A Novel Attribute-based Symmetric Multiple Instance Learning for Histopathological Image Analysis

Trung Vu, *Graduate Student Member, IEEE*, Phung Lai, *Member, IEEE*, Raviv Raich, *Senior Member, IEEE*, Anh Pham, *Member, IEEE*, Xiaoli Z. Fern, *Member, IEEE*, and UK Arvind Rao, *Senior Member, IEEE*

***Abstract*— Histopathological image analysis is a challenging task due to a diverse histology feature set as well as due to the presence of large non-informative regions in whole slide images. In this paper, we propose a multiple-instance learning (MIL) method for image-level classification as well as for annotating relevant regions in the image. In MIL, a common assumption is that negative bags contain only negative instances while positive bags contain one or more positive instances. This asymmetric assumption may be inappropriate for some application scenarios where negative bags also contain representative negative instances. We introduce a novel symmetric MIL framework associating each instance in a bag with an attribute which can be either negative, positive, or irrelevant. We extend the notion of relevance by introducing control over the number of relevant instances. We develop a probabilistic graphical model that incorporates the aforementioned paradigm and a corresponding computationally efficient inference for learning the model parameters and obtaining an instance level attribute-learning classifier. The effectiveness of the proposed method is evaluated on available histopathology datasets with promising results.**

***Index Terms*— Histopathological image analysis, multiple instance learning, symmetric setting, attribute learning, cardinality constraints, dynamic programming**

I. Introduction

Histopathological image analysis is a critical task in cancer diagnosis. Generally, this process is performed by pathologists who are capable of identifying problem-specific cues in a digital image, or a whole slide image (WSI), in order to classify it into one of the disease categories. In recent years, there have been an increasing interest in the application of automatic histopathological image analysis using machine learning algorithms [1], [2]. The advantages of this approach include (i) reducing variability in human interpretations and hence improving classification accuracy, (ii) eliminating a significant amount of trivial cases to ease the burden on pathologists, and (iii) providing quantitative image analysis in the context of one specific disease.

Most conventional approaches to automatic histopathological image analysis fall into the category of fully supervised learning, where training labels are available for the WSI and all of its patches (small blocks extracted from the image). These methods often rely on feature extraction techniques that are customized for a variety of problems, namely texture features [3], [4], spatial features [5], [6], graph-based features [7], [8], and morphological features [9], [10]. Those features can then be used by various classification algorithms such as random forest, support vector machines (SVM), and convolutional neural networks (CNN). Recently, automatic feature discovery framework has also been proposed by Vu *et al.* [11]. Their discriminative feature-oriented dictionary learning (DFDL) method was shown to outperform many competing methods, particularly in low training scenarios. Nevertheless, one major disadvantage of the fully-supervised approach is the labeling cost. Since each WSI typically comprises hundreds of patches, it requires a large amount of labor to create even a small number of training data. Moreover, labeling a histopathology image at the region-level could be a challenging task with inherent uncertainty, even for experts in the field. To address these issues, many researchers have been studying weakly-supervised learning that focuses on coarse-grain annotations, i.e., only WSI labels are given.

Multiple Instance Learning (MIL) is a framework for weakly supervised learning (limited supervision) that relies on a training set of bags of instances labeled at the bag level only. In our scenario, each WSI (bag) contains a collection of tissue segments (instances), but the annotation of cancer-type is only available at the image/bag level. The goal is to develop a classifier to predict both bag level and instance level labels. Various MIL-based approaches have been applied successfully in biochemistry [12], [13], image classification and segmentation [14]–[16], text categorization [17], [18], object recognition, tracking and localization [19]–[21], behavior coding [22], anomaly detection [23], and co-saliency

Manuscript received March 05, 2020; accepted April 7, 2020. was supported in part by the National Science Foundation under Grant CCF-1254218, Grant DBI-1356792, and Grant IIS-1055113, in part by the American Cancer Society Research Scholar under Grant RSG-RSG-16-005-01, and in part by the National Institutes of Health under Grant 1R37CA21495501A1. *(Trung Vu and Phung Lai contributed equally to this work.) (Corresponding author: Trung Vu.)*

Trung Vu, Phung Lai, Raviv Raich, Anh Pham, and Xiaoli Fern are with School of EECS, Oregon State University, Corvallis, OR 97331-5501 (e-mail: {vutru, laith, raich, phaman, xfern}@oregonstate.edu).

UK Arvind Rao is with the Department of Computational Medicine and Bioinformatics, and Department of Radiation Oncology, The University of Michigan, Ann Arbor, MI 48109 (e-mail: ukarvind@umich.edu).

detection [24]. In histopathological image classification, only a limited number of MIL approaches have been studied in literature. In [25], Dundar *et al.* introduced a multiple instance learning approach (MILSVM) based on the implementation of the large margin principle with different loss functions defined for positive and negative samples. Later on, Xu *et al.* [26] adopted the clustering concept into MIL to propose an integrated framework of segmentation, clustering, and classification named multiple clustered instance learning (MCIL). The authors also extended their work by taking into consideration the contextual prior in the MIL training stage to reduce the intrinsic ambiguity. Most recently, a comparison of general MIL-based methods on histopathological image classification, namely, mi-SVM and MI-SVM [27], miGraph and MIGraph [28] has also been reported in [29]. All of the aforementioned methods, nonetheless, deal with predicting the presence or absence of cancer, and are based on an asymmetric assumption that all instances in a negative bag are negative while each positive bag contains at least one positive instance. This commonly-used MIL assumption may be not suitable for other MIL setting such as predicting the cancer type based on histopathology images. In this case, only a fraction of the tissue segments can be useful towards recognizing the cancer type of each WSI, and such MIL setting is symmetric in that both positive and negative bags contain stereotypical instances featured for each cancer type along with irrelevant ones.

