags that only contain negative instances. However, this would make them indistinguishable during bag labeling from negative bags, which by definition only contain negative instances. Interestingly, this assumption is not required if we are only interested in learning an instance-level concept.

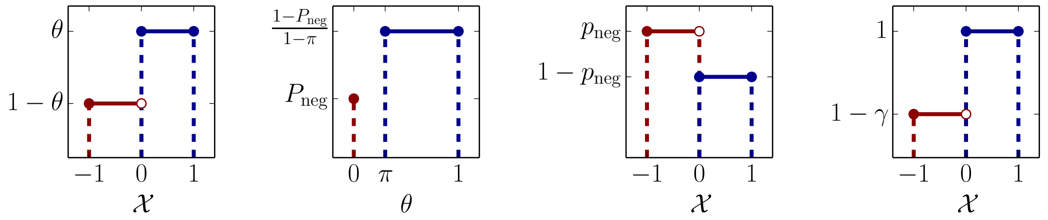
Now, we can formally define MI-GEN, the set of generative processes for MI data consistent with the assumptions described above,

Definition 1 (MI-GEN) *Given any $\gamma \in (0,1]$ and $\pi \in [0,1]$,* MI-GEN(γ, π) *is the set of all tuples $(D_{\mathcal{X}}, D_{\mathcal{B}}, f, F)$, each consisting of an instance distribution $D_{\mathcal{X}}$ (with corresponding* $\mathrm{P}(x)$*), bag distribution $D_{\mathcal{B}}$ (with corresponding* $\mathrm{P}(B)$*), instance-labeling function f, and bag-labeling function F, that satisfy the conditions:*

1. $\mathrm{P}(x) = \int_{\mathcal{B}} \mathrm{P}(x \mid B)\,\mathrm{d}\,\mathrm{P}(B)$
2. $\forall x : f(x) = 1 \implies \mathrm{P}\left[F(B) = 0 \mid x\right] = 0$
3. $\forall x : f(x) = 0 \implies \mathrm{P}\left[F(B) = 0 \mid x\right] \geq \gamma$
4. $\forall B : F(B) = 1 \implies \mathrm{P}\left[f(x) = 1 \mid B\right] \geq \pi.$

For simplicity, we will write MI-GEN(γ) for the case when $\pi = 0$, which corresponds to the weakest Condition 4. That is, for any fixed γ, MI-GEN$(\gamma) \supseteq$ MI-GEN(γ, π) for every $\pi \geq 0$. Such notation will be used when discussing instance-concept learnability, which does not require the $\pi > 0$ assumption. That is, instance-concept learning under our model is naturally tolerant to "bag label noise" of the form where positive bags contain only negative instances.

Finally, note that for any $\gamma \in (0,1]$, $\pi \in [0,1]$, MI-GEN$(\gamma, \pi) \supseteq$ MI-GEN$(1,1)$. That is, $\gamma = \pi = 1$ corresponds to the strongest constraints on the generative process. Even in this case, for *any* $D_{\mathcal{X}}$ and f, there exist $D_{\mathcal{B}}$ and F such that $(D_{\mathcal{X}}, D_{\mathcal{B}}, f, F) \in$ MI-GEN$(1,1)$. In particular, given a point mass δ_x centered on x, we can define $D_{\mathcal{B}}$ so that $\mathrm{P}_{\mathcal{B}}(\delta_x) = \mathrm{P}_{\mathcal{X}}(x)$ and F such that $F(\delta_x) = f(x)$. This choice of $(D_{\mathcal{B}}, F)$ corresponds to supervised learning



(a) $\text{P}(x \mid B_\theta)$

(b) $D_{\mathcal{B}}$, $\text{P}(B_\theta)$

(c) $D_{\mathcal{X}}$, $\text{P}(x)$

(d) $c(x)$

Figure 2: An example generative process for MI data. Each bag distribution (a) is parameterized by θ, and the distribution over bags (b) corresponds to a distribution over values of θ. The resulting distribution over instance (c) is derived in Equation 4 and Equation 5. The probability of instances appearing in positive bags (d) is derived in Equation 6 and Equation 7.

expressed in our generative model. That is, sampling from our generative process in that case is indistinguishable from sampling directly from $D_{\mathcal{X}}$ with labels assigned according to f. Below, we discuss the relationship between our generative model and other proposed models for MI learning.

2.2 An Example of the Generative Process

As a concrete example, suppose the instance space is the closed real-valued interval $\mathcal{X} = [-1, 1]$ and each bag B_θ is a distribution parameterized by a single real-valued parameter $\theta \in [0, 1]$. As illustrated in Figure 2(a), the bag distribution $\text{P}(x \mid B_\theta)$ assigns $(1-\theta)$ of the probability mass uniformly to the interval $[-1, 0)$, and θ of the mass uniformly to the interval $[0, 1]$. Each value of θ corresponds to a different bag, which is a different distribution over instances.

In this example, a distribution over bags is essentially a distribution over the bag parameter θ. Such a distribution is illustrated in Figure 2(b), and assigns P_{neg} of the mass to the set $\{0\}$ and the remaining $1 - P_{\text{neg}}$ portion of the mass uniformly to the interval $[\pi, 1]$. The probability of sampling a bag, $\text{P}(B_\theta)$, corresponds to the probability of sampling the corresponding value of θ. Similarly, a bag-labeling function F can be defined in terms of θ as follows:

$$F(B_\theta) = \begin{cases} 0 & \text{if } \theta = 0 \\ 1 & \text{if } \theta > 0. \end{cases} \tag{3}$$

Thus, for this example, $P_{\text{neg}} = \text{P}\left[F(B_\theta) = 0\right]$.

For the sake of the example, we choose the instance-labeling function to be

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

This choice is consistent with the bag-labeling function defined in Equation 3, since $F(B_\theta) = 0$ implies $\theta = 0$, which implies that the probability of sampling a positively labeled $x \in [0, 1]$ is zero as required.

If we marginalize out the bag distribution, we obtain the single instance distribution in Figure 2(c). Analytically, for any $x_- \in [-1, 0)$, we have

$$\begin{aligned} \mathrm{P}(x_-) &= \int_0^1 \mathrm{P}(x_- \mid B_\theta)\,\mathrm{P}(B_\theta)\,\mathrm{d}\theta \\ &= 1 \cdot P_{\text{neg}} + \int_\pi^1 (1-\theta)\frac{1-P_{\text{neg}}}{1-\pi}\,\mathrm{d}\theta \\ &= P_{\text{neg}} + \tfrac{1}{2}(1-P_{\text{neg}})(1-\pi) \triangleq p_{\text{neg}}. \end{aligned} \tag{4}$$

Similarly, for $x_+ \in [0, 1]$,

$$\begin{aligned} \mathrm{P}(x_+) &= \int_0^1 \mathrm{P}(x_+ \mid B_\theta)\,\mathrm{P}(B_\theta)\,\mathrm{d}\theta \\ &= 0 \cdot P_{\text{neg}} + \int_\pi^1 \theta\frac{1-P_{\text{neg}}}{1-\pi}\,\mathrm{d}\theta \\ &= \tfrac{1}{2}(1-P_{\text{neg}})(1+\pi) = 1 - p_{\text{neg}}. \end{aligned} \tag{5}$$

Since probability density functions exist for this example, we can analytically compute $c(x)$ given the following expression:

$$c(x) = \mathrm{P}\left[F(B) = 1 \mid x\right] = \frac{\int_{\mathcal{B}_+} \mathrm{P}(x \mid B)\,\mathrm{d}\,\mathrm{P}(B)}{\mathrm{P}(x)},$$

where $\mathcal{B}_+ = \{B : F(B) = 1\}$. As described, for positive instances $x_+ \in [0, 1]$, we have

$$c(x_+) = \frac{\int_\pi^1 \mathrm{P}(x_+ \mid B_\theta)\,\mathrm{P}(B_\theta)\,\mathrm{d}\theta}{\tfrac{1}{2}(1-P_{\text{neg}})(1+\pi)} = 1, \tag{6}$$

since positive instances always appear in positive bags. On the other hand, for negative instances,

$$\begin{aligned} c(x_-) &= \frac{\int_\pi^1 \mathrm{P}(x_- \mid B_\theta)\,\mathrm{P}(B_\theta)\,\mathrm{d}\theta}{P_{\text{neg}} + \tfrac{1}{2}(1-P_{\text{neg}})(1-\pi)} \\ &= \frac{\tfrac{1}{2}(1-P_{\text{neg}})(1-\pi)}{P_{\text{neg}} + \tfrac{1}{2}(1-P_{\text{neg}})(1-\pi)} \triangleq 1-\gamma. \end{aligned} \tag{7}$$

The resulting values of $c(x)$ are shown in Figure 2(d). Note that for this generative process, except for the trivial case in which $P_{\text{neg}} = 0$, $1 - \gamma = c(x_-) < 1$, so $\gamma > 0$. Thus, the assumption that negative instances appear in negative bags is automatically satisfied for the example in Figure 2. By construction, this example also satisfies the $\pi > 0$ assumption since there is zero probability of sampling a bag with $\theta \in (0, \pi)$ mass over positive bags. Hence, this example is an element of MI-GEN.

2.3 The Empirical Bag-Labeling Function

In MI-GEN, the instance- and bag-labeling functions are defined independently at the level of the generative model, and must satisfy the relationships indicated in Definition 1.

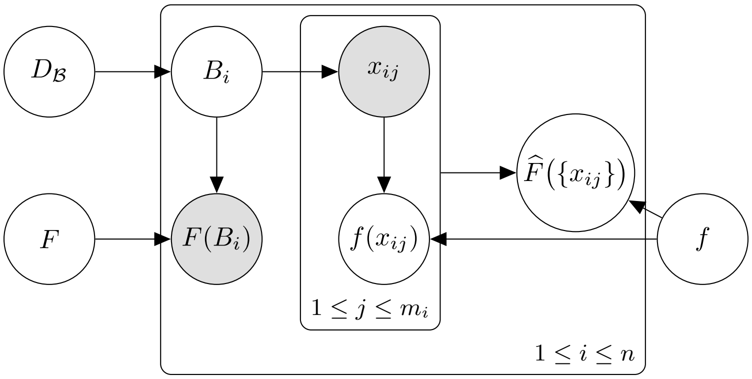


Figure 3: An illustration of the instance-, bag-, and empirical bag-labeling functions in MI-GEN.

However, in the standard MI setting, bag labels are typically viewed as being derived from instance labels. That is, if a bag is a set, then it is positive if it contains at least one positive instance, and negative otherwise.

Unlike the plate model in Figure 1(a), we do not typically observe bags directly in the MI setting. In the typical case, we only have access to samples $X_i = \{x_{ij}\}_{j=1}^{m_i}$, each drawn independently according to the distribution corresponding to each bag B_i so that $\left\{\left(\{x_{ij}\}_{j=1}^{m_i}, F(B_i)\right)\right\}_{i=1}^{n}$ is the observed MI data set, as shown in Figure 1(c). Each bag can be a different size, but we will use $m_l \leq m_i \leq m_u$ to denote the lower and upper bounds on bag sizes, respectively.

If we think of "bags" (in the sense of the standard generative model) of instances $\{x_{ij}\}_{j=1}^{m_i}$ as empirical samples drawn from the underlying bag distributions B_i in our model, then it is possible that samples from positive bags do not contain any positive instances. Hence, such "bags" would be negative in the sense of the standard model. To more harmoniously account for the standard notion of bag labels within our model, we introduce the empirical bag-labeling function, $\widehat{F} : \mathcal{X}^* \to \{0, 1\}$:

$$\widehat{F}(X_i) = \max_j f(x_{ij}), \tag{8}$$

where $X_i = \{x_{ij}\}_{j=1}^{m_i}$ is any finite set of instances.

We can think of $\widehat{F}$ as the bag labels that would be assigned by an oracle that had perfect information about the instance-labeling function f, but only an empirical sample from each bag. An illustration of the empirical bag-labeling function is shown in Figure 3. Figure 3 is a version of Figure 1(c) that shows the contributions of the instance-labeling and empirical bag-labeling functions. For every instance x_{ij} in an empirical bag sample, f assigns the label $f(x_{ij})$ to x_{ij}. On the other hand, $\widehat{F}$ is a function of the entire bag sample $X_i = \{x_{ij}\}_{j=1}^{m_i}$ as well as the instance-labeling function f, as specified in Equation 8.

The labeling functions F and $\widehat{F}$ will always agree on negative bags, since only negative instances are observed in negative bags. However, there might be some discrepancy between

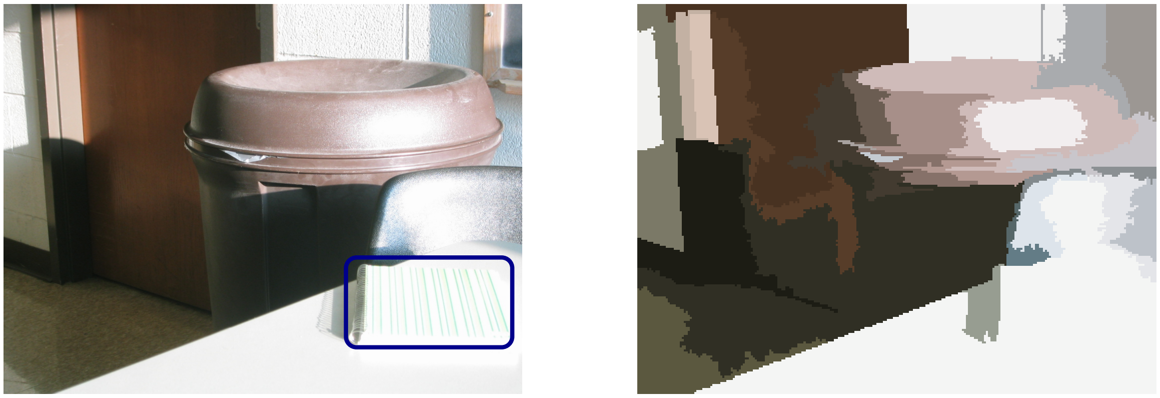


Figure 4: An example from the CBIR domain when a positive image does not contain a positive instance (the notebook, annotated in blue) after segmentation.

F and $\widehat{F}$ on positive bags if only negative instances are sampled within a positive bag. We return to characterizing the discrepancy between F and $\widehat{F}$ in Section 3.2.

The discrepancy between F and $\widehat{F}$ is essentially bag-label noise that naturally results from our generative model. Previous generative models do not account for this potential source of label noise, despite its presence in some domains. For example, in the drug activity prediction domain, even if it is known that a molecule activates a receptor, a sample of conformations from this molecule might not contain the particular positive conformation that causes activation. Likewise, for the CBIR domain, extracting a set of objects from images is often performed using segmentation achieved through local optimization (Andrews et al., 2003; Carson et al., 2002). Therefore, it is possible that no single instance in a bag generated from a positive image will correspond to the positive instance. Figure 4 shows an example from the SIVAL data set when, due to lighting conditions, the positive "notebook" instance in the image is grouped with the table during segmentation (Settles et al., 2008). This kind of "noise" is naturally captured by our generative model as the discrepancy between $\widehat{F}$ and F.

2.4 Relationship to Prior Models

The most general model in which instance learnability results have been previously shown is the "IID r-tuple" model (Blum and Kalai, 1998). The model, illustrated in Figure 5(a), assumes that each bag is generated by randomly sampling r instances in every bag from the same underlying instance distribution, $D_{\mathcal{X}}$. However, this is an unrealistic assumption for many domains. For example, consider the drug activity prediction setting. In this domain, that would mean that the conformations of every molecule are sampled independently from the same distribution, which is not true as it requires that different molecules share the same conformations. Likewise, for CBIR, the IID assumption asserts that all segments in *all* images are sampled from the same distribution, when the distributions over objects/segments clearly change between images.

To show that our model is more general than the IID r-tuple model, we now describe how to simulate this model within our model. First, we define each bag to be a probability distribution parameterized by an r-tuple of instances $B_{(x_1,\dots,x_r)}$. This distribution will be a weighted sum of point masses over each of the r instances: $\mathrm{P}(x \mid B_{(x_1,\dots,x_r)}) = \frac{1}{r}\sum_{i=1}^{r} \delta_{x_i}(x)$. Then, for any distribution $D_{\mathcal{X}}$ over instances (with $\mathrm{P}(x)$) and instance-labeling function f, we let the distribution over bags $D_{\mathcal{B}}$ be defined as $\mathrm{P}(B_{(x_1,\dots,x_r)}) \triangleq \prod_{i=1}^{r} \mathrm{P}(x_i)$, which is the probability that the corresponding r-tuple would have been sampled from $D_{\mathcal{X}}^r$, and the bag-labeling function F to be $F(B_{(x_1,\dots,x_r)}) = \max_{1\le i\le r} f(x_i)$. Let $p_{\text{neg}} = \mathrm{P}\left[f(x) = 0\right]$, then we claim that the $(D_{\mathcal{X}}, D_{\mathcal{B}}, f, F)$ described above is in MI-GEN $\left(p_{\text{neg}}^{r-1}, \frac{1}{r}\right)$.

First, we need to show that $D_{\mathcal{B}}$ as defined satisfies Condition 1 of Definition 1:

$$\begin{aligned}
\mathrm{P}(x) &\overset{?}{=} \int_{\mathcal{B}} \mathrm{P}(x \mid B)\,\mathrm{d}\,\mathrm{P}(B)\\
&= \int_{\mathcal{B}} \frac{1}{r}\sum_{i=1}^{r} \delta_{x_i}(x)\,\mathrm{d}\,\mathrm{P}(B_{(x_1,\dots,x_r)})\\
&= \frac{1}{r}\sum_{i=1}^{r} \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} \delta_{x_i}(x)\,\mathrm{d}\,\mathrm{P}(x_r)\cdots\,\mathrm{d}\,\mathrm{P}(x_1)\\
&= \frac{1}{r}\sum_{i=1}^{r} \left(\prod_{j\neq i}\int_{\mathcal{X}} \mathrm{d}\,\mathrm{P}(x_j)\right)\left(\int_{\mathcal{X}} \delta_{x_i}(x)\,\mathrm{d}\,\mathrm{P}(x_i)\right)\\
&= \frac{1}{r}\sum_{i=1}^{r} \left(1^{r-1}\right)\mathrm{P}(x) = \mathrm{P}(x).
\end{aligned}$$

So sampling instances under our two-step generative process is equivalent to sampling according to the original instance distribution.