In this paper, we consider the binary classification problem of cancer types based on histopathology images. Our contribution in this paper is as follows. First, we introduce a novel symmetric MIL where both negative and positive bags contain relevant and irrelevant instances. Second, we propose a probabilistic graphical model, named Attribute-based Symmetric Multiple Instance Learning (AbSMIL), that incorporates cardinality constraints on the relevant instances in each bag to leverage possible prior knowledge and develop a forward-backward dynamic programming algorithm for learning model parameters. To facilitate efficient inference, the online learning version of our algorithm is also presented. Finally, the advantages of the proposed framework are demonstrated by experiments on instance annotation and bag level label prediction using classical multi-instance image recognition datasets as well as histopathology datasets.

II. Related work

In MIL setting, it is important to make assumptions regarding the relationship between the instances within a bag and the class label of the bag. Most of the MIL algorithms follow the standard assumption that each positive bag contains at least one positive instance and negative bags contain only negative instances. In one of the early works, Maron and Lozano-Pérez [30] introduced the concept point that is close to one instance in each positive bag and far from all instances in negative bags. The diverse density (DD) framework [31], [32] is then developed based on the idea of finding the best candidate concept. Later on, Andrew *et al.* [27] extended SVM approach to mi-SVM and MI-SVM by finding a separating hyperplane such that at least one instance in every positive bag is located on one side and all instances in negative bags are located on the other side of the hyperplane. In a different approach, Zhou *et al.* [28] proposed MIGraph and miGraph methods that map every bag to a graph and explicitly model the relationships among the instances within a bag. A number of single-instance algorithms have also been adapted to a multiple-instance context such as MIL-Boost [33], KI-SVM [34], Latent SVM [35], MI-CRF [36]. They all maintain the classical asymmetric assumption in their algorithms.

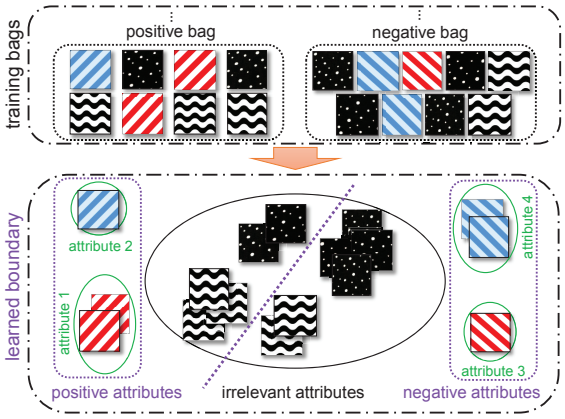

Fig. 1. The setting of the proposed attribute-based symmetric MIL framework. The goal is to classify positive and negative bags and to figure out distinctly positive and negative relevant attributes (stripes going from *bottom-left to top-right* or *bottom-right to top-left*, respectively) and irrelevant attributes (*dotted and wave*) in each bag category. The proposed model helps control the relevant instance proportions in the bags.

For some MIL applications where the class of a bag is defined by instances belonging to more than one concept, the standard assumption may be viewed as too strict. Therefore, researchers have recently shown a general interest in other more loose assumptions such as the collective assumption [37], [38]. In this paper, we consider the problem of predicting cancer types in histopathology images and propose a symmetric assumption that treats classes equally (see Fig. 1). In a positive bag, there is at least one relevant instance that demonstrates certain attributes associated with the positive class (e.g., stripes going from bottom-left to top-right). Similarly, a negative bag also contains at least one relevant instance that demonstrates certain negative attributes (e.g., stripes going from bottom-right to top-left). Finally, both bags contain some irrelevant instances whose attributes do not contribute to the difference between the two classes (e.g., dotted and wave). The goal is to learn the positive, negative, and irrelevant attributes and use them to address various classification tasks within this framework.

Our assumption is motivated by multi-instance multi-label learning (MIML), a generalization of single-instance binary classifiers to the multiple-instance case. In MIML setting, each instance is associated with a latent instance label and each bag is the union of its instance labels [39]. With the presence of novel class instances [40], the bag label only involves the known instance classes and does not provide information about the presence or absence of the novel class. This consideration is similar to our symmetric MIL setting where only a small

portion of the relevant instances are labeled and irrelevant instances are often ignored by pathologists. However, a simple reduction of the model for MIML would fail in the cancer classification problem as the bag labels in MIL are binary but not a subset of the class labels. Recently, You *et al.* [41] introduced cardinality constraints to the MIML setting and demonstrated that optimizing the control over the maximum number of instances per bag can significantly improve the performance of the model. Motivated