Condition 2 of Definition 1 is trivially satisfied, since by the definition of F, positive instances never appear in negative bags. To show that Condition 3 holds, we must compute the probability that negative instances appear in a negative bag. Using the definition of conditional probability, this is,

$$\mathrm{P}\left[F(B) = 0 \mid x\right] = \frac{\int_{\mathcal{B}_-} \mathrm{P}(x \mid B)\,\mathrm{d}\,\mathrm{P}(B)}{\mathrm{P}(x)}.$$

Using the fact that in a negative bag $B_{(x_1,\dots,x_r)}$, all instances must be negative, we can compute the numerator for a negative instance as

$$\begin{aligned}\int_{\mathcal{B}_-} \mathrm{P}(x \mid B)\,\mathrm{d}\,\mathrm{P}(B) &= \int_{\mathcal{B}_-} \frac{1}{r}\sum_{i=1}^{r} \delta_{x_i}(x)\,\mathrm{d}\,\mathrm{P}(B_{(x_1,\dots,x_r)})\\ &= \frac{1}{r}\sum_{i=1}^{r}\int_{\mathcal{X}_-}\cdots\int_{\mathcal{X}_-} \delta_{x_i}(x)\,\mathrm{d}\,\mathrm{P}(x_r)\cdots\,\mathrm{d}\,\mathrm{P}(x_1)\\ &= \frac{1}{r}\sum_{i=1}^{r}\left(\prod_{j\neq i}\int_{\mathcal{X}_-}\mathrm{d}\,\mathrm{P}(x_j)\right)\left(\int_{\mathcal{X}_-}\delta_{x_i}(x)\,\mathrm{d}\,\mathrm{P}(x_i)\right)\\ &= \frac{1}{r}\sum_{i=1}^{r}\left(p_{\text{neg}}^{r-1}\right)\mathrm{P}(x) = p_{\text{neg}}^{r-1}\,\mathrm{P}(x).\end{aligned}$$

Thus, $\mathrm{P}\left[F(B) = 0 \mid x\right] = \frac{p_{\text{neg}}^{r-1}\,\mathrm{P}(x)}{\mathrm{P}(x)} = p_{\text{neg}}^{r-1}$. Since this probability is the same across all negative instances, this means that $\gamma = p_{\text{neg}}^{r-1}$. This quantity is positive as long as $p_{\text{neg}} > 0$. Otherwise, all instances are positive, so the $\gamma > 0$ assumption is vacuously satisfied.

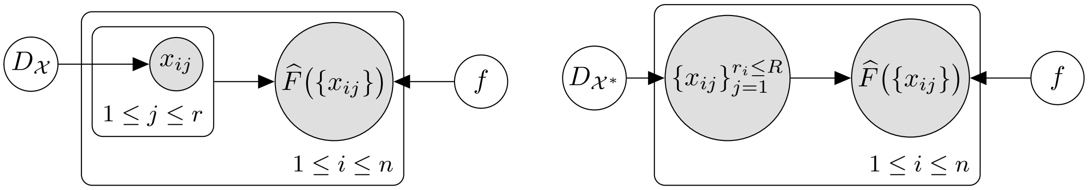
Finally, to show that Condition 4 of Definition 1 is satisfied, we see that for a positive bag, B_i,

$$\begin{aligned}\mathrm{P}\left[f(x) = 1 \mid B\right] &= \int_{\mathcal{X}} f(x)\left(\tfrac{1}{r}\sum_{i=1}^{r}\delta_{x_i}(x)\right)\mathrm{d}x\\ &= \frac{1}{r}\sum_{i=1}^{r} f(x_i) \geq \frac{1}{r} = \pi,\end{aligned}\tag{9}$$

since at least one instance in the bag is such that $f(x) = 1$. Therefore, the IID r-tuple model is a special case of our model in which γ and π are positive, and determined by the fraction of negative instances and bag size r.

Another generative model, used to show the learnability of bag-level concepts (Sabato and Tishby, 2012), allows arbitrary distributions over r-tuples. The model further relaxes the r-tuple model by allowing bag sizes to vary from 1 to R, some maximum bag size. The model is illustrated in Figure 5(b), where $D_{\mathcal{X}^*}$ denotes the distribution over tuples of size at most R. However, this model is also restrictive for many problem domains like drug activity prediction, since it enforces that bag sizes are finite and bounded, whereas molecules can exist in infinitely many conformations.

We can also represent the generative model of Sabato and Tishby (2012) in a similar way as for the IID r-tuple model. We simplify the space of bags to be atomic distributions over $r \leq R$ tuples, and allow an arbitrary distribution $D_{\mathcal{B}}$ over bags rather than requiring that $\mathrm{P}(B_{(x_1,\dots,x_r)}) = \prod_{i=1}^{r}\mathrm{P}(x_i)$. Now, $D_{\mathcal{X}}$ is not fixed, so we can define it in terms of Condition 1 of MI-GEN so that that condition is automatically satisfied. The bag-labeling function F is still defined in terms of the arbitrary instance-labeling function f, so Condition 2 is still trivially satisfied. Furthermore, by similar reasoning as in Equation 9, $\pi = \frac{1}{R}$ in this generative model, so Condition 4 is satisfied. However, the $\gamma > 0$ assumption (Condition 3) is no longer automatically satisfied by this generative process, since arbitrary distributions over tuples are allowed. Hence, while Sabato and Tishby (2012) analyze bag



(a) The IID r-tuple model

(b) The arbitrary R-tuple model

Figure 5: Previous generative models for MI data.

concept learnability with MI-GEN $\left(0, \frac{1}{R}\right)$, we require MI-GEN $\left(\gamma, \frac{1}{R}\right) \subset$ MI-GEN $\left(0, \frac{1}{R}\right)$ for instance concept learnability.

Babenko et al. (2011) propose treating bags in the MI setting as *manifolds* in the instance space $\mathcal{X}$. While this allows describing a bag with an infinite number of instances, it assigns an equal "weight" to every instance. However, for a domain like drug activity prediction, a molecule is more likely to exist in certain conformations than in others. The varying weight of instances is naturally handled in our setting, but we do not treat bags as manifolds over instances, so our results may not apply to the generative process in which bags are manifolds.

Some prior work proposes additional generative models for MIL, some of which model bags as distributions over instances. For example, some work uses Gaussian distributions (Maron and Lozano-Pérez, 1998; Xu, 2003) or mixture models (Foulds and Smyth, 2011) to represent distributions over instances, while other work uses more complex graphical (Yang et al., 2009; Adel et al., 2013; Kandemir and Hamprecht, 2014) or hierarchical Bayesian (Kuck and de Freitas, 2005) models. However, our work differs from these prior investigations in two important ways. First, we focus on the theoretical properties of our generative model, whereas prior work has only empirically explored the performance of algorithms tailored to specific generative models. Secondly, while prior generative models require that bags or instances are sampled from specific, parametric probability distributions, our generative model does not require such assumptions. Thus, the theoretical results presented below apply to more general scenarios than the models previously explored.

2.5 Applicability to Problem Domains

At the beginning of this section, we motivated MI-GEN using the 3D-QSAR domain. In this section, we elaborate on how bags can naturally be viewed as distributions in various other problem domains and which labeling tasks illustrated in Figure 3 are of interest in each domain. Of course, as described in Section 2.4, standard generative models are special cases of MI-GEN, so previous applications of MIL for which it is most natural to think of bags as finite sets of instances can still be incorporated in this model.

2.5.1 Drug Activity Prediction

For the drug activity prediction or the 3D-QSAR problem, it is natural to think of each molecule as a distribution over conformations. While it is natural to view learning molecule-level activity F as the ultimate goal of 3D-QSAR, it is also important to learn the instance-labeling function f. Knowing whether an individual conformation binds to a receptor provides information about the structure of the receptor's binding site, which is practically difficult to measure directly. Hence, learning both instance- and bag-labeling functions are important in the 3D-QSAR domain.

2.5.2 Text Categorization

While it is popular to represent documents as a flat "bag of words" using a single feature vector comprised of word frequencies (Salton and McGill, 1983), prior work has acknowledged the benefits of representing document-specific structure. In particular, latent Dirichlet allocation (LDA) models each document as a mixture of distributions over words (Blei et al., 2003). Of course, LDA can also applied to a coarser-grained representation in which documents are distributions over n-grams or paragraphs, which are like individual instances in the MI setting (Blei et al., 2003). Other work attempts to infer the level of granularity in a document in addition to modeling the distributions over the discovered "segments" (Du et al., 2013). Hence, treating documents as distributions is already a natural and popular representation for text. On the other hand, LDA treats each document distribution as taking a specific parametric form, whereas our results and analysis do not make any parametric assumptions about bag-level distributions.

As for 3D-QSAR, both document-level and instance-level categorization is important in the text categorization domain. For example, if certain types of documents like survey articles discuss various subjects, then it might be important to determine not just that the document as a whole discusses a particular subject, but also which specific passage or paragraph discusses the subject.

2.5.3 Content-Based Image Retrieval

Applying our generative framework to the CBIR task requires viewing images as distributions over objects such that the objects in each image are a sample from the corresponding distribution. As with LDA for the text categorization domain, analogous probabilistic models have been proposed for categorizing natural scenes (Fei-Fei and Perona, 2005). Thus, while treating images as distributions is not unprecedented, our analysis is novel in that it discuses learnability under such a model without assuming that image distributions take a specific parametric form.

Furthermore, as for the other domains discussed, the bag-labeling function F is not the only latent variable of interest in CBIR. In additional to labeling new images, a CBIR system might be interested in determining the location of the object of interest within an image, which requires learning the instance-labeling function f.

3. Learning Accurate Concepts from MI Data

In this section, we describe new theoretical results that highlight the advantages of the generative model proposed in Section 2. In particular, the new generative model allows new results about instance- and bag-concept learnability that previously only held under a much stronger set of assumptions. As we describe in Section 4, additional theoretical results imply the surprising but testable ability of standard supervised approaches to learn

Table 1: A summary of the learnability results in Section 3 and Section 4.

		Accuracy	**AUC**
Instance	f	Theorem 1	Theorem 4
Bag	$\widehat{F}$	Theorem 2	Theorem 5
	F	Theorem 3	Theorem 6

Table 2: Legend of the basic notation used in Section 3.

Symbol	**Description/Definition**	
$\mathcal{X}$	Space of instances	
$\mathcal{B}$	Space of bags (distributions over instances)	
$\mathcal{X}^*$	Set of bag samples (sets of instances)	
x_{ij}	Instance $x_{ij} \in \mathcal{X}$	
B_i	Bag $B_i \in \mathcal{B}$	
X_i	Bag sample $\{x_{ij}\}_{j=1}^{m_i} \in \mathcal{X}^*$, $x_{ij} \sim B_i$	$(m_l \leq \lvert X_i \rvert \leq m_u)$
m_i	Bag Sample Size	$m_i = \lvert X_i \rvert$
f	Instance-Labeling Concept	
F	Bag-Labeling Concept	
$\widehat{F}$	Empirical Bag-Labeling Concept	$\widehat{F}(X_i) \triangleq \max_j f(x_{ij})$
g	Instance-Labeling Hypothesis	
$\widehat{G}$	Empirical Bag-Labeling Hypothesis	$\widehat{G}(X_i) \triangleq \max_j g(x_{ij})$
$\mathcal{F}$	Instance-Labeling Concept Class	
VC($\mathcal{F}$)	Vapnik–Chervonenkis (VC) Dimension (Vapnik and Chervonenkis, 1971)	

to rank instances and bags from MI data. Table 1 summarizes the theoretical contributions made in this and the following sections, which demonstrate the learnability of the instance concept f, empirical bag-labeling function $\widehat{F}$, and bag-labeling function F with respect to both accuracy and ranking as measured by area under ROC (AUC). The results in this section and the following section use a model of the instance labeling function f to derive models for the bag-labeling functions $\widehat{F}$ and F.

Defining the ability of an algorithm to learn a good approximation of a target concept requires some metric by which the quality of the approximation is to be measured. Traditionally, the quality of a classifier is measured in terms of expected 0–1 loss. We begin by investigating the ability of algorithms to learn accurate concepts from MI data in this sense. While there is only one learning task in the supervised setting, there are now both instance- and bag-concept learning tasks in the MI setting, which we explore separately in the following sections. Table 2 shows the notation used for the concepts in this section.

3.1 Learning Accurate Instance Concepts

The probably approximately correctly (PAC) framework describes one sense in which it is possible to learn accurate concepts from supervised data. Since the generative process described in Section 2 differs from that for supervised learning, we must restate what it means to "PAC" learn an accurate instance concept under this model.

In the supervised setting, the learnability of some fixed concept class $\mathcal{F}$ is discussed without making any assumptions about the distribution over instances. The definition of MI-GEN in Definition 1 similarly allows any instance distribution, with which many bag distributions are consistent in the sense of Condition 1. To ensure that the target concept f is a member of the concept class $\mathcal{F}$, we must further restrict the set of models allowed by the generative process as follows:

Definition 2 (MI-GEN$_{\mathcal{F}}$) *For any $\gamma \in (0, 1]$ and $\pi \in [0, 1]$:*

$$\text{MI-GEN}_{\mathcal{F}}(\gamma, \pi) \triangleq \big\{(D_{\mathcal{X}}, D_{\mathcal{B}}, f, F) \in \text{MI-GEN}(\gamma, \pi) : f \in \mathcal{F}\big\}.$$

Now, we can formally define PAC learnability for the MI setting:

Definition 3 (Instance MI PAC-learning) *We say that an algorithm $\mathcal{A}$ MI PAC-learns instance concept class $\mathcal{F}$ from MI data when for any $(D_{\mathcal{X}}, D_{\mathcal{B}}, f, F) \in \text{MI-GEN}_{\mathcal{F}}(\gamma)$ with $\gamma > 0$, and $\epsilon_{\text{I}}, \delta > 0$, $\mathcal{A}$ requires $O\big(\mathsf{poly}(\frac{1}{\gamma}, \frac{1}{\epsilon_{\text{I}}}, \frac{1}{\delta})\big)$ bag-labeled instances sampled independently from the MI generative process in Figure 1(b) to produce an instance hypothesis g with risk $R_f(g) < \epsilon_{\text{I}}$ with probability at least $1 - \delta$ over samples.*[2]

Note that because our generative model allows us to discuss the marginalized instance distribution $D_{\mathcal{X}}$, the risk $R_f(g) = \mathrm{E}_{x \sim D_{\mathcal{X}}}\left[\mathbb{1}[f(x) \neq g(x)]\right]$ is measured with respect to this distribution as in the supervised setting. Now we show that instance concepts are MI PAC-learnable in the sense of Definition 3:

Theorem 1 *An instance concept class $\mathcal{F}$ with VC dimension $\mathrm{VC}(\mathcal{F})$ is Instance MI PAC-learnable using $O\left(\frac{1}{\epsilon_{\text{I}}\gamma}\left(\mathrm{VC}(\mathcal{F}) \log \frac{1}{\epsilon_{\text{I}}\gamma} + \log \frac{1}{\delta}\right)\right)$ examples.*

Proof By Condition 1 in Definition 1, we can treat bag-labeled instances as being drawn from the underlying instance distribution $D_{\mathcal{X}}$. Instances are observed with some label noise with respect to true labels given by f. Since positive instances never appear in negative bags (by Condition 2 of Definition 1), noise on instances is one-sided. If every negative instance appears in negative bags at least some γ fraction of the time (by Condition 3), then the maximum one-sided noise rate is $\eta = 1 - \gamma$. Since $\gamma > 0$, $\eta < 1$, which is required for learnability. Under our generative assumptions, the noise is "semi-random" in that noise rate might vary across instances, but is bounded by $\eta < 1$. Thus, learning an instance concept is equivalent to learning from data with one-sided label noise in this sense.

The result of Simon (2012) shows that in the presence of one-sided, semi-random noise, when a concept class $\mathcal{F}$ has a VC dimension of $\mathrm{VC}(\mathcal{F})$, $\mathcal{F}$ is PAC-learnable from $O\left(\frac{1}{\epsilon_{\text{I}}(1-\eta)}\left(\mathrm{VC}(\mathcal{F}) \log \frac{1}{\epsilon_{\text{I}}(1-\eta)} + \log \frac{1}{\delta}\right)\right)$ examples using a "minimum one-sided disagreement" strategy. This strategy entails choosing a classifier that minimizes the number of disagreements on positively-labeled examples while perfectly classifying all negatively-labeled examples. This strategy also works in the special case that all instances and bags are positive ($\eta = 0$, or $\gamma = 1$, since there are no negative instances). Substituting $1 - \gamma$ for η in the

[2] Alternate definitions of PAC-learnability require that $\mathcal{A}$ also take at most polynomial *time* to produce hypothesis g. We defer the discussion of time complexity to Section 3.5.

bound of Simon (2012) yields the bound in terms of γ. ■

We note that Theorem 1 allows for "noisy" positive bags without positive instances ($\pi = 0$), since the additional bag-level noise is essentially absorbed into η.

3.2 Learning Accurate Bag Concepts

As for instance concept learnability, we must formally define what we mean to learn accurate bag concepts in the MI setting. As described in Section 2, there are two bag-labeling functions we might be interested in learning. In our generative model, we assume that the MI relationship between bag and instance labels holds at the level of the generative process. That is, bags are directly assigned labels by a bag concept F. On the other hand, given a set of instances sampled from a bag, we might be interested in learning the more traditional bag-labeling concept in the MI setting, $\widehat{F}(X_i) = \max_j f(x_{ij})$, which we have called the empirical bag-labeling function (Equation 8). We will first analyze learnability with respect to the empirical bag-labeling function and then extend this result to the true bag-labeling function.

We can define the risk of a bag-labeling concept $\widehat{G}$ with respect to the underlying empirical bag-labeling concept $\widehat{F}$ as follows:

$$\begin{aligned} R_{\widehat{F}}(\widehat{G}) &= \mathrm{E}\left[\mathbb{1}[\widehat{F}(X) \neq \widehat{G}(X)]\right] \\ &= \int_{\mathcal{B}} \int_{\mathcal{X}^*} \mathbb{1}[\widehat{F}(X) \neq \widehat{G}(X)] \,\mathrm{d}\,\mathrm{P}(X \mid B) \,\mathrm{d}\,\mathrm{P}(B), \end{aligned} \tag{10}$$

where $\mathrm{P}(X \mid B)$ is the probability of sampling the set of instances X from bag B. Since we assume that instances are sampled IID according to B, $\mathrm{P}(X \mid B) = \prod_{x \in X} \mathrm{P}(x \mid B)$. Given a formal definition of the risk of an empirical bag-labeling function, we can define learnability with respect to this function below:

Definition 4 (Empirical Bag MI PAC-learning) *We say that an algorithm $\mathcal{A}$ MI PAC-learns empirical bag-labeling functions derived from instance concept class $\mathcal{F}$ when for any $(D_{\mathcal{X}}, D_{\mathcal{B}}, f, F) \in$ MI-GEN$_{\mathcal{F}}(\gamma)$ with $\gamma > 0$, and $\epsilon_{\mathrm{B}}, \delta > 0$, $\mathcal{A}$ requires $O\big(\mathsf{poly}(\frac{1}{\gamma}, \frac{1}{\epsilon_{\mathrm{B}}}, \frac{1}{\delta})\big)$ bag-labeled instances sampled independently from the MI generative process in Figure 1(b) to produce an empirical bag-labeling function $\widehat{G}$ with risk $R_{\widehat{F}}(\widehat{G}) < \epsilon_{\mathrm{B}}$ with probability at least $1 - \delta$ over samples.*

To show empirical bag concept learnability under our generative model, we will show that by learning an accurate enough instance concept g, the resulting empirical bag-labeling concept given by $\widehat{G}(X_i) = \max_j g(x_{ij})$ will have low risk with respect to $\widehat{F}$. Thus, we start with a bound on $R_{\widehat{F}}(\widehat{G})$ in terms of $R_f(g)$.

Lemma 1 *Let $R_f(g)$ be the risk of an instance labeling concept g, and $R_{\widehat{F}}(\widehat{G})$ be the risk of the empirical bag-labeling function $\widehat{G}(X_i) = \max_j g(x_{ij})$. Then if bag sample sizes are bounded by m_u ($\forall i : |X_i| \leq m_u$), $R_{\widehat{F}}(\widehat{G}) \leq m_u R_f(g)$.*

Proof See Appendix A. ■

Given the bound demonstrated in Lemma 1, we can derive the following result:

Theorem 2 *Empirical bag-labeling functions derived from instance concept class $\mathcal{F}$ with VC dimension $\mathrm{VC}(\mathcal{F})$ are PAC-learnable from MI data using*

$$O\left(\tfrac{m_u}{\epsilon_{\mathrm{B}}\gamma}\left(\mathrm{VC}(\mathcal{F})\log\tfrac{m_u}{\epsilon_{\mathrm{B}}\gamma}+\log\tfrac{1}{\delta}\right)\right)$$

examples.

Proof The general strategy is to learn an approximation g for $f \in \mathcal{F}$ using minimum one-sided disagreement as mentioned in the proof of Theorem 1 and then to derive an empirical bag-labeling function $\widehat{G}$ from g.

For a desired bound ϵ_{B} on $R_{\widehat{F}}(\widehat{G})$, by using $\epsilon_{\mathrm{I}} = \frac{\epsilon_{\mathrm{B}}}{m_u}$ in Theorem 1, this ensures that the resulting instance classifier is such that $R_f(g) < \frac{\epsilon_{\mathrm{B}}}{m_u}$ with high probability. Combined with the result in Lemma 1, this implies that $R_{\widehat{F}}(\widehat{G}) \le m_u R_f(g) < m_u\left(\frac{\epsilon_{\mathrm{B}}}{m_u}\right) = \epsilon_{\mathrm{B}}$, so $R_{\widehat{F}}(\widehat{G}) < \epsilon_{\mathrm{B}}$ as desired. Substituting $\epsilon_{\mathrm{I}} = \frac{\epsilon_{\mathrm{B}}}{m_u}$ into the bound in Theorem 1 gives the bound as stated in Theorem 2. ■

Again, Theorem 2 allows for noisy positive bags without positive instances ($\pi = 0$). Furthermore, in the special case when every bag sample is a singleton $X = \{x\}$, then $m_u = 1$ and $\widehat{F}(\{x\}) = f(x)$. Thus, the instance concept learnability result in Theorem 1 is really just a special case of learning an empirical bag-labeling function with bags of size 1 as in Theorem 2.

Next, we turn our attention to learning the underlying bag-labeling function F. We will still use the instance-labeling function g to derive this bag-labeling function. It is possible to consider learning F without the use of an instance concept g, as we discuss in Section 3.3. During both training *and* testing, we are only given access to a sample X_i from each bag B_i with which we can estimate $F(B_i)$. Therefore, we will again learn an empirical bag labeling function $\widehat{G}(X_i)$. However, now we will assess the quality of $\widehat{G}$ with respect to the underlying bag-labeling function F as follows:

$$\begin{aligned} R_F(\widehat{G}) &= \mathrm{E}\left[\mathbb{1}[F(B) \neq \widehat{G}(X)]\right] \\ &= \int_{\mathcal{B}}\int_{\mathcal{X}^*} \mathbb{1}[F(B) \neq \widehat{G}(X)]\,\mathrm{d}\,\mathrm{P}(X \mid B)\,\mathrm{d}\,\mathrm{P}(B). \end{aligned} \tag{11}$$

The definition of bag concept learnability then takes the same form as that in Definition 4 with the risk as given in Equation 11. As we will show in Lemma 2, we now also require the further assumption that π, minimum fraction of positive instances in positive bags, is nonzero.

Definition 5 (Bag MI PAC-learning) *We say that an algorithm $\mathcal{A}$ MI PAC-learns bag-labeling functions derived from instance concept class $\mathcal{F}$ when for any $(D_{\mathcal{X}}, D_{\mathcal{B}}, f, F) \in$ MI-GEN$_{\mathcal{F}}(\gamma, \pi)$ with $\gamma, \pi > 0$, and $\epsilon_{\mathrm{B}}, \delta > 0$, algorithm $\mathcal{A}$ requires $O\big(\mathsf{poly}(\frac{1}{\gamma}, \frac{1}{\pi}, \frac{1}{\epsilon_{\mathrm{B}}}, \frac{1}{\delta})\big)$ bag-labeled instances sampled independently from the MI generative process in Figure 1(b) to produce an empirical bag-labeling function $\widehat{G}$ with risk $R_F(\widehat{G}) < \epsilon_{\mathrm{B}}$ with probability at least $1 - \delta$ over samples.*

In order to show learnability of the bag-labeling concept F, we adopt a similar strategy as for Theorem 2 in which we first learn an instance-labeling concept g, then use g to derive an empirical bag-labeling concept $\widehat{G}$. Since Theorem 2 shows that we can a learn a concept $\widehat{G}$ that accurately models $\widehat{F}$, what remains to be shown is that $\widehat{F}$ is an accurate model of F under some additional conditions. First, we prove the following lemma, which decomposes the risk $R_F(\widehat{G})$ into the discrepancy between $\widehat{G}$ and $\widehat{F}$, and the discrepancy between $\widehat{F}$ and F.

Lemma 2 *For any empirical bag-labeling concept $\widehat{G}$,*

$$R_F(\widehat{G}) \leq R_{\widehat{F}}(\widehat{G}) + R_F(\widehat{F}).$$

Proof See Appendix A. ■

Now, we derive a bound on the discrepancy between the empirical bag-labeling function $\widehat{F}$ and the underlying bag-labeling function F. Since this discrepancy arises when we do not sample a positive instance within a positive bag, the bound depends on the minimum bag sample size and the minimum fraction π of positive instances in every positive bag.

Lemma 3 *Suppose bag samples are of size at least m_l ($\forall i : m_l \leq |X_i|$), then $R_F(\widehat{F}) \leq (1-\pi)^{m_l}$.*

Proof See Appendix A. ■

Finally, we can now show the following learnability result with respect to the underlying bag-labeling function. However, note that in Lemma 3, the error $R_F(\widehat{F})$ decreases with the minimum bag size m_l. Thus, in order to achieve low error with respect to F, we must ensure that bags in the test set are sufficiently large. Therefore, the following result is stated under the additional condition that the test bag sample sizes m_i satisfy some constraints. Note that these constraints arise naturally from the process that samples instances from bag distributions.

Theorem 3 *Bag-labeling functions derived from instance concept class $\mathcal{F}$ with VC dimension $\mathrm{VC}(\mathcal{F})$ are PAC-learnable from MI data using*

$$O\left(\tfrac{1}{\epsilon_{\mathrm{B}}^2\gamma\pi}\left(\mathrm{VC}(\mathcal{F})\log\tfrac{1}{\epsilon_{\mathrm{B}}\gamma\pi} + \log\tfrac{1}{\delta}\right)\right) \tag{12}$$

examples when test bag sample sizes are bounded by $m_l \leq m \leq m_u$ and m_l is large enough such that $m_l \geq \frac{1}{\pi}\log\frac{2}{\epsilon_{\mathrm{B}}}$.

Proof Intuitively, we can learn an instance-labeling function g according to Theorem 1 and then use the resulting empirical bag-labeling function $\widehat{G}$. By combining the previously stated results, we can bound $R_F(\widehat{G})$ as

$$\begin{aligned}
R_F(\widehat{G}) &\leq R_{\widehat{F}}(\widehat{G}) + R_F(\widehat{F}) && \text{(by Lemma 2)}\\
&\leq m_u R_f(g) + R_F(\widehat{F}) && \text{(by Lemma 1)}\\
&\leq m_u R_f(g) + (1-\pi)^{m_l}. && \text{(by Lemma 3)}
\end{aligned}$$

In the case that $\pi = 1$, then the second term in the sum is zero. Otherwise, suppose the minimum bag size is such that

$$m_l \geq \tfrac{1}{\pi} \log \tfrac{2}{\epsilon_{\mathrm{B}}} \geq \frac{\log \frac{\epsilon_{\mathrm{B}}}{2}}{\log(1-\pi)} = \log_{1-\pi} \tfrac{\epsilon_{\mathrm{B}}}{2},$$

where the second inequality follows from the fact that $\pi \leq -\log(1-\pi)$ for $\pi \in (0,1)$. Therefore, since $(1-\pi) < 1$, we have that

$$(1-\pi)^{m_l} \leq (1-\pi)^{\log_{1-\pi} \frac{\epsilon_{\mathrm{B}}}{2}} = \tfrac{\epsilon_{\mathrm{B}}}{2}.$$

Furthermore, when learning the instance concept g, we can choose ϵ_{I} to be such that $\epsilon_{\mathrm{I}} = \frac{\epsilon_{\mathrm{B}}}{2m_u}$. Since g will be such that $R_f(g) < \epsilon_{\mathrm{I}}$ with probability $(1-\delta)$, with the same probability we have that

$$\begin{aligned} R_F(\widehat{G}) &\leq m_u R_f(g) + (1-\pi)^{m_l} \\ &< m_u \left(\tfrac{\epsilon_{\mathrm{B}}}{2m_u}\right) + \tfrac{\epsilon_{\mathrm{B}}}{2} = \epsilon_{\mathrm{B}}. \end{aligned}$$

Substituting the expression for ϵ_{I} in terms of ϵ_{B} into the bound in Theorem 1 gives the sample complexity bound:

$$O\left(\tfrac{m_u}{\epsilon_{\mathrm{B}}\gamma}\left(\mathrm{VC}(\mathcal{F}) \log \tfrac{m_u}{\epsilon_{\mathrm{B}}\gamma} + \log \tfrac{1}{\delta}\right)\right),$$

which is the same bound as stated in Theorem 2.

It is also possible to sub-sample large bags such that there is also a conservative upper bound on sample size $m_u = O\left(\frac{1}{\epsilon_{\mathrm{B}}\pi}\right)$, which is consistent with $m_l \geq \frac{1}{\pi} \log \frac{2}{\epsilon_{\mathrm{B}}}$. Then, we can derive an expression for learnability in terms of π. Substituting this bound into that of Theorem 2 gives the second sample complexity bound as stated in Equation 12. ■

3.3 Discussion

The results presented in Section 3.1 and Section 3.2 follow the same basic strategy. First, minimum one-sided disagreement is used to learn an accurate instance concept g in the presence of one-sided noise on bag-labeled instances. Then, for the bag-labeling task, instance labels are aggregated using an empirical bag-labeling function $\widehat{G}$ to approximate the empirical bag-labeling function $\widehat{F}$ or the underlying bag-labeling function F. The idea of combining instance labels to produce a bag-labeling function is used by many existing MI algorithms.

However, under the generative model that treats bags as distributions, the bag-labeling results derived in Section 3.2 are somewhat counterintuitive. On the one hand, if bags are distributions from which we observe samples, then the larger the samples, the more information an algorithm has about the underlying bag distribution. Intuitively, it seems that the better an algorithm can estimate the underlying bag distribution, which is the object of interest for classification, the better it can learn a concept to label new bags. On

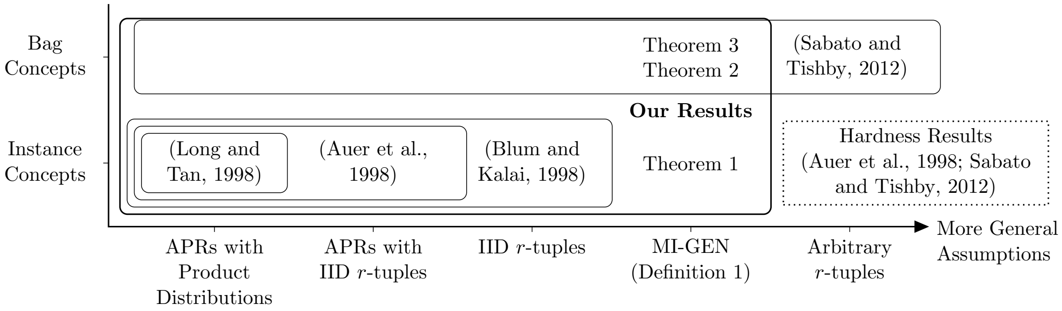


Figure 6: Relation to prior learnability results.

the other hand, the result in Theorem 2 suggests that it is harder to learn from larger bag sizes, since roughly $O(m_u \log m_u)$ more examples are required to learn an accurate concept.

Essentially, the source of this incoherence in reasoning is the use of an instance-labeling concept g to derive the bag-labeling concept $\widehat{G}$. In the process of combining instance labels to label a bag, small errors in the instance labeling function g compound quickly. For example, g must label *all* instances in a negative bag correctly for $\widehat{G}$ to label the bag correctly. As bag size increases, it becomes less likely that g will agree with f across all instances.

Therefore, despite our positive results suggesting that learning an accurate instance-labeling function is sufficient to learn an accurate bag-labeling function, in practice, it is possible to imagine an alternative approach in which bag-labeling functions are learned directly by representing bags in a supervised fashion. Several practical approaches, such as bag-level kernels (Gärtner et al., 2002; Zhou et al., 2009) and embeddings (Chen et al., 2006; Foulds, 2008) do precisely this, and essentially turn the bag-label learning problem into a supervised learning problem. Further investigating the trade-offs between techniques that classify bags using instance classifiers and those that directly learn bag classifiers is an interesting direction for future work.

3.4 Relation to Prior Learnability Results

An overview of our results in the context of prior work is shown in Figure 6. Early work on instance learnability shows that axis-parallel rectangles (APRs) are learnable from MI data, but under the restrictive assumption that each bag contains r IID instances sampled from a product distribution (Long and Tan, 1998). Later work by Auer et al. (1998) extends these results to the case when the instance distribution is no longer a product distribution, but the instances are still sampled IID from a single distribution across bags. The most recent results on instance concept learnability in the MI setting are described by Blum and Kalai (1998). Like the proof of Theorem 1, Blum and Kalai (1998) also reduce the problem of learning instance concepts to learning from noisy examples. However, the proof in Blum and Kalai (1998) requires that the label noise on negative examples be uniformly random. This condition is met under the strong assumption made in that work, that instances in all bags are drawn IID from the same distribution over instances. On the other hand, the result in Theorem 1 applies to our more general model in which the noise rate can vary across instances. Hence, our results rely on the recent work of Simon (2012), which shows that it is possible to learn from instances corrupted with semi-random one-sided noise.

The bag learnability results in Section 3.2 show that an accurate bag concept can be learned by learning an accurate instance concept and deriving a bag concept by combining instance labels within a bag. Other recent work on bag concept learnability takes a different approach. The strategy of Sabato and Tishby (2012) is to directly learn empirical bag-labeling concepts using empirical risk minimization (ERM). That is, they suppose that an algorithm selects an instance-labeling function $g \in \mathcal{F}$ that minimizes $R_{\widehat{F}}(\widehat{G})$. Since general sample complexity bounds exist for ERM in terms of capacity measures such as VC dimension of a hypothesis class (Blumer et al., 1989), Sabato and Tishby (2012) proceed by proving that the capacity of the function class $\left\{\widehat{G} : \widehat{G}(X_i) = \max_j g(x_{ij}),\, g \in \mathcal{F}\right\}$ is bounded in terms of the capacity of $\mathcal{F}$. In fact, the results of Sabato and Tishby (2012) apply to more general cases in which the combining function used to derive a bag-labeling function from an instance-labeling function is other than the max function. However, the results of Sabato and Tishby (2012) do not prove positive results about instance-labeling concepts.

As indicated in Figure 6, the results in Section 3.2 are not a strict generalization of those in Sabato and Tishby (2012), nor are those in Sabato and Tishby (2012) a generalization of those in Section 3.2. In particular, since MI-GEN treats bags as distributions, the results in Section 3.2 apply to cases not considered in Sabato and Tishby (2012), in which bags are assumed to have finite size. On the other hand, while our generative model can encapsulate aspects of the generative model in Sabato and Tishby (2012) (see Section 2.4), arbitrary distributions over r-tuples are not permitted. However, it may be that we are able to prove positive instance learnability results precisely *because* of this constraint on the generative process.

Other recent work has discussed the difficulty, both theoretically and in practice, to relate the performance of the same classifier on the instance- and bag-labeling tasks (Tragante do Ó et al., 2011). In contrast, Lemma 1 illustrates a clear connection between the accuracy of an instance concept and that of the resulting empirical bag concept. This new connection is made possible by the relationship between bag and instance distributions in our generative model, as highlighted in Lemma 1. In particular, Condition 1 of Definition 1 is employed to marginalize out the effect of individual bag distributions so that error on bags can be expressed directly in terms of the error on instances. There are at least two reasons why this relationship is not obvious from empirical results. First, contrasting the sample complexity expressions Theorem 1 and Theorem 3, we see that larger samples are required to learn accurate bag concepts. Thus, for a fixed sample size, there might be a significant discrepancy in performance of a single algorithm on the two learning tasks. Secondly, the relationship we demonstrate holds when an accurate instance concept is learned and *then* applied to the bag-labeling task. However, in practice, many algorithms attempt to learn an instance function to label bags using ERM at the *bag-level.* It is not clear that the instance-labeling function found via bag-level ERM in this way will successfully label instances.

3.5 Relation to Prior Hardness Results

The positive learnability results in Section 3.1 and Section 3.2 do not contradict existing hardness results about learning in the MI setting. Essentially, most hardness results are shown under the scenarios that lie on the far right of Figure 6. For example, Sabato and Tishby (2012) observe that if only positive bags are generated, then learning the bag-labeling function is trivial, but no label information about instances is provided. In this case, learning instance labels is equivalent to learning in the *unsupervised learning* setting, for which no PAC-style guarantees can be made. However, the additional assumptions in MI-GEN preclude the case when only positive bags appear, since the negative instances would never appear in negative bags as required by Condition 3 in Definition 1.

Similarly, under the weak assumption in which arbitrary distributions over r-tuples are allowed, Auer et al. (1998) show that that efficiently PAC-learning MI instance concepts is impossible (unless $\mathsf{NP} = \mathsf{RP}$[3]). While the results on instance and bag learnability stemming from Theorem 1 show that a polynomial number of examples can be used to learn accurate concepts, they do not bound the computational complexity of learning from the examples. In particular, minimum one-sided disagreement is known to be NP-hard for certain concept classes and loss functions (Simon, 2012). Therefore, for some concept classes, instance and bag concepts are not *efficiently* PAC-learnable: learnable with a polynomial number of examples *in polynomial time*.

The apparent contradiction between our learnability results and the hardness results of Auer et al. (1998) is resolved by observing that MI-GEN precludes the scenario used to reduce learning disjunctive normal form (DNF) formulae to learning APRs from MI data. In the reduction used by Auer et al. (1998), each instance corresponds to a (variable assignment, clause) pair, and a bag is formed for each variable assignment by including a pair with that variable assignment for each clause. Bags are sampled uniformly over all variable assignments. Suppose a particular variable assignment v satisfies the first clause c_1, but not the second clause c_2. Then the instance (v, c_1) is positive, but (v, c_2) is negative. However, (v, c_2) only ever appears in bags along with (v, c_1); that is, in positive bags. This violates the condition that $\gamma > 0$, or that negative instances appear with some probability in negative bags, so our results do not apply to this hard scenario.

Similarly, our generative model precludes scenarios used to show the hardness of learning hyperplane concepts for MI data (Kundakcioglu et al., 2010; Diochnos et al., 2012; Doran and Ray, 2013). It is unknown whether there is an algorithm to efficiently learn hyperplanes that minimize one-sided disagreement. However, even ERM under 0–1 loss is NP-hard for the concept class of hyperplanes (Ben-David et al., 2003), which are widely used in practice for supervised learning. Thus, while previous results have characterized the hardness of MI learning as resulting from arbitrary distributions across bags, our results suggest that the hardness arises from cases in which $\gamma = 0$, or when negative instances only occur in negative bags.

[3] RP is the class of decision problems for which a probabilistic Turing machine terminates in polynomial time, always returns `NO` when the answer is `NO`, and returns `YES` with probability at least $\frac{1}{2}$ when the answer is `YES`.

3.6 Must Instances be Dependent Samples?

As observed in prior work, most real-world examples of MIL have bags that contain non-IID instances (Zhou et al., 2009). Thus, our assumption that bag samples X_i are drawn IID according to their corresponding bag distributions B_i might seem unrealistic. However, note that our generative model *does* allow for dependencies between instances at the level of bag distributions, B_i. That is, although the samples X_i are drawn from bag distributions independently, we can use such independent samples to *approximate* the behavior of empirical bag-labeling functions on *non*independent samples.

The arbitrary R-tuple model, as illustrated in Figure 5(b), allows for arbitrary distributions over tuples of size at most R, which can be used to represent any generative model in which there is a relationship between instances in bags (i.e., bags in which instances are non-IID). As described in Section 2.4, it is possible to represent this model within MI-GEN where each bag is an atomic distribution over the instances in the tuple and the distribution over bags corresponds to the original distribution over tuples. Given this representation, $\pi = \frac{1}{R}$ in our model. In the traditional MI setting, we would directly observe these R instances. Our generative model, on the other hand, assumes that we perform the equivalent of repeatedly sampling an instance from these R instances uniformly and independently at random. In this case, we have the following result:

Corollary 1 (MI PAC-learning from Dependent Instances) *When distributions of instances in bags are defined by a set of R dependent instances sampled from a distribution over R tuples, bag-labeling functions derived from instance concept class $\mathcal{F}$ with VC dimension $\mathrm{VC}(\mathcal{F})$ are PAC-learnable from MI data using*

$$O\left(\tfrac{R}{\epsilon_{\mathrm{B}}\gamma}\log\tfrac{1}{\epsilon_{\mathrm{B}}}\left(\mathrm{VC}(\mathcal{F})\log\tfrac{R}{\epsilon_{\mathrm{B}}\gamma}+\log\tfrac{1}{\delta}\right)\right)$$

examples with test bags of size $m = \left\lceil R\log\frac{2}{\epsilon_{\mathrm{B}}}\right\rceil$ drawn independently with replacement from the R dependent instances.

Proof Following the same line of reasoning as in Theorem 3, we can derive the sample complexity bound

$$O\left(\tfrac{m}{\epsilon_{\mathrm{B}}\gamma}\left(\mathrm{VC}(\mathcal{F})\log\tfrac{m}{\epsilon_{\mathrm{B}}\gamma}+\log\tfrac{1}{\delta}\right)\right), \tag{13}$$

when there are m instances per bag. Choosing $m = \left\lceil R\log\frac{2}{\epsilon_{\mathrm{B}}}\right\rceil$ satisfies the conditions of that theorem, since $\pi = \frac{1}{R}$. Substituting m into Equation 13 implies the sample complexity as stated in the corollary. ■

Thus, even though the bag distribution is representable using only R dependent instances, when sampling independently, we must sample a factor of $O\left(\log\frac{1}{\epsilon_{\mathrm{B}}}\right)$ more instances to ensure that we can learn an accurate bag-labeling concept with high probability.

4. Learning to Rank from MI Data

Learnability results are often stated as in Section 3 with respect to the accuracy metric. However, other metrics often provide a more useful characterization of algorithm performance in practice. For example, for the 3D-QSAR problem, it is not necessary to accurately

Table 3: Legend of the basic notation used in Section 4.

Symbol	Description/Definition	
$\mathcal{X}$	Space of instances	
$\mathcal{B}$	Space of bags (distributions over instances)	
$\mathcal{X}^*$	Set of bag samples (sets of instances)	
x_{ij}	Instance $x_{ij} \in \mathcal{X}$	
B_i	Bag $B_i \in \mathcal{B}$	
X_i	Bag sample $\{x_{ij}\}_{j=1}^{m_i} \in \mathcal{X}^*$, $x_{ij} \sim B_i$	$(m_l \leq \lvert X_i \rvert \leq m_u)$
m_i	Bag Sample Size	$m_i = \lvert X_i \rvert$
c	p-concept for bag-labeled instances	$c(x) \triangleq \mathrm{P}\left[F(B) = 1 \mid x\right]$
h	Instance-Labeling p-concept	
$\widehat{H}$	Empirical Bag-Labeling p-concept	$\widehat{H}(X_i) \triangleq \max_j h(x_{ij})$
p_{neg}, p	$p_{\text{neg}} \triangleq \mathrm{P}\left[f(x) = 0\right]$	$p \triangleq \min\{p_{\text{neg}}, 1 - p_{\text{neg}}\}$
$\widehat{P}_{\text{neg}}$, $\widehat{P}$	$\widehat{P}_{\text{neg}} \triangleq \mathrm{P}\left[\widehat{F}(X) = 0\right]$	$\widehat{P} \triangleq \min\{\widehat{P}_{\text{neg}}, 1 - \widehat{P}_{\text{neg}}\}$
P_{neg}, P	$P_{\text{neg}} \triangleq \mathrm{P}\left[F(B) = 0\right]$	$P \triangleq \min\{P_{\text{neg}}, 1 - P_{\text{neg}}\}$
$\mathcal{C}$	Instance-Labeling p-concept Class	
$\mathrm{PD}(\mathcal{C})$	Pseudo-Dimension of $\mathcal{C}$ (Haussler, 1992)	

predict the activity of every molecule. Instead, a classifier can produce a ranked list indicating its confidence that each molecule is active. The set of active molecules with the highest predicted activity can then be investigated further by chemists. Unlike the prior work on learning accurate concepts from MI data as shown in Figure 6, there has been virtually no prior work on learning to rank in the MI setting. That is, although ranking algorithms have been developed for MIL (Bergeron et al., 2008), there is no formal analysis of the performance of such approaches. We provide such an analysis in this section.

In the 3D-QSAR example, a desirable property of a classifier is that it appropriately *ranks* bags or instances. That is, it assigns a higher real-valued confidence that a conformation is positive to actual positive conformations than to negative conformations. The AUC metric is commonly used to measure the ranking performance of a classifier. We show in this section that classifiers with high AUC are also learnable from MI data under our generative model. Furthermore, we show that learning high-AUC concepts from MI data is easier than learning accurate concepts in the sense that it can be achieved using standard ERM approaches. This suggests that standard supervised algorithms can learn high-AUC concepts from MI data generated according to MI-GEN, a surprising hypothesis that we evaluate in the final section.

4.1 Learning High-AUC Instance Concepts

Prior work has shown that the AUC is equivalent to the probability that a randomly selected positive example will be assigned a higher confidence than a randomly selected negative example (Hanley and McNeil, 1982). We can define a corresponding instance AUC error of a real-valued hypothesis h as $1 - \text{AUC}$, or the probability that a negative instance is assigned a higher confidence than a positive instance:

$$\begin{aligned} R_f^{\text{AUC}}(h) &= \int_{\mathcal{X}} \int_{\mathcal{X}} \mathbb{1}[h(x_-) > h(x_+)] \, \mathrm{d}\,\mathrm{P}(x_+ \mid f(x_+) = 1) \, \mathrm{d}\,\mathrm{P}(x_- \mid f(x_-) = 0) \\ &= \frac{\int_{\mathcal{X}_-} \int_{\mathcal{X}_+} \mathbb{1}[h(x_-) > h(x_+)] \, \mathrm{d}\,\mathrm{P}(x_+) \, \mathrm{d}\,\mathrm{P}(x_-)}{\mathrm{P}\left[f(x) = 1\right] \mathrm{P}\left[f(x) = 0\right]} \\ &= \frac{1}{(1 - p_{\text{neg}}) p_{\text{neg}}} \int_{\mathcal{X}_-} \int_{\mathcal{X}_+} \mathbb{1}[h(x_-) > h(x_+)] \, \mathrm{d}\,\mathrm{P}(x_+) \, \mathrm{d}\,\mathrm{P}(x_-). \end{aligned} \tag{14}$$

The first step follows from the definition of conditional probability, and we introduce $p_{\text{neg}} = \mathrm{P}\left[f(x) = 0\right]$ for notational convenience (see Table 3 for a list of notation used in this section). By definition, this quantity is zero in the cases when either all instance are positive or all instances are negative.

Given the formal definition of AUC, we can begin to describe how it is possible to learn high-AUC instance concepts from MI data. Since a classifier's confidence values are relevant for the AUC metric, we will consider the hypothesis class corresponding to a classifier to be a p-concept class $\mathcal{C}$. The p-concept model is a model for binary classification in which a p-concept $c : \mathcal{X} \to [0, 1]$ represents the probability that an instance $\mathcal{X}$ is observed with a positive label (Kearns and Schapire, 1994). For high-AUC instance learnability, we will show that it is sufficient to learn a p-concept $h \in \mathcal{C}$ that models the p-concept $c(x) = \mathrm{P}\left[F(B) = 1 \mid x\right]$, the probability of observing instance x in a positive bag as defined in Equation 2.

To ensure that the target concept c is also a member of $\mathcal{C}$, we must formally restrict the set of bag labeling functions and distributions that are permitted by the generative model as follows:

Definition 6 (MI-GEN$_{\mathcal{C}}$) *For any $\gamma \in (0, 1]$ and $\pi \in [0, 1]$:*

$$\text{MI-GEN}_{\mathcal{C}}(\gamma, \pi) \triangleq \Big\{ (D_{\mathcal{X}}, D_{\mathcal{B}}, f, F) \in \text{MI-GEN}(\gamma, \pi) : \big(x \mapsto \mathrm{P}\left[F(B) = 1 \mid x\right]\big) \in \mathcal{C} \Big\}.$$

Learnability of a p-concept with high AUC is then defined with respect to p-concept class $\mathcal{C}$:

Definition 7 (Instance MI AUC-PAC-learning) *We say that an algorithm $\mathcal{A}$ MI AUC-PAC-learns instance p-concept class $\mathcal{C}$ from MI data when for any $(D_{\mathcal{X}}, D_{\mathcal{B}}, f, F) \in \text{MI-GEN}_{\mathcal{C}}(\gamma)$ with $\gamma > 0$, and $\epsilon_{\text{I}}, \delta > 0$, algorithm $\mathcal{A}$ requires $O\big(\mathsf{poly}(\frac{1}{\gamma}, \frac{1}{\epsilon_{\text{I}}}, \frac{1}{\delta})\big)$ bag-labeled instances sampled independently from the MI generative process in Figure 1(b) to produce an instance p-concept hypothesis h with risk $R_f^{\text{AUC}}(h) < \epsilon_{\text{I}}$ with probability at least $1 - \delta$ over samples.*

Whereas learning accurate instance concepts as in Definition 3 required the use of minimum one-sided disagreement, we show in Theorem 4 that it is possible to learn high-AUC

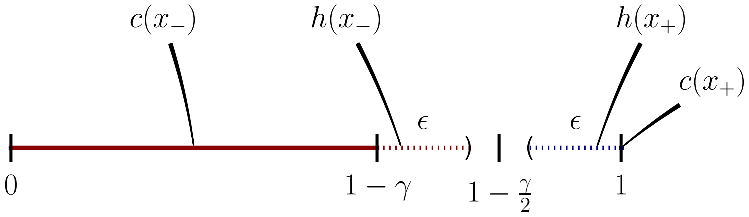


Figure 7: The intuition behind Theorem 4. A hypothesis h that closely approximates c will correctly rank instances with high probability.

concepts using ERM. In particular, the strategy used in the following theorem is to learn a p-concept h that models the concept c defined in Equation 2 using standard ERM. The intuition is that c already achieves perfect AUC; that is, $R_f^{\text{AUC}}(c) = 0$. The reason is that for any negative instance x_- and positive instance x_+, $c(x_-) \leq 1 - \gamma < 1 = c(x_+)$, see Figure 7 for an illustration. If we learn a p-concept h that closely approximates c to within some ϵ, then with high probability, h will also correctly rank instances.

Stating the learnability of a p-concept with ERM requires use of the pseudo-dimension of the concept class $\mathcal{C}$, just as VC dimension can be used to characterize the capacity of a deterministic concept class. The pseudo-dimension is similar to the VC dimension, but uses a different notion of "shattering." In particular, for a set of points with real-valued labels, $\{(x_i, y_i)\}_{i=1}^n$, a p-concept class $\mathcal{C}$ shatters the points if for any binary labeling of the points $\{b_i\}$, there exists some $c \in \mathcal{C}$ such that $c(x_i) \geq y_i$ if $b_i = 1$ and $c(x_i) < y_i$ if $b_i = 0$ (Haussler, 1992). The pseudo-dimension of $\mathcal{C}$, denoted $\text{PD}(\mathcal{C})$, is the size of the largest set such that $\mathcal{C}$ shatters some set of that size.

Theorem 4 *An instance p-concept class $\mathcal{C}$ with pseudo-dimension $\text{PD}(\mathcal{C})$ is Instance MI AUC-PAC-learnable using $O\left(\frac{1}{(\epsilon_{\text{I}}\gamma p)^4}\left(\text{PD}(\mathcal{C})\log\frac{1}{\epsilon_{\text{I}}\gamma p} + \log\frac{1}{\delta}\right)\right)$ examples with standard ERM approaches, where $p = \min\{p_{neg}, 1 - p_{neg}\}$.*

Proof See Appendix A. ■

Comparing Theorem 4 with Theorem 1 on learning accurate instance concepts, we see that neither results require that positive instances appear in positive bags ($\pi > 0$). In both cases, the addition label noise affects γ, but is tolerated by the underlying algorithm. The key difference between these results is that high-AUC concepts can be learned via standard ERM approaches, whereas accurate concept learning requires minimum one-sided disagreement. Additionally, the sample complexity bound in Theorem 4 contains an additional factor p that accounts for class imbalance. Intuitively, this factor appears because it is difficult to learn to effectively rank instances from different classes when one class appears very infrequently in the training set (p is small).

4.2 Learning High-AUC Bag Concepts

As for accuracy, we might be interested in learning either high-AUC instance *or* bag concepts from MI data. Following a similar strategy as employed in Section 3.2 for learning accurate bag concepts, here we will consider two measures of bag-level performance of a bag concept $\widehat{H}$ derived from an instance concept h. The same combining function as in Section 3, $\widehat{H}(X_i) = \max_j h(x_{ij})$, is commonly used to derive real-valued bag-labeling functions in prior work (Ray and Craven, 2005). Following the analysis in Section 3.2, we will measure performance of $\widehat{H}$ with respect to both $\widehat{F}$, the empirical bag-labeling function, and later F, the underlying bag-labeling function.

For the empirical bag-labeling function, $\widehat{F}$, the intuitive definition of AUC is the probability that a bag-level hypothesis $\widehat{H}$ assigns a higher value to a bag sample given that it is labeled positive by $\widehat{F}$ (that is, containing at least one positive instance) than another bag sample labeled negative by $\widehat{F}$ (containing no positive instances). Formally, we can define the corresponding AUC-based risk as follows:

$$\begin{aligned} R^{\mathrm{AUC}}_{\widehat{F}}(\widehat{H}) &= \frac{\begin{array}{c}\int_{\mathcal{B}} \int_{\mathcal{B}} \int_{\mathcal{X}_-^*} \int_{\mathcal{X}_+^*} \mathbb{1}\big[\widehat{H}(X_-) > \widehat{H}(X_+)\big] \dots \\ \dots \,\mathrm{d}\,\mathrm{P}(X_+ \mid B_+)\,\mathrm{d}\,\mathrm{P}(X_- \mid B_-)\,\mathrm{d}\,\mathrm{P}(B_+)\,\mathrm{d}\,\mathrm{P}(B_-)\end{array}}{\mathrm{P}\left[\widehat{F}(X) = 1\right] \mathrm{P}\left[\widehat{F}(X) = 0\right]} \\ &= \frac{\begin{array}{c}\int_{\mathcal{B}} \int_{\mathcal{B}} \int_{\mathcal{X}_-^*} \int_{\mathcal{X}_+^*} \mathbb{1}\big[\widehat{H}(X_-) > \widehat{H}(X_+)\big] \dots \\ \dots \,\mathrm{d}\,\mathrm{P}(X_+ \mid B_+)\,\mathrm{d}\,\mathrm{P}(X_- \mid B_-)\,\mathrm{d}\,\mathrm{P}(B_+)\,\mathrm{d}\,\mathrm{P}(B_-)\end{array}}{(1 - \widehat{P}_{\mathrm{neg}})\widehat{P}_{\mathrm{neg}}}. \end{aligned} \tag{15}$$

Above, $\mathcal{X}_-^*$ is the set of all negative bag samples, and $\mathcal{X}_+^*$ the set of all positive bag samples. The notation $\widehat{P}_{\mathrm{neg}} = \Pr\left[\widehat{F}(X) = 0\right]$ is used for convenience. Now, we can define learnability with respect to this metric:

Definition 8 (Empirical Bag MI AUC-PAC-learning) *We say that an algorithm $\mathcal{A}$ MI AUC-PAC-learns empirical bag-labeling functions derived from p-concept class $\mathcal{C}$ when for any $(D_{\mathcal{X}}, D_{\mathcal{B}}, f, F) \in$ MI-GEN$_{\mathcal{C}}(\gamma)$ with $\gamma > 0$, and $\epsilon_{\mathrm{B}}, \delta > 0$, algorithm $\mathcal{A}$ requires $O\big(\mathsf{poly}(\frac{1}{\gamma}, \frac{1}{\epsilon_{\mathrm{B}}}, \frac{1}{\delta})\big)$ bag-labeled instances sampled independently from the MI generative process in Figure 1(b) to produce an empirical bag-labeling function $\widehat{H}$ with risk $R^{\mathrm{AUC}}_{\widehat{F}}(\widehat{H}) < \epsilon_{\mathrm{B}}$ with probability at least $1 - \delta$ over samples.*

We will now show learnability of empirical bag-labeling functions by reducing the problem to learning an accurate model of the p-concept c. Hence, the approach of the proof follows that for learning accurate empirical bag-labeling functions.

Theorem 5 *Empirical bag-labeling functions derived from p-concept class $\mathcal{C}$ with pseudo-dimension* PD$(\mathcal{C})$ *are AUC-PAC-learnable from MI data using*

$$O\left(\tfrac{m_u^4}{\left(\epsilon_{\mathrm{B}}\gamma\widehat{P}\right)^4}\left(\mathrm{PD}(\mathcal{C}) \log \tfrac{m_u}{\left(\epsilon_{\mathrm{B}}\gamma\widehat{P}\right)} + \log \tfrac{1}{\delta}\right)\right)$$

examples with standard ERM approaches, where

$$\widehat{P} \triangleq \min\{\widehat{P}_{neg}, 1 - \widehat{P}_{neg}\} \geq \min\{P_{neg}, 1 - p_{neg}\},$$

and m_u is an upper bound on bag sample size.

Proof See Appendix A. ■

Note that Theorem 4 is a special case of Theorem 5 when $m_u = 1$. In this case $\widehat{P}_{\text{neg}} = p_{\text{neg}}$ when samples all have size 1, so $\widehat{P} = p$ and the sample complexity is the same.

Now, we can examine AUC-learnability with respect to the true bag-labeling function, F. To define AUC with respect to F, we measure the probability that a sample X_+ is labeled higher by $\widehat{H}$ than X_- is, given that X_+ is sampled from a positive bag and X_- is sampled from a negative bag. Formally, the AUC risk of $\widehat{H}$ with respect to F is

$$\begin{aligned} R_F^{\text{AUC}}(\widehat{H}) &= \frac{\begin{gathered}\int_{\mathcal{B}_-}\int_{\mathcal{B}_+}\int_{\mathcal{X}^*}\int_{\mathcal{X}^*} \mathbb{1}\left[\widehat{H}(X_-) > \widehat{H}(X_+)\right] \dots \\ \dots \,\mathrm{d}\,\mathrm{P}(X_+ \mid B_+)\,\mathrm{d}\,\mathrm{P}(X_- \mid B_-)\,\mathrm{d}\,\mathrm{P}(B_+)\,\mathrm{d}\,\mathrm{P}(B_-)\end{gathered}}{\mathrm{P}\left[F(B) = 1\right]\mathrm{P}\left[F(B) = 0\right]} \\ &= \frac{\begin{gathered}\int_{\mathcal{B}_-}\int_{\mathcal{B}_+}\int_{\mathcal{X}^*}\int_{\mathcal{X}^*} \mathbb{1}\left[\widehat{H}(X_-) > \widehat{H}(X_+)\right] \dots \\ \dots \,\mathrm{d}\,\mathrm{P}(X_+ \mid B_+)\,\mathrm{d}\,\mathrm{P}(X_- \mid B_-)\,\mathrm{d}\,\mathrm{P}(B_+)\,\mathrm{d}\,\mathrm{P}(B_-)\end{gathered}}{(1 - P_{\text{neg}})P_{\text{neg}}}. \end{aligned} \tag{16}$$

The notation $P_{\text{neg}} = \mathrm{P}\left[F(B) = 0\right]$ is used to denote the probability of sampling a negative bag from the distribution over bags. We define the risk to be zero in the case that this probability is equal to either 0 or 1.

As for accuracy, the risk of an empirical bag-labeling function now depends on how representative a sample is of the underlying bag. Thus, in the definition of AUC-learnability with respect to F (Definition 9), we now again require an additional assumption that positive instances appear some $\pi > 0$ fraction of the time in positive bags.

Definition 9 (Bag MI AUC-PAC-learning) *We say that an algorithm $\mathcal{A}$ MI AUC-PAC-learns bag-labeling functions derived from the p-concept class $\mathcal{C}$ when for any tuple $(D_{\mathcal{X}}, D_{\mathcal{B}}, f, F) \in \text{MI-GEN}_{\mathcal{C}}(D_{\mathcal{X}}, f, \gamma, \pi)$ with $\gamma, \pi > 0$, and $\epsilon_{\text{B}}, \delta > 0$, algorithm $\mathcal{A}$ requires $O\big(\mathsf{poly}(\frac{1}{\gamma}, \frac{1}{\pi}, \frac{1}{\epsilon_{\text{B}}}, \frac{1}{\delta})\big)$ bag-labeled instances sampled independently from the MI generative process in Figure 1(b) to produce an empirical bag-labeling function $\widehat{H}$ with risk $R_F^{\text{AUC}}(\widehat{H}) < \epsilon_{\text{B}}$ with probability at least $1 - \delta$ over samples.*

Again, we will learn an instance p-concept h that models c, and then show that a sufficiently accurate p-concept can produce an empirical bag-labeling function $\widehat{H}$ that models F with high AUC. To do this, we will show that the AUC error of $\widehat{H}$ with respect to F, $R_F^{\text{AUC}}(\widehat{H})$, is bounded in terms of the AUC error of $\widehat{H}$ with respect to $\widehat{F}$, $R_{\widehat{F}}^{\text{AUC}}(\widehat{H})$.

Lemma 4 *Suppose bag samples are of size at least m_l ($\forall i : m_l \leq |X_i|$), then $R_F^{\text{AUC}}(\widehat{H}) \leq \frac{1}{P_{neg}} R_{\widehat{F}}^{\text{AUC}}(\widehat{H}) + (1 - \pi)^{m_l}$.*

Proof See Appendix A. ■

Finally, given the bound in Lemma 4, we can derive a result on learning high-AUC bag concepts with respect to the underlying bag-labeling function F. As with the results in Theorem 3, we state the results conditioned on the fact that bag sizes m_i respect some constraints to account for the error that naturally results from insufficiently large samples of instances in positive bags.

Theorem 6 *Bag-labeling functions derived from p-concept class $\mathcal{C}$ with pseudo-dimension* $\mathrm{PD}(\mathcal{C})$ *are AUC-PAC-learnable from MI data using*

$$O\left(\frac{1}{\left(\epsilon_{\mathrm{B}}^{2}\gamma\pi\widehat{P}P_{neg}\right)^{4}}\left(\mathrm{PD}(\mathcal{C})\log\frac{1}{\left(\epsilon_{\mathrm{B}}\gamma\pi\widehat{P}P_{neg}\right)}+\log\frac{1}{\delta}\right)\right). \tag{17}$$

examples using standard ERM approaches when bag sample sizes are bounded by $m_l \leq m \leq m_u$ *and* $m_l \geq \frac{1}{\pi}\log\frac{2}{\epsilon_{\mathrm{B}}}$, *where* $\widehat{P} = \min\{\widehat{P}_{neg}, 1 - \widehat{P}_{neg}\}$.

Proof See Appendix A. ■

4.3 Discussion

The results on AUC learnability in the MI setting are surprising, because they imply the testable hypothesis that *standard supervised approaches* can be used to learn about instance and bag labels in the MI setting. The work that introduced the MI setting evaluated the performance, in terms of *accuracy*, of supervised approaches on MI problems and found them to perform poorly (Dietterich et al., 1997). Later empirical work found that supervised algorithms actually performed quite well on MI problems, *in terms of AUC* (Ray and Craven, 2005). This apparent discrepancy can be explained with the results in this section. Supervised approaches can perform well in terms of AUC on MI problems, but not, it seems, with respect to accuracy.

While the results in Section 3 do not formally show that supervised approaches cannot learn high-accuracy concepts, we conjecture that this is the case due to the one-sided noise inherent in learning to discriminate classes. As illustrated in Figure 7, the fact that negative instances appear some $\gamma > 0$ fraction of the time in negative bags means that learning to approximate c can be used to rank instances. However, accurately labeling instances using an approximation of c requires choosing a *threshold* to discriminate between positive and negative instances. If the value of γ were known in advance, then such a threshold might be selected at $1-\frac{\gamma}{2}$, for example. However, without knowledge of γ or other further assumptions about the generative process, proving that such a threshold might be selected accurately is a direction for future work.

5. Empirical Evaluation

Because the results in this work imply the surprising fact that standard supervised algorithms can be used to learn concepts with high-AUC, but not high accuracy, from MI data, we explicitly evaluate this hypothesis using real-world MI data sets. As always, there are some differences between theory and practice that might confound the experimental results. Below we first explain these two key differences and argue why they do not threaten the validity of our experimental results. Then, we discuss the remainder of our experimental methodology and results.

5.1 Single Instance Learning

In these experiments, we use single-instance learning (SIL) to apply supervised algorithms to MI data. The SIL procedure takes an MI data set and applies to every instance the label of its bag. Hence, like in the generative model described in Section 2 used to show the theoretical results in this work, the SIL training set consists of bag-labeled instances. However, unlike in the generative model used in this work, SIL samples more than one instance per bag. As a result, SIL potentially introduces some "correlation" between instances in the training set. Figure 1 provies a comparison of the generative model of Section 2 (Figure 1(b)) and that of SIL (Figure 1(c)).

We could make SIL more closely resemble our generative model by randomly discarding all but one instance in every bag. However, this would dramatically reduce the size of most practical MI data sets and would needlessly "throw away" the information associated with the discarded instances. Instead, we use all instances in the data set, and ignore the fact that they are potentially correlated, thereby assuming that every instance is sampled from an independent bag. The correlation could change the training distribution over instances, but this should *hurt* the performance of the supervised algorithm if it has any significant effect at all. Therefore, comparing SIL to MI-specific algorithms provides a comparison that is fair to the MI approaches.

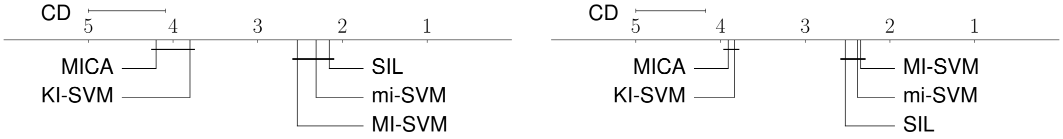
5.2 Risk Minimization Approaches

The results on AUC learnability for MI data use results on learning via empirical risk minimization (ERM). ERM requires that some concept class $\mathcal{C}$ is fixed in advance, and a hypothesis $h \in \mathcal{C}$ that minimizes empirical risk (in terms of accuracy) is selected. In practice, however, $\mathcal{C}$ might not be known *a priori*. Thus, structural risk minimization (SRM) strategies are often used in practice, which simultaneously select a hypothesis h that minimizes empirical risk while controlling the capacity of the class from which h is selected. The standard support vector machine (SVM) is an SRM approach, where the parameter C, selected via cross-validation, controls the trade-off between risk minimization and regularization. Although our theoretical result holds for ERM, we will use the SRM-based SVM for these experiments. The same SRM strategy is used across all of the baseline algorithms.

Similarly, the SVM outputs confidence values that range from $(-\infty, \infty)$ rather than from $[0, 1]$. Thus, the SVM technically does not learn a p-concept. However, prior work has shown how it is possible to fit a logistic regression model to an SVM's outputs to derive associated probabilities (Platt, 1999). However, since rescaling the data does not affect the relative rankings of the real-valued outputs produced by the SVM, the AUC of the classifier does not change. Accordingly, we report results using the raw confidence values produced by the SVM in these experiments.

5.3 Methodology

To evaluate our hypothesis that a supervised SVM can perform well with respect to AUC for learning instance- and bag-labeling functions, we use a total of 55 real-world data sets across a variety of problem domains, including 3D-QSAR (Dietterich et al., 1997), CBIR (Andrews et al., 2003; Maron and Ratan, 1998; Rahmani et al., 2005), text categorization (Andrews et al., 2003; Settles et al., 2008), and audio classification (Briggs et al., 2012). Of the 55 data sets, 45 of them have instance labels, which are only used to test the instance-level performance of classifiers, not for training.



(a) Instance AUC

(b) Bag AUC

Figure 8: The average ranks (lower is better) of approaches on the instance- and bag-labeling tasks, evaluated using AUC. Statistically insignificant differences in performance are indicated with horizontal lines.

The SIL approach combined with a standard supervised SVM is compared with four popular baseline MI SVM approaches: mi-SVM, MI-SVM (Andrews et al., 2003), MICA (Mangasarian and Wild, 2008), and the "instance" variant of KI-SVM (Li et al., 2009), which have been specifically designed to learn bag or instance labels from MI data. Prior empirical results show that these approaches constitute the state-of-the-art in instance-based MI SVM approaches (Doran and Ray, 2013).

The experiments used for this work were implemented in Python using NumPy (Ascher et al., 2001) and SciPy (Jones et al., 2001) for general matrix computations and the CVXOPT library (Dahl and Vandenberghe, 2009) for solving quadratic programs (QPs). We use the authors' original MATLAB code, found at `http://lamda.nju.edu.cn/code_KISVM.ashx`, for the key instance SVM (KI-SVM) approach (Liu et al., 2012). We evaluate algorithms using 10-fold stratified cross-validation, with 5-fold inner-validation used to select parameters using random search (Bergstra and Bengio, 2012). Parameter selection is performed with respect to bag-level labels (since instance-level labels are unavailable at training time, even during cross-validation). We use the radial basis function (RBF) kernel with all algorithms, with scale parameter $\gamma \in [10^{-6}, 10^{1}]$, and regularization–loss trade-off parameter $C \in [10^{-2}, 10^{5}]$. The L_2 norm is used for regularization in all algorithms.

To statistically compare the classifiers, we use the approach described by Demšar (2006). We use the nonparametric Friedman test to reject the null hypothesis that the algorithms perform equally at an $\alpha = 0.001$ significance level. Finally, we plot the average ranks using a *critical difference* diagram, which uses the Nemenyi test to identify statistically equivalent groups of classifiers at an $\alpha = 0.05$ significance level.

5.4 Results and Discussion

The results are summarized using critical difference diagrams in Figure 8. Using AUC to measure performance, the ranks of the approaches are averaged across the 45 instance-labeled data sets for the instance-level metrics and across the 55 data sets for the bag-level metrics. Lower ranks indicate better performance. Full results can be found in Table 4 and Table 5.

Prior work has found that with respect to accuracy, the naïve SIL approach applied to MI data does not perform well (Dietterich et al., 1997; Doran and Ray, 2014). On the other hand, with respect to AUC, the relative performance of SIL increases significantly, and SIL

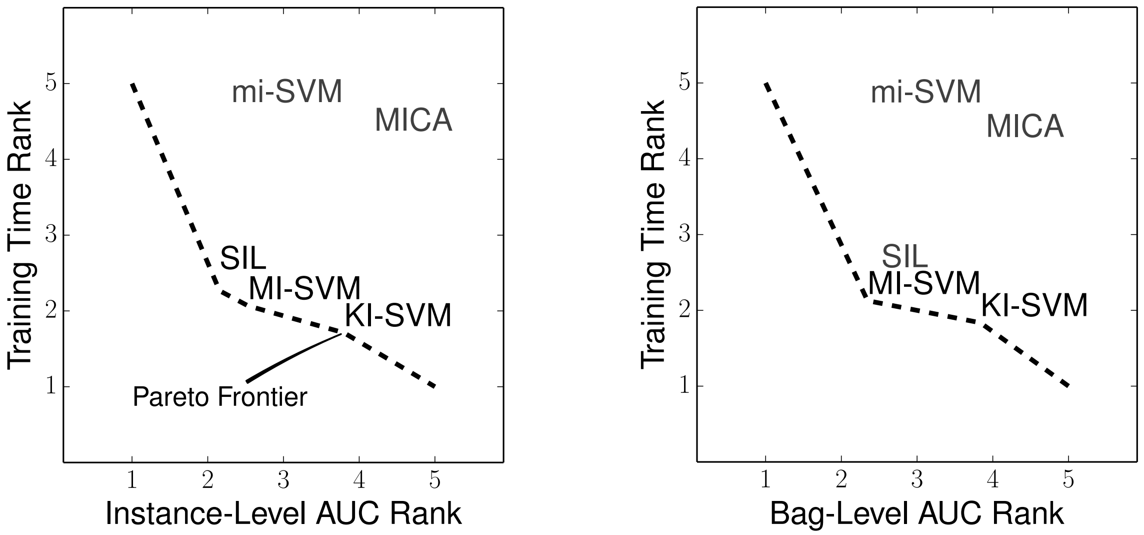


Figure 9: Comparison of supervised and MI-specific approaches in terms of running time and classification performance (AUC). Lower ranks correspond to better performance and faster training time. The Pareto frontier shows algorithms that are not dominated by any other algorithm along both dimensions.

performs as well the best MI approaches. For instance-level AUC, SIL is the highest-ranked approach. For bag-level AUC, SIL is not the best approach on average, but it is statistically equivalent to the top MI approaches. Since more samples are required to learn a bag-level concept using SIL, it could be that performance would improve even more with a larger training sample.

These surprising results support the theoretical framework described in Section 2. In particular, the experimental results suggest that the assumptions made by our generative model hold in practice in many cases. For example, we claim that the assumption that negative instances appear in negative bags ($\gamma > 0$) is weak and reasonable for many MI domains. In CBIR, negative background segments are likely to appear at least some of the time in images without the object of interest. The experimental results provide empirical support for this claim across the four domains on which we evaluate the classifiers. Of course, there might be domains for which this assumption does not hold. Determining whether any learnability results can be derived under weaker assumptions is an interesting question for future work.

There are also several ways that the theoretical and empirical results in this work can inform future work on MI learning. As mentioned earlier, the first work on MIL used accuracy as a performance measure, and found the SIL approach to be inaccurate in the MI setting for labeling bags (Dietterich et al., 1997). As a result, subsequent studies rarely used it as a baseline when evaluating new MI techniques. However, our results suggest that SIL should be used as a baseline when evaluating new MI approaches, especially if the intended application involves ranking bags or instances.

For researchers looking to learn high-AUC instance concepts from MI data, our results suggest that supervised approaches often suffice for this purpose in practice. Since supervised approaches are typically more computationally efficient than their MI counterparts, our theoretical and empirical justification for using supervised approaches with MI data provides a valuable practical benefit. The results in Figure 9 support this claim. The training time required by the algorithms for each data set is ranked with 1 corresponding to the fastest algorithm, and these ranks are averaged across data sets. Then, the combined performance of each approach in terms of both AUC and training time is shown in Figure 9. The Pareto frontier, the set of algorithms for which there does not exist any other algorithm that has better performance along both dimensions, is indicated in the figure. SIL is at or near the Pareto frontier for both instance- and bag-labeling.

For learning high-AUC bag-labeling concepts, MI algorithms still had a slight (but statistically insignificant) advantage over SIL in terms of classifier performance and training time. However, we have observed that even better performance can be attained by applying standard supervised approaches directly to the bag-level learning task using kernel methods (Doran and Ray, 2013).

Table 4: Instance-level AUC results for Section 5.4. The best result is indicated in boldface.

Dataset	SIL	MI-SVM	mi-SVM	KI-SVM	MICA
SIVAL01	0.758	0.872	0.836	0.758	**0.898**
SIVAL02	**0.867**	0.841	0.782	0.761	0.815
SIVAL03	0.676	0.588	0.795	0.690	**0.934**
SIVAL04	0.647	0.651	**0.859**	0.595	0.836
SIVAL05	0.954	0.810	**0.961**	0.906	0.754
SIVAL06	0.619	0.489	0.603	0.516	**0.703**
SIVAL07	0.895	0.784	**0.903**	0.780	0.725
SIVAL08	**0.868**	0.852	0.768	0.556	0.759
SIVAL09	**0.829**	0.730	0.824	0.771	0.581
SIVAL10	0.882	0.788	**0.948**	0.686	0.721
SIVAL11	**0.965**	0.795	0.952	0.746	0.600
SIVAL12	0.566	0.541	0.515	**0.690**	0.623
Newsgroups01	**0.980**	0.953	0.968	0.834	0.584
Newsgroups02	**0.904**	0.899	0.864	0.850	0.572
Newsgroups03	**0.866**	0.783	0.782	0.686	0.576
Newsgroups04	**0.923**	0.883	0.885	0.846	0.612
Newsgroups05	**0.951**	0.922	0.906	0.796	0.537
Newsgroups06	0.946	**0.948**	0.895	0.824	0.587
Newsgroups07	**0.907**	0.853	0.835	0.827	0.604
Newsgroups08	0.753	0.881	**0.909**	0.808	0.551
Newsgroups09	0.711	0.962	**0.979**	0.869	0.560
Newsgroups10	0.660	**0.947**	0.908	0.746	0.565
Newsgroups11	0.728	0.971	**0.980**	0.968	0.702
Newsgroups12	0.958	**0.961**	0.942	0.767	0.536

continued...

Table 4: Instance-level AUC results (continued).

Dataset	SIL	MI-SVM	mi-SVM	KI-SVM	MICA
Newsgroups13	**0.970**	0.939	0.911	0.920	0.608
Newsgroups14	0.823	**0.903**	0.884	0.902	0.614
Newsgroups15	0.736	0.949	**0.955**	0.930	0.553
Newsgroups16	0.454	0.938	0.906	**0.940**	0.600
Newsgroups17	**0.946**	0.913	0.921	0.854	0.565
Newsgroups18	**0.964**	0.914	0.922	0.797	0.605
Newsgroups19	**0.931**	0.914	0.826	0.803	0.558
Newsgroups20	0.573	0.884	**0.914**	0.912	0.556
Birdsong01	0.762	**0.925**	0.907	0.704	0.708
Birdsong02	**0.895**	0.884	0.849	0.574	0.748
Birdsong03	**0.782**	0.729	0.673	0.636	0.599
Birdsong04	**0.966**	0.927	0.932	0.905	0.858
Birdsong05	**0.686**	0.439	0.641	0.422	0.498
Birdsong06	0.627	**0.741**	0.581	0.540	0.719
Birdsong07	0.782	0.570	**0.857**	0.441	0.814
Birdsong08	**0.836**	0.615	0.796	0.552	0.774
Birdsong09	0.920	**0.940**	0.915	0.889	0.702
Birdsong10	0.858	0.859	**0.879**	0.763	0.757
Birdsong11	**0.989**	0.971	0.970	0.982	0.712
Birdsong12	**0.954**	0.907	0.918	0.490	0.745
Birdsong13	0.799	0.640	**0.806**	0.605	0.589

Table 5: Bag-level AUC results for Section 5.4. The best result is indicated in boldface.

Dataset	SIL	MI-SVM	mi-SVM	KI-SVM	MICA
musk1	0.922	0.845	**0.943**	0.836	0.849
musk2	0.897	**0.949**	0.661	0.665	0.913
elephant	**0.919**	0.912	0.916	0.676	0.871
fox	**0.662**	0.589	0.632	0.500	0.615
tiger	**0.859**	0.856	0.853	0.673	0.688
field	**0.923**	0.871	0.908	0.687	0.847
flower	0.907	0.873	**0.921**	0.810	0.759
mountain	0.916	0.915	**0.935**	0.759	0.830
SIVAL01	0.626	**0.954**	0.875	0.643	0.933
SIVAL02	0.826	**0.952**	0.731	0.747	0.863
SIVAL03	0.785	0.666	0.716	0.708	**0.906**
SIVAL04	0.657	0.683	0.831	0.697	**0.957**
SIVAL05	0.985	0.964	**1.000**	0.938	0.914
SIVAL06	0.648	0.756	0.753	0.542	**0.918**
SIVAL07	0.793	**0.993**	0.972	0.969	0.974

continued. . .

Table 5: Bag-level AUC results (continued).

Dataset	SIL	MI-SVM	mi-SVM	KI-SVM	MICA
SIVAL08	0.874	**0.998**	0.812	0.488	0.865
SIVAL09	0.907	**0.979**	0.822	0.768	0.709
SIVAL10	0.819	0.772	**0.930**	0.643	0.785
SIVAL11	0.981	**1.000**	0.987	0.817	0.736
SIVAL12	0.601	0.621	0.516	0.692	**0.710**
Newsgroups01	0.928	**0.931**	0.870	0.746	0.535
Newsgroups02	0.873	0.794	**0.878**	0.826	0.538
Newsgroups03	0.755	0.715	**0.805**	0.640	0.517
Newsgroups04	0.765	**0.767**	0.727	0.631	0.511
Newsgroups05	0.776	**0.842**	0.760	0.800	0.538
Newsgroups06	0.814	**0.862**	0.837	0.741	0.521
Newsgroups07	0.802	0.798	0.789	**0.844**	0.576
Newsgroups08	0.674	0.810	**0.837**	0.759	0.532
Newsgroups09	0.728	**0.925**	0.918	0.784	0.552
Newsgroups10	0.752	**0.904**	0.900	0.696	0.550
Newsgroups11	0.709	**0.975**	0.957	0.816	0.669
Newsgroups12	0.844	0.805	**0.858**	0.695	0.530
Newsgroups13	**0.971**	0.911	0.914	0.930	0.602
Newsgroups14	0.648	0.825	0.861	**0.869**	0.600
Newsgroups15	0.742	0.886	0.913	**0.928**	0.555
Newsgroups16	0.564	**0.860**	0.538	0.838	0.543
Newsgroups17	0.630	**0.787**	0.751	0.674	0.533
Newsgroups18	0.864	**0.874**	0.797	0.771	0.558
Newsgroups19	0.754	**0.812**	0.745	0.640	0.548
Newsgroups20	0.575	**0.807**	0.757	0.770	0.520
OHSUMED1	**0.958**	0.914	0.954	0.779	0.816
OHSUMED2	0.740	0.638	**0.775**	0.520	0.715
Birdsong01	0.894	0.974	**0.976**	0.807	0.824
Birdsong02	**0.902**	0.895	0.895	0.570	0.827
Birdsong03	0.908	**0.916**	0.909	0.747	0.755
Birdsong04	**0.999**	0.998	0.989	0.992	0.945
Birdsong05	0.664	0.623	0.542	0.529	**0.699**
Birdsong06	0.926	0.915	0.964	0.728	**0.975**
Birdsong07	0.931	0.928	**0.934**	0.617	0.919
Birdsong08	0.913	0.901	**0.921**	0.723	0.908
Birdsong09	**0.984**	0.941	0.962	0.897	0.787
Birdsong10	0.947	**0.975**	0.956	0.819	0.862
Birdsong11	0.991	0.988	0.991	**0.995**	0.803
Birdsong12	**0.996**	0.961	0.993	0.609	0.737
Birdsong13	0.980	0.948	**0.985**	0.885	0.780

6. Conclusion

In this work, we describe a new generative model for the MI setting in which bags are viewed as distributions over instances. The sets of instances observed in a training sample are then viewed as samples from each underlying bag distribution. We then introduce several additional assumptions that we show entail instance and bag concept learnability. We discuss the relationship between the proposed model and those found in prior work.

Next, we describe new positive learnability results for learning instance or bag concepts from data generated by MI-GEN. We describe how our generative model allows for learnability while excluding scenarios used to show hardness under other generative models. Nevertheless, our generative process extends prior results on instance concept learnability that are over a decade old (Blum and Kalai, 1998). We also show that MI-GEN can incorporate the non-IID instance assumption within bag-specific distributions over instances, so assuming that samples from individual bags are drawn independently is not a restrictive assumption in our model.

Finally, we argue that for many real-world applications of MIL, it is sufficient to *rank* instances or bags rather than assign accurate binary labels. Accordingly, we derived results demonstrating the ability to learn high-AUC rankings of instances or bags from data generated by a process in MI-GEN. The surprising aspect of these results is that such rankings can be found via *standard supervised approaches*. We evaluate this surprising hypothesis empirically and find that supervised approaches *can* in fact learn to rank from MI data in practice. Thus, the empirical results support the assumptions made by MI-GEN.

Our work provides a starting point for many future investigations of learning in the MI framework. For example, we plan to extend our learnability results to the multi-class and multi-label settings. Furthermore, we plan to investigate generalizations of our generative model that allow for other previously studied instance- and bag-label relationships (Scott et al., 2005). Finally, some recent work has investigated learnability of real-valued bag-level concepts under a similar generative process for MI regression (Szabó et al., 2015). We plan to investigate instance-level learning in the MI regression setting.

Acknowledgements

We thank the anonymous reviewers for their feedback. G. Doran was supported by GAANN grant P200A090265 from the US Department of Education and NSF grant CNS-1035602. S. Ray was partially supported by CWRU award OSA110264. This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University.

Appendix A. Detailed Proofs

Lemma 1 *Let $R_f(g)$ be the risk of an instance labeling concept g, and $R_{\widehat{F}}(\widehat{G})$ be the risk of the empirical bag-labeling function $\widehat{G}(X_i) = \max_j g(x_{ij})$. Then if bag sample sizes are bounded by m_u ($\forall i : |X_i| \le m_u$), $R_{\widehat{F}}(\widehat{G}) \le m_u R_f(g)$.*

Proof First, observe that when all elements of an empirical bag X_i are labeled correctly by g, $\widehat{F}(X_i) = \widehat{G}(X_i)$, so when $\widehat{F}(X_i) \neq \widehat{G}(X_i)$, at least one instance in X_i is labeled incorrectly by g. In set notation, this implication is equivalent to the statement

$$\left\{X_i : \widehat{F}(X_i) \neq \widehat{G}(X_i)\right\} \subseteq \left\{X_i : \big(f(x_{i1}) \neq g(x_{i1})\big) \vee \ldots \vee \big(f(x_{im}) \neq g(x_{im})\big)\right\}.$$

Using indicator function ($\mathbb{1}[\cdot]$) notation, the statement above implies

$$\begin{aligned}
\mathbb{1}\big[\widehat{F}(X_i) \neq \widehat{G}(X_i)\big] &\leq \mathbb{1}\big[\big(f(x_{i1}) \neq g(x_{i1})\big) \vee \ldots \vee \big(f(x_{im}) \neq g(x_{im})\big)\big] \\
&= \mathbb{1}\left[\textstyle\bigvee_{x_{ij} \in X_i} \big(f(x_{ij}) \neq g(x_{ij})\big)\right] \\
&\leq \sum_{x_{ij} \in X_i} \mathbb{1}\left[f(x_{ij}) \neq g(x_{ij})\right].
\end{aligned}$$

Using this inequality in the definition of risk for empirical bag-labeling functions (Equation 10) yields

$$\begin{aligned}
R_{\widehat{F}}(\widehat{G}) &= \int_{\mathcal{B}} \int_{\mathcal{X}^*} \mathbb{1}\big[\widehat{F}(X_i) \neq \widehat{G}(X_i)\big] \,\mathrm{d}\,\mathrm{P}(X_i \mid B) \,\mathrm{d}\,\mathrm{P}(B) \\
&\leq \int_{\mathcal{B}} \int_{\mathcal{X}^*} \sum_{x_{ij} \in X_i} \mathbb{1}\left[f(x_{ij}) \neq g(x_{ij})\right] \,\mathrm{d}\,\mathrm{P}(X_i \mid B) \,\mathrm{d}\,\mathrm{P}(B).
\end{aligned}$$

By the independence of the instances $x_{ij} \in X_i$, and the bound m_u on bag sample sizes, we can rewrite the inner integral to conclude that

$$\begin{aligned}
R_{\widehat{F}}(\widehat{G}) &\leq \int_{\mathcal{B}} m_u \int_{\mathcal{X}} \mathbb{1}\left[f(x) \neq g(x)\right] \,\mathrm{d}\,\mathrm{P}(x \mid B) \,\mathrm{d}\,\mathrm{P}(B) \\
&= m_u \int_{\mathcal{X}} \mathbb{1}\left[f(x) \neq g(x)\right] \,\mathrm{d}\,\mathrm{P}(x) \\
&= m_u R_f(g).
\end{aligned}$$

Exchanging the order of the integrals and marginalizing out the individual bag distributions to obtain an integral with respect to the instance distribution follows from Condition 1 in Definition 1. ■

Lemma 2 *For any empirical bag-labeling concept $\widehat{G}$,*

$$R_F(\widehat{G}) \leq R_{\widehat{F}}(\widehat{G}) + R_F(\widehat{F}).$$

Proof First, note that if $\widehat{G}(X) = \widehat{F}(X)$ and $\widehat{F}(X) = F(B)$, then $\widehat{G}(X) = F(B)$. Thus, if $\widehat{G}(X) \neq F(B)$, then either $\widehat{G}(X) \neq \widehat{F}(X)$ or $\widehat{F}(X) \neq F(B)$. In set notation, this is equivalent to the statement

$$\left\{(X, B) : \widehat{G}(X) \neq F(B)\right\} \subseteq \left\{(X, B) : \left(\widehat{G}(X) \neq \widehat{F}(X)\right) \vee \left(\widehat{F}(X) \neq F(B)\right)\right\}.$$

Using indicator function notation, the statement above implies

$$\begin{aligned}
\mathbb{1}\big[\widehat{G}(X) \neq F(B)\big] &\leq \mathbb{1}\big[\big(\widehat{G}(X) \neq \widehat{F}(X)\big) \vee \big(\widehat{F}(X) \neq F(B)\big)\big] \\
&\leq \mathbb{1}\big[\big(\widehat{G}(X) \neq \widehat{F}(X)\big)\big] + \mathbb{1}\big[\big(\widehat{F}(X) \neq F(B)\big)\big].
\end{aligned}$$

Finally, substituting the expression above into the definitions of risk yields

$$\begin{aligned}
R_F(\widehat{G}) &= \int_{\mathcal{B}} \int_{\mathcal{X}^*} \mathbb{1}\big[\widehat{G}(X) \neq F(B)\big] \, \mathrm{d}\,\mathrm{P}(X \mid B) \, \mathrm{d}\,\mathrm{P}(B) \\
&\leq \int_{\mathcal{B}} \int_{\mathcal{X}^*} \mathbb{1}\big[\big(\widehat{G}(X) \neq \widehat{F}(X)\big)\big] \, \mathrm{d}\,\mathrm{P}(X \mid B) \, \mathrm{d}\,\mathrm{P}(B) \\
&\quad + \int_{\mathcal{B}} \int_{\mathcal{X}^*} \mathbb{1}\big[\big(\widehat{F}(X) \neq F(B)\big)\big] \, \mathrm{d}\,\mathrm{P}(X \mid B) \, \mathrm{d}\,\mathrm{P}(B) \\
&= R_{\widehat{F}}(\widehat{G}) + R_F(\widehat{F}).
\end{aligned}$$

■

Lemma 3 *Suppose bag samples are of size at least m_l ($\forall i : m_l \leq |X_i|$), then $R_F(\widehat{F}) \leq (1-\pi)^{m_l}$.*

Proof Given the definition of $R_F(\widehat{F})$, we can decompose it as such

$$\begin{aligned}
R_F(\widehat{F}) &= \int_{\mathcal{B}} \int_{\mathcal{X}^*} \mathbb{1}[F(B_i) \neq \widehat{F}(X_i)] \, \mathrm{d}\,\mathrm{P}(X_i \mid B_i) \, \mathrm{d}\,\mathrm{P}(B_i) \\
&= \int_{\mathcal{B}_+} \int_{\mathcal{X}^*} \mathbb{1}[F(B_i) \neq \widehat{F}(X_i)] \, \mathrm{d}\,\mathrm{P}(X_i \mid B_i) \, \mathrm{d}\,\mathrm{P}(B_i) \\
&\quad + \int_{\mathcal{B}_-} \int_{\mathcal{X}^*} \mathbb{1}[F(B_i) \neq \widehat{F}(X_i)] \, \mathrm{d}\,\mathrm{P}(X_i \mid B_i) \, \mathrm{d}\,\mathrm{P}(B_i).
\end{aligned}$$

On the set of negative bags $\mathcal{B}_-$, F and $\widehat{F}$ always agree, since only negative instances are sampled within negative bags. Therefore, the second term of the decomposition can be eliminated and we are left with

$$R_F(\widehat{F}) = \int_{\mathcal{B}_+} \int_{\mathcal{X}^*} \mathbb{1}[F(B_i) \neq \widehat{F}(X_i)] \, \mathrm{d}\,\mathrm{P}(X_i \mid B_i) \, \mathrm{d}\,\mathrm{P}(B_i).$$

Now, we observe that for a positive bag B_i, the only way that F and $\widehat{F}$ can disagree is if every instance in X_i is negative. Using basic properties of indicator functions (namely, that $\mathbb{1}\left[\bigwedge_i E_i\right] = \prod_i \mathbb{1}[E_i]$), we can use this fact to rewrite the expression above as

$$\begin{aligned}
R_F(\widehat{F}) &= \int_{\mathcal{B}_+} \int_{\mathcal{X}^*} \mathbb{1}[F(B_i) \neq \widehat{F}(X_i)] \, \mathrm{d}\,\mathrm{P}(X_i \mid B_i) \, \mathrm{d}\,\mathrm{P}(B_i) \\
&= \int_{\mathcal{B}_+} \int_{\mathcal{X}^*} \mathbb{1}\left[\textstyle\bigwedge_{x_{ij} \in X_i} \big(f(x_{ij}) = 0\big)\right] \, \mathrm{d}\,\mathrm{P}(X_i \mid B_i) \, \mathrm{d}\,\mathrm{P}(B_i) \\
&= \int_{\mathcal{B}_+} \int_{\mathcal{X}^*} \prod_{x_{ij} \in X_i} \mathbb{1}\left[f(x_{ij}) = 0\right] \, \mathrm{d}\,\mathrm{P}(X_i \mid B_i) \, \mathrm{d}\,\mathrm{P}(B_i).
\end{aligned}$$

Since the instances $x_{ij} \in X_i$ are independent, we can rewrite the integral as

$$\begin{aligned}
R_F(\widehat{F}) &= \int_{\mathcal{B}_+} \int_{\mathcal{X}^*} \prod_{x_{ij} \in X_i} \mathbb{1}\left[f(x_{ij}) = 0\right] \mathrm{d}\,\mathrm{P}(X_i \mid B_i)\, \mathrm{d}\,\mathrm{P}(B_i) \\
&= \int_{\mathcal{B}_+} \prod_{x_{ij} \in X_i} \left(\int_{\mathcal{X}} \mathbb{1}\left[f(x_{ij}) = 0\right] \mathrm{d}\,\mathrm{P}(x_{ij} \mid B_i) \right) \mathrm{d}\,\mathrm{P}(B_i) \\
&\leq \int_{\mathcal{B}_+} \prod_{x_{ij} \in X_i} (1 - \pi)\ \mathrm{d}\,\mathrm{P}(B_i) \\
&\leq \int_{\mathcal{B}_+} (1 - \pi)^{m_l}\, \mathrm{d}\,\mathrm{P}(B_i) \\
&= (1 - \pi)^{m_l} \int_{\mathcal{B}_+} \mathrm{d}\,\mathrm{P}(B_i) \\
&\leq (1 - \pi)^{m_l}.
\end{aligned}$$

■

Theorem 4 *An instance p-concept class $\mathcal{C}$ with pseudo-dimension* $\mathrm{PD}(\mathcal{C})$ *is Instance MI AUC-PAC-learnable using* $O\left(\frac{1}{(\epsilon_{\mathrm{I}}\gamma p)^4}\left(\mathrm{PD}(\mathcal{C}) \log \frac{1}{\epsilon_{\mathrm{I}}\gamma p} + \log \frac{1}{\delta}\right)\right)$ *examples with standard ERM approaches, where* $p = \min\{p_{neg}, 1 - p_{neg}\}$*.*

Proof For any $c \in \mathcal{C}$, we can use ERM with respect to the quadratic loss function to learn a hypothesis h such that $\mathrm{E}\left[(h(x) - c(x))^2\right] < \epsilon$ with probability $1 - \delta$ across samples. By Jensen's inequality, this bounds the expected absolute deviation between h and c:

$$\mathrm{E}\left[\,|h(x) - c(x)|\,\right] \leq \sqrt{\mathrm{E}\left[(h(x) - c(x))^2\right]} < \sqrt{\epsilon}.$$

Then, by Markov's inequality, this expression bounds the probability over examples that $|h(x) - c(x)|$ exceeds some constant t:

$$\mathrm{P}\left[\,|h(x) - c(x)| > t\,\right] \leq \frac{\mathrm{E}\left[\,|h(x) - c(x)|\,\right]}{t} < \frac{\sqrt{\epsilon}}{t}. \tag{18}$$

Therefore, with high probability, $|h(x) - c(x)|$ is small for small ϵ.

Now, we can proceed by following the intuition illustrated in Figure 7. In particular, we will show that the AUC risk is bounded when h and c agree on examples with high probability. First, suppose $|h(x) - c(x)| \leq \frac{\gamma}{2}$ for both of a pair (x_+, x_-) of positive and negative instances. Then for the negative instance, x_-, by Definition 1, Condition 3,

$$h(x_-) \leq c(x_-) + \tfrac{\gamma}{2} \leq (1 - \gamma) + \tfrac{\gamma}{2} = 1 - \tfrac{\gamma}{2}.$$

Similarly, for the positive instance, x_+, by Definition 1, Condition 2,

$$h(x_+) \geq c(x_+) - \tfrac{\gamma}{2} = 1 - \tfrac{\gamma}{2}.$$

Hence, we have that $h(x_-) \leq h(x_+)$.

By contraposition of the conclusion above, if $h(x_-) > h(x_+)$, then it is either the case that $|h(x_-) - c(x_-)| > \frac{\gamma}{2}$ or that $|h(x_+) - c(x_+)| > \frac{\gamma}{2}$. In set theoretic terms, this means

$$\big\{(x_+, x_-) : h(x_-) > h(x_+)\big\} \subseteq \big\{(x_+, x_-) : |h(x_-) - c(x_-)| > \tfrac{\gamma}{2} \vee |h(x_+) - c(x_+)| > \tfrac{\gamma}{2}\big\}$$

In indicator function notation, this implies

$$\begin{aligned} \mathbb{1}\big[h(x_-) > h(x_+)\big] &\leq \mathbb{1}\big[\,|h(x_-) - c(x_-)| > \tfrac{\gamma}{2} \vee |h(x_+) - c(x_+)| > \tfrac{\gamma}{2}\big] \\ &\leq \mathbb{1}\big[\,|h(x_-) - c(x_-)| > \tfrac{\gamma}{2}\big] + \mathbb{1}\big[\,|h(x_+) - c(x_+)| > \tfrac{\gamma}{2}\big]. \end{aligned}$$

Substituting this expression into the definition of $R_f^{\mathrm{AUC}}(h)$ (Equation 14) yields

$$\begin{aligned} R_f^{\mathrm{AUC}}(h) &= \frac{\int_{\mathcal{X}_-}\int_{\mathcal{X}_+} \mathbb{1}[h(x_-) > h(x_+)]\,\mathrm{d}\,\mathrm{P}(x_+)\,\mathrm{d}\,\mathrm{P}(x_-)}{(1 - p_{\mathrm{neg}})p_{\mathrm{neg}}} \\ &\leq \frac{\int_{\mathcal{X}_-}\int_{\mathcal{X}_+} \mathbb{1}\big[\,\big|h(x_-) - c(x_-) > \tfrac{\gamma}{2}\big|\,\big]\,\mathrm{d}\,\mathrm{P}(x_+)\,\mathrm{d}\,\mathrm{P}(x_-)}{(1 - p_{\mathrm{neg}})p_{\mathrm{neg}}} \\ &\quad + \frac{\int_{\mathcal{X}_-}\int_{\mathcal{X}_+} \mathbb{1}\big[\,|h(x_+) - c(x_+)| > \tfrac{\gamma}{2}\big]\,\mathrm{d}\,\mathrm{P}(x_+)\,\mathrm{d}\,\mathrm{P}(x_-)}{(1 - p_{\mathrm{neg}})p_{\mathrm{neg}}} \\ &= \frac{\int_{\mathcal{X}_-} \mathbb{1}\big[\,\big|h(x_-) - c(x_-) > \tfrac{\gamma}{2}\big|\,\big]\,\mathrm{d}\,\mathrm{P}(x_-)}{p_{\mathrm{neg}}} \\ &\quad + \frac{\int_{\mathcal{X}_+} \mathbb{1}\big[\,|h(x_+) - c(x_+)| > \tfrac{\gamma}{2}\big]\,\mathrm{d}\,\mathrm{P}(x_+)}{1 - p_{\mathrm{neg}}}. \end{aligned}$$

Then, using the definition $p = \min\{p_{\mathrm{neg}}, 1 - p_{\mathrm{neg}}\}$, this becomes

$$\begin{aligned} R_f^{\mathrm{AUC}}(h) &\leq \frac{\int_{\mathcal{X}_-} \mathbb{1}\big[\,\big|h(x_-) - c(x_-) > \tfrac{\gamma}{2}\big|\,\big]\,\mathrm{d}\,\mathrm{P}(x_-)}{p} \\ &\quad + \frac{\int_{\mathcal{X}_+} \mathbb{1}\big[\,|h(x_+) - c(x_+)| > \tfrac{\gamma}{2}\big]\,\mathrm{d}\,\mathrm{P}(x_+)}{p} \\ &= \frac{\int_{\mathcal{X}} \mathbb{1}\big[\,\big|h(x) - c(x) > \tfrac{\gamma}{2}\big|\,\big]\,\mathrm{d}\,\mathrm{P}(x)}{p} = \frac{\mathrm{P}\big[\,|h(x) - c(x)| > \tfrac{\gamma}{2}\big]}{p}. \end{aligned}$$

Finally, using the inequality derived in Equation 18, we have

$$R_f^{\mathrm{AUC}}(h) \leq \frac{\mathrm{P}\big[\,|h(x) - c(x)| > \tfrac{\gamma}{2}\big]}{p} < \frac{2\sqrt{\epsilon}}{\gamma p}.$$

Therefore, it is sufficient to choose $\epsilon = \frac{(\epsilon_{\mathrm{I}}\gamma p)^2}{4}$ when learning h via ERM as so that $R_f^{\mathrm{AUC}}(h) < \epsilon_{\mathrm{I}}$.

Finally, the sample complexity bound results from substituting $\epsilon = \frac{(\epsilon_{\mathrm{I}}\gamma p)^2}{4}$ into the existing bound $O\big(\frac{1}{\epsilon^2}\big(\mathrm{PD}(\mathcal{C})\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\big)\big)$ for learning p-concepts using ERM (Kearns and Schapire, 1994). ■

Theorem 5 *Empirical bag-labeling functions derived from p-concept class $\mathcal{C}$ with pseudo-dimension* $\mathrm{PD}(\mathcal{C})$ *are AUC-PAC-learnable from MI data using*

$$O\left(\frac{m_u^4}{\left(\epsilon_{\mathrm{B}}\gamma\widehat{P}\right)^4}\left(\mathrm{PD}(\mathcal{C})\log\frac{m_u}{\left(\epsilon_{\mathrm{B}}\gamma\widehat{P}\right)}+\log\tfrac{1}{\delta}\right)\right)$$

examples with standard ERM approaches, where

$$\widehat{P}\triangleq\min\{\widehat{P}_{neg},1-\widehat{P}_{neg}\}\geq\min\{P_{neg},1-p_{neg}\},$$

and m_u is an upper bound on bag sample size.

Proof As in Theorem 4, we will learn a p-concept h to model c accurately with high probability. Then, given bag samples X_+ with at least one positive instance and X_- with all negative instances, suppose that $|h(x)-c(x)|\leq\frac{\gamma}{2}$ for *all* instances across both samples. Then by the same argument as in Theorem 4 as illustrated in Figure 7, at least one instance in X_+ is assigned a label by h that is at least $1-\frac{\gamma}{2}$, and all instances in X_- are assigned a label by h of at most $1-\frac{\gamma}{2}$. Therefore, the maximum label assigned in X_+, $\widehat{H}(X_+)$, is greater than or equal to the maximum label in X_-, $\widehat{H}(X_-)$.

By contraposition, if $\widehat{H}(X_-)>\widehat{H}(X_+)$, then the label $h(x)$ of *some* instance x in either X_+ or X_- deviates by more than $\frac{\gamma}{2}$ from $c(x)$. That is,

$$\begin{aligned}&\big\{(X_+,X_-):\widehat{H}(X_-)>\widehat{H}(X_+)\big\}\\&\qquad\subseteq\left\{(X_+,X_-):\left(\bigvee_{x\in X_+}|h(x)-c(x)|>\tfrac{\gamma}{2}\right)\vee\left(\bigvee_{x\in X_-}|h(x)-c(x)|>\tfrac{\gamma}{2}\right)\right\}\end{aligned}$$

Therefore, in indicator function notation,

$$\mathbb{1}\big[\widehat{H}(X_-)>\widehat{H}(X_+)\big]\leq\sum_{x\in X_+}\mathbb{1}\big[\,|h(x)-c(x)|>\tfrac{\gamma}{2}\big]+\sum_{x\in X_-}\mathbb{1}\big[\,|h(x)-c(x)|>\tfrac{\gamma}{2}\big].$$

Using the inequality above in the definition of $R^{\mathrm{AUC}}_{\widehat{F}}(\widehat{H})$ in Equation 15 gives

$$\begin{aligned}R^{\mathrm{AUC}}_{\widehat{F}}(\widehat{H})\leq{}&\frac{\begin{gathered}\int_{\mathcal{B}}\int_{\mathcal{B}}\int_{\mathcal{X}^*_-}\int_{\mathcal{X}^*_+}\sum_{x\in X_+}\mathbb{1}\big[\,|h(x)-c(x)|>\tfrac{\gamma}{2}\big]\dots\\\dots\,\mathrm{d}\,\mathrm{P}(X_+\mid B_+)\,\mathrm{d}\,\mathrm{P}(X_-\mid B_-)\,\mathrm{d}\,\mathrm{P}(B_+)\,\mathrm{d}\,\mathrm{P}(B_-)\end{gathered}}{(1-\widehat{P}_{\mathrm{neg}})\widehat{P}_{\mathrm{neg}}}\\&+\frac{\begin{gathered}\int_{\mathcal{B}}\int_{\mathcal{B}}\int_{\mathcal{X}^*_-}\int_{\mathcal{X}^*_+}\sum_{x\in X_-}\mathbb{1}\big[\,|h(x)-c(x)|>\tfrac{\gamma}{2}\big]\dots\\\dots\,\mathrm{d}\,\mathrm{P}(X_+\mid B_+)\,\mathrm{d}\,\mathrm{P}(X_-\mid B_-)\,\mathrm{d}\,\mathrm{P}(B_+)\,\mathrm{d}\,\mathrm{P}(B_-)\end{gathered}}{(1-\widehat{P}_{\mathrm{neg}})\widehat{P}_{\mathrm{neg}}}\end{aligned}$$

Since the integrands above only depend on X_+ and X_-, we can rewrite the expression using the fact that

$$\begin{aligned}\int_{\mathcal{X}^*_-}\mathrm{d}\,\mathrm{P}(X_-\mid B_-)\,\mathrm{d}\,\mathrm{P}(B_-)&=\mathrm{P}\left[\widehat{F}(X)=0\right]=\widehat{P}_{\mathrm{neg}}\\\int_{\mathcal{X}^*_+}\mathrm{d}\,\mathrm{P}(X_+\mid B_+)\,\mathrm{d}\,\mathrm{P}(B_+)&=\mathrm{P}\left[\widehat{F}(X)=1\right]=1-\widehat{P}_{\mathrm{neg}}.\end{aligned}$$

The result is

$$\begin{aligned}
R^{\mathrm{AUC}}_{\widehat{F}}(\widehat{H}) &\leq \frac{\int_{\mathcal{B}} \int_{\mathcal{X}^*_+} \sum_{x \in X_+} \mathbb{1}\left[\, |h(x) - c(x)| > \frac{\gamma}{2}\right] \mathrm{d}\,\mathrm{P}(X_+ \mid B_+)\,\mathrm{d}\,\mathrm{P}(B_+)}{(1 - \widehat{P}_{\mathrm{neg}})} \\
&\quad + \frac{\int_{\mathcal{B}} \int_{\mathcal{X}^*_-} \sum_{x \in X_-} \mathbb{1}\left[\, |h(x) - c(x)| > \frac{\gamma}{2}\right] \mathrm{d}\,\mathrm{P}(X_- \mid B_-)\,\mathrm{d}\,\mathrm{P}(B_-)}{\widehat{P}_{\mathrm{neg}}} \\
&\leq \frac{\int_{\mathcal{B}} \int_{\mathcal{X}^*_+} \sum_{x \in X_+} \mathbb{1}\left[\, |h(x) - c(x)| > \frac{\gamma}{2}\right] \mathrm{d}\,\mathrm{P}(X_+ \mid B_+)\,\mathrm{d}\,\mathrm{P}(B_+)}{\widehat{P}} \\
&\quad + \frac{\int_{\mathcal{B}} \int_{\mathcal{X}^*_-} \sum_{x \in X_-} \mathbb{1}\left[\, |h(x) - c(x)| > \frac{\gamma}{2}\right] \mathrm{d}\,\mathrm{P}(X_- \mid B_-)\,\mathrm{d}\,\mathrm{P}(B_-)}{\widehat{P}} \\
&= \frac{\int_{\mathcal{B}} \int_{\mathcal{X}^*} \sum_{x \in X} \mathbb{1}\left[\, |h(x) - c(x)| > \frac{\gamma}{2}\right] \mathrm{d}\,\mathrm{P}(X \mid B)\,\mathrm{d}\,\mathrm{P}(B)}{\widehat{P}}.
\end{aligned}$$

By the independence of instances $x \in X$, the upper bound m_u on bag size, and Condition 1 in Definition 1, we can rewrite the expression above as

$$\begin{aligned}
R^{\mathrm{AUC}}_{\widehat{F}}(\widehat{H}) &\leq \frac{m_u}{\widehat{P}} \int_{\mathcal{B}} \int_{\mathcal{X}} \mathbb{1}\left[\, |h(x) - c(x)| > \frac{\gamma}{2}\right] \mathrm{d}\,\mathrm{P}(x \mid B)\,\mathrm{d}\,\mathrm{P}(B) \\
&= \frac{m_u}{\widehat{P}} \int_{\mathcal{X}} \mathbb{1}\left[\, |h(x) - c(x)| > \frac{\gamma}{2}\right] \mathrm{d}\,\mathrm{P}(x) \\
&= \frac{m_u}{\widehat{P}} \,\mathrm{P}\left[\, |h(x) - c(x)| > \frac{\gamma}{2}\right].
\end{aligned}$$

Then, by Markov's inequality in Equation 18,

$$R^{\mathrm{AUC}}_{\widehat{F}}(\widehat{H}) \leq \frac{m_u}{\widehat{P}} \,\mathrm{P}\left[\, |h(x) - c(x)| > \frac{\gamma}{2}\right] < \frac{2m_u\sqrt{\epsilon}}{\gamma\widehat{P}}. \tag{19}$$

Therefore, choosing $\epsilon = \frac{\left(\epsilon_{\mathrm{B}}\gamma\widehat{P}\right)^2}{4m_u^2}$, is sufficient to learn $\widehat{H}$ with $R^{\mathrm{AUC}}_{\widehat{F}}(\widehat{H}) < \epsilon_{\mathrm{B}}$. Substituting this ϵ into the bound of Kearns and Schapire (1994) gives the sample complexity of learning $\widehat{H}$ as stated in the theorem.

Finally, we show that $\widehat{P} \geq \min\{P_{\mathrm{neg}}, 1 - p_{\mathrm{neg}}\}$ as asserted in the theorem, which demonstrates that $\widehat{P}$ is independent of the bag size m (so there is no hidden dependence on bag size). First, observe that $\widehat{P}_{\mathrm{neg}} \geq P_{\mathrm{neg}}$. The reason is that whenever a negative *bag* is sampled, a sample of only negative instances is guaranteed to be sampled from the bag. Thus, the probability of a negative sample of instances is at least the probability of sampling a negative bag.

Additionally, $1 - \widehat{P}_{\mathrm{neg}} \geq 1 - p_{\mathrm{neg}}$. This is true because the probability of a sample containing a positive instance is at least the probability that the very first instance sampled is positive, which is $1 - p_{\mathrm{neg}}$.

Combining the observations above, we get

$$\begin{aligned}
\widehat{P} &\triangleq \min\{\widehat{P}_{\mathrm{neg}}, 1 - \widehat{P}_{\mathrm{neg}}\} \\
&\geq \min\{P_{\mathrm{neg}}, 1 - p_{\mathrm{neg}}\}.
\end{aligned}$$

■

Lemma 4 *Suppose bag samples are of size at least m_l ($\forall i : m_l \leq |X_i|$), then $R_F^{\mathrm{AUC}}(\widehat{H}) \leq \frac{1}{P_{neg}} R_{\widehat{F}}^{\mathrm{AUC}}(\widehat{H}) + (1-\pi)^{m_l}$.*

Proof We can derive the inequality in Lemma 4 by transforming the definition of $R_F^{\mathrm{AUC}}(\widehat{H})$ in Equation 16 to that of $R_{\widehat{F}}^{\mathrm{AUC}}(\widehat{H})$ in Equation 15. Starting from $R_F^{\mathrm{AUC}}(\widehat{H})$, we get

$$\begin{aligned}
R_F^{\mathrm{AUC}}(\widehat{H}) &= \frac{\begin{array}{c}\int_{\mathcal{B}_-}\int_{\mathcal{B}_+}\int_{\mathcal{X}^*}\int_{\mathcal{X}^*} \mathbb{1}\left[\widehat{H}(X_-) > \widehat{H}(X_+)\right]\dots \\ \dots \,\mathrm{d}\,\mathrm{P}(X_+ \mid B_+)\,\mathrm{d}\,\mathrm{P}(X_- \mid B_-)\,\mathrm{d}\,\mathrm{P}(B_+)\,\mathrm{d}\,\mathrm{P}(B_-)\end{array}}{(1-P_{\mathrm{neg}})P_{\mathrm{neg}}} \\
&= \frac{\begin{array}{c}\int_{\mathcal{B}_-}\int_{\mathcal{B}_+}\int_{\mathcal{X}^*}\int_{\mathcal{X}^*_+} \mathbb{1}\left[\widehat{H}(X_-) > \widehat{H}(X_+)\right]\dots \\ \dots \,\mathrm{d}\,\mathrm{P}(X_+ \mid B_+)\,\mathrm{d}\,\mathrm{P}(X_- \mid B_-)\,\mathrm{d}\,\mathrm{P}(B_+)\,\mathrm{d}\,\mathrm{P}(B_-)\end{array}}{(1-P_{\mathrm{neg}})P_{\mathrm{neg}}} && (A) \\
&\quad + \frac{\begin{array}{c}\int_{\mathcal{B}_-}\int_{\mathcal{B}_+}\int_{\mathcal{X}^*}\int_{\mathcal{X}^*_-} \mathbb{1}\left[\widehat{H}(X_-) > \widehat{H}(X_+)\right]\dots \\ \dots \,\mathrm{d}\,\mathrm{P}(X_+ \mid B_+)\,\mathrm{d}\,\mathrm{P}(X_- \mid B_-)\,\mathrm{d}\,\mathrm{P}(B_+)\,\mathrm{d}\,\mathrm{P}(B_-)\end{array}}{(1-P_{\mathrm{neg}})P_{\mathrm{neg}}}. && (B)
\end{aligned}$$

Starting with (B), we see that since $\mathbb{1}\left[\widehat{H}(X_-) > \widehat{H}(X_+)\right] \leq 1$, we can rewrite this term as

$$\begin{aligned}
\text{(B)} &\leq \frac{\int_{\mathcal{B}_-}\int_{\mathcal{B}_+}\int_{\mathcal{X}^*}\int_{\mathcal{X}^*_-} \,\mathrm{d}\,\mathrm{P}(X_+ \mid B_+)\,\mathrm{d}\,\mathrm{P}(X_- \mid B_-)\,\mathrm{d}\,\mathrm{P}(B_+)\,\mathrm{d}\,\mathrm{P}(B_-)}{(1-P_{\mathrm{neg}})P_{\mathrm{neg}}} \\
&= \frac{\int_{\mathcal{B}_+}\int_{\mathcal{X}^*_-} \,\mathrm{d}\,\mathrm{P}(X_+ \mid B_+)\,\mathrm{d}\,\mathrm{P}(B_+)}{(1-P_{\mathrm{neg}})} \\
&\leq \frac{\int_{\mathcal{B}_+}(1-\pi)^{m_l}\,\mathrm{d}\,\mathrm{P}(B_+)}{(1-P_{\mathrm{neg}})} = (1-\pi)^{m_l}.
\end{aligned}$$

The second step follows from the fact that $\int_{\mathcal{X}^*_-} \,\mathrm{d}\,\mathrm{P}(X_+ \mid B_+)$ is the probability of sampling only negative instances within a positive bag of size at least m_l, which is at most $(1-\pi)^{m_l}$.

Continuing with term (A), we can rewrite this as

$$\begin{aligned}
\text{(A)} &= \frac{\begin{array}{c}\int_{\mathcal{B}_-}\int_{\mathcal{B}_+}\int_{\mathcal{X}^*_-}\int_{\mathcal{X}^*_+} \mathbb{1}\left[\widehat{H}(X_-) > \widehat{H}(X_+)\right]\dots \\ \dots \,\mathrm{d}\,\mathrm{P}(X_+ \mid B_+)\,\mathrm{d}\,\mathrm{P}(X_- \mid B_-)\,\mathrm{d}\,\mathrm{P}(B_+)\,\mathrm{d}\,\mathrm{P}(B_-)\end{array}}{(1-P_{\mathrm{neg}})P_{\mathrm{neg}}} && (C) \\
&\quad + \frac{\begin{array}{c}\int_{\mathcal{B}_-}\int_{\mathcal{B}_+}\int_{\mathcal{X}^*_+}\int_{\mathcal{X}^*_+} \mathbb{1}\left[\widehat{H}(X_-) > \widehat{H}(X_+)\right]\dots \\ \dots \,\mathrm{d}\,\mathrm{P}(X_+ \mid B_+)\,\mathrm{d}\,\mathrm{P}(X_- \mid B_-)\,\mathrm{d}\,\mathrm{P}(B_+)\,\mathrm{d}\,\mathrm{P}(B_-)\end{array}}{(1-P_{\mathrm{neg}})P_{\mathrm{neg}}}. && (D)
\end{aligned}$$

Now, we see that (D) $= 0$, since it involves an integral over bags with positive instances in negative bags, which occurs with probability zero by Condition 2 in Definition 1. For (C), since $0 \leq \mathbb{1}\big[\widehat{H}(X_-) > \widehat{H}(X_+)\big]$, we can bound this term by taking the outermost integrals with respect to the entire bag space:

$$\begin{aligned}
(\mathrm{C}) &\leq \frac{\begin{array}{c}\int_{\mathcal{B}}\int_{\mathcal{B}}\int_{\mathcal{X}_-^*}\int_{\mathcal{X}_+^*}\mathbb{1}\big[\widehat{H}(X_-) > \widehat{H}(X_+)\big]\dots\\ \dots\,\mathrm{d}\,\mathrm{P}(X_+ \mid B_+)\,\mathrm{d}\,\mathrm{P}(X_- \mid B_-)\,\mathrm{d}\,\mathrm{P}(B_+)\,\mathrm{d}\,\mathrm{P}(B_-)\end{array}}{(1-P_{\mathrm{neg}})P_{\mathrm{neg}}}\\
&\leq \left(\frac{(1-\widehat{P}_{\mathrm{neg}})\widehat{P}_{\mathrm{neg}}}{(1-P_{\mathrm{neg}})P_{\mathrm{neg}}}\right)\frac{\begin{array}{c}\int_{\mathcal{B}}\int_{\mathcal{B}}\int_{\mathcal{X}_-^*}\int_{\mathcal{X}_+^*}\mathbb{1}\big[\widehat{H}(X_-) > \widehat{H}(X_+)\big]\dots\\ \dots\,\mathrm{d}\,\mathrm{P}(X_+ \mid B_+)\,\mathrm{d}\,\mathrm{P}(X_- \mid B_-)\,\mathrm{d}\,\mathrm{P}(B_+)\,\mathrm{d}\,\mathrm{P}(B_-)\end{array}}{(1-\widehat{P}_{\mathrm{neg}})\widehat{P}_{\mathrm{neg}}}.
\end{aligned}$$

Using the definition of $R_{\widehat{F}}^{\mathrm{AUC}}(\widehat{H})$ in Equation 15, we get that

$$(\mathrm{C}) \leq \left(\tfrac{(1-\widehat{P}_{\mathrm{neg}})\widehat{P}_{\mathrm{neg}}}{(1-P_{\mathrm{neg}})P_{\mathrm{neg}}}\right) R_{\widehat{F}}^{\mathrm{AUC}}(\widehat{H}) \leq \tfrac{1}{P_{\mathrm{neg}}} R_{\widehat{F}}^{\mathrm{AUC}}(\widehat{H}).$$

The second inequality results by observing that samples of only negative instances can be sampled within negative *or* positive bags, so $P_{\mathrm{neg}} \leq \widehat{P}_{\mathrm{neg}} \leq 1$ and $1 - P_{\mathrm{neg}} \geq 1 - \widehat{P}_{\mathrm{neg}}$.

Combining the terms above, we have that

$$\begin{aligned}
R_F^{\mathrm{AUC}}(\widehat{H}) &= (\mathrm{A}) + (\mathrm{B}) = \big((\mathrm{C}) + (\mathrm{D})\big) + (\mathrm{B})\\
&\leq \tfrac{1}{P_{\mathrm{neg}}} R_{\widehat{F}}^{\mathrm{AUC}}(\widehat{H}) + (1-\pi)^{m_l}.
\end{aligned}$$

■

Theorem 6 *Bag-labeling functions derived from p-concept class $\mathcal{C}$ with pseudo-dimension* $\mathrm{PD}(\mathcal{C})$ *are AUC-PAC-learnable from MI data using*

$$O\left(\tfrac{1}{\left(\epsilon_{\mathrm{B}}^2\gamma\pi\widehat{P}P_{neg}\right)^4}\left(\mathrm{PD}(\mathcal{C})\log\tfrac{1}{\left(\epsilon_{\mathrm{B}}\gamma\pi\widehat{P}P_{neg}\right)} + \log\tfrac{1}{\delta}\right)\right). \tag{17}$$

examples using standard ERM approaches when bag sample sizes are bounded by $m_l \leq m \leq m_u$ and $m_l \geq \frac{1}{\pi}\log\frac{2}{\epsilon_{\mathrm{B}}}$, where $\widehat{P} = \min\{\widehat{P}_{neg}, 1 - \widehat{P}_{neg}\}$.

Proof As in Theorem 5, we will use ERM to learn a *p*-concept h to model c accurately with high probability. Then, we can bound $R_F^{\mathrm{AUC}}(\widehat{H})$ using

$$\begin{aligned}
R_F^{\mathrm{AUC}}(\widehat{H}) &\leq \frac{1}{P_{\mathrm{neg}}} R_{\widehat{F}}^{\mathrm{AUC}}(\widehat{H}) + (1-\pi)^{m_l} && \text{(by Lemma 4)}\\
&\leq \frac{2m_u\sqrt{\epsilon}}{\gamma\widehat{P}P_{\mathrm{neg}}} + (1-\pi)^{m_l}. && \text{(by Theorem 5, Equation 19)}
\end{aligned}$$

In the case that $\pi = 1$, then the second term in the sum is zero. Otherwise, suppose the minimum bag size is such that

$$m_l \geq \tfrac{1}{\pi}\log\tfrac{2}{\epsilon_{\mathrm{B}}} \geq \frac{\log\frac{\epsilon_{\mathrm{B}}}{2}}{\log(1-\pi)} = \log_{1-\pi}\tfrac{\epsilon_{\mathrm{B}}}{2},$$

where the second inequality follows from the fact that $\pi \leq -\log(1-\pi)$ for $\pi \in (0,1)$. Therefore, since $(1-\pi) < 1$, we have that

$$(1-\pi)^{m_l} \leq (1-\pi)^{\log_{1-\pi}\frac{\epsilon_{\mathrm{B}}}{2}} = \tfrac{\epsilon_{\mathrm{B}}}{2}.$$

Furthermore, when learning the instance p-concept h, we can choose ϵ to be such that $\epsilon = \left(\frac{\epsilon_{\mathrm{B}}\gamma\widehat{P}P_{\mathrm{neg}}}{4m_u}\right)^2$. By the bound on $R_F^{\mathrm{AUC}}(\widehat{H})$, with probability $(1-\delta)$, we have

$$\begin{aligned} R_F^{\mathrm{AUC}}(\widehat{H}) &\leq \frac{2m_u\sqrt{\epsilon}}{\gamma\widehat{P}P_{\mathrm{neg}}} + (1-\pi)^{m_l} \\ &\leq \frac{\epsilon_{\mathrm{B}}}{2} + \frac{\epsilon_{\mathrm{B}}}{2} = \epsilon_{\mathrm{B}}. \end{aligned}$$

Substituting the expression for ϵ in terms of ϵ_{B} into the bound for learning p-concepts using ERM (Kearns and Schapire, 1994) gives a sample complexity of

$$O\left(\frac{m_u^4}{\left(\epsilon_{\mathrm{B}}\gamma\widehat{P}P_{\mathrm{neg}}\right)^4}\left(\mathrm{PD}(\mathcal{C})\log\frac{m_u}{\left(\epsilon_{\mathrm{B}}\gamma\widehat{P}P_{\mathrm{neg}}\right)} + \log\tfrac{1}{\delta}\right)\right).$$

By sub-sampling large bags, we can place a conservative upper bound on bag sample size $m_u = O\left(\frac{1}{\epsilon_{\mathrm{B}}\pi}\right)$. Then by substituting this bound into that above gives the bound as stated in Equation 17. ■

References

T. Adel, B. Smith, R. Urner, D. Stashuk, and D. J. Lizotte. Generative multiple-instance learning models for quantitative electromyography. In *Uncertainty in Artificial Intelligence*, 2013.

S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems*, pages 561–568, 2003.

D. Ascher, P. F. Dubois, K. Hinsen, J. Hugunin, and T. Oliphant. *Numerical Python*. Lawrence Livermore National Laboratory, Livermore, CA, 2001. `http://numpy.scipy.org/`.

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

B. Babenko, N. Verma, P. Dollár, and S. Belongie. Multiple instance learning with manifold bags. In *Proceedings of the International Conference on Machine Learning*, pages 81–88, 2011.

S. Ben-David, N. Eiron, and P. M. Long. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66(3):496–514, 2003.

C. Bergeron, J. Zaretzki, C. Breneman, and K. P. Bennett. Multiple instance ranking. In *Proceedings of the International Conference on Machine Learning*, pages 48–55, 2008.

J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, 2012.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

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

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik–Chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36(4): 929–965, 1989.

F. Briggs, X. Z. Fern, and R. Raich. Rank-loss support instance machines for MIML instance annotation. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 534–542, 2012.

C. Carson, S. Belongie, H. Greenspan, and J. Malik. Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1026–1038, 2002.

Y. Chen, J. Bi, and J. Z. Wang. MILES: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12): 1931–1947, 2006.

J. Dahl and L. Vandenberghe. CVXOPT: A Python package for convex optimization, 2009. `http://abel.ee.ucla.edu/cvxopt`.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

J. Diestel and J. J. Uhl. *Vector Measures*. Mathematical surveys and monographs. American Mathematical Society, 1977.

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

D. Diochnos, R. Sloan, and G. Turán. On multiple-instance learning of halfspaces. *Information Processing Letters*, 2012.

G. Doran and S. Ray. A theoretical and empirical analysis of support vector machine methods for multiple-instance classification. *Machine Learning Journal*, pages 1–24, 2013.

G. Doran and S. Ray. Learning instance concepts from multiple-instance data with bags as distributions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1802–1808, 2014.

L. Du, W. L. Buntine, and M. Johnson. Topic segmentation with a structured topic model. In *Proceedings of the Human Language Technologies Conference of the North American Chapter of the Association for Computational Linguistics*, pages 190–200, 2013.

L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 524–531, 2005.

J. Foulds and P. Smyth. Multi-instance mixture models and semi-supervised learning. In *SIAM International Conference on Data Mining*. SIAM, 2011.

J. R. Foulds. Learning instance weights in multi-instance learning. Master's thesis, The University of Waikato, 2008.

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

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

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

E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001. `http://www.scipy.org/`.

M. Kandemir and F. A. Hamprecht. Instance label prediction by Dirichlet process multiple instance learning. In *Uncertainty in Artificial Intelligence*, 2014.

M. J. Kearns and R. E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.

H. Kuck and N. de Freitas. Learning about individuals from group statistics. In *Uncertainty in Artificial Intelligence*, 2005.

O. Kundakcioglu, O. Seref, and P. Pardalos. Multiple instance learning via margin maximization. *Applied Numerical Mathematics*, 60(4):358–369, 2010.

Y.-F. Li, J. T. Kwok, I. W. Tsang, and Z.-H. Zhou. A convex method for locating regions of interest with multi-instance learning. In *Machine Learning and Knowledge Discovery in Databases*, pages 15–30. Springer, 2009.

G. Liu, J. Wu, and Z.-H. Zhou. Key instance detection in multi-instance learning. In *Proceedings of the Asian Conference on Machine Learning*, pages 253–268, 2012.

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

O. Mangasarian and E. Wild. Multiple instance classification via successive linear programming. *Journal of Optimization Theory and Applications*, 137:555–568, 2008.

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

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

J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, 1999.

R. Rahmani, S. A. Goldman, H. Zhang, J. Krettek, and J. E. Fritts. Localized content based image retrieval. In *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 227–236. ACM, 2005.

S. Ray and M. Craven. Supervised versus multiple instance learning: an empirical comparison. In *Proceedings of the International Conference on Machine Learning*, pages 697–704, 2005.

S. Sabato and N. Tishby. Multi-instance learning with any hypothesis class. *Journal of Machine Learning Research*, 13:2999–3039, 2012.

G. Salton and M. McGill. *Introduction to modern information retrieval.* McGraw-Hill Computer Science Series. McGraw-Hill, 1983.

S. Scott, J. Zhang, and J. Brown. On generalized multiple-instance learning. *International Journal of Computational Intelligence and Applications*, 5(1):21–35, 2005.

B. Settles, M. Craven, and S. Ray. Multiple-instance active learning. In *Advances in Neural Information Processing Systems*, pages 1289–1296, 2008.

H. U. Simon. PAC-learning in the presence of one-sided classification noise. *Annals of Mathematics and Artificial Intelligence*, pages 1–18, 2012.

Z. Szabó, A. Gretton, B. Póczos, and B. Sriperumbudur. Two-stage sampled learning theory on distributions. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2015.

V. Tragante do Ó, D. Fierens, and H. Blockeel. Instance-level accuracy versus bag-level accuracy in multi-instance learning. In *Proceedings of the 23rd Benelux Conference on Artificial Intelligence*, 2011.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

X. Xu. Statistical learning in multiple instance problems. Master's thesis, The University of Waikato, 2003.

S.-H. Yang, H. Zha, and B.-G. Hu. Dirichlet-Bernoulli alignment: A generative model for multi-class multi-label multi-instance corpora. In *Advances in Neural Information Processing Systems*, pages 2143–2150, 2009.

Z. Zhou, Y. Sun, and Y. Li. Multi-instance learning by treating instances as non-IID samples. In *Proceedings of the International Conference on Machine Learning*, pages 1249–1256, 2009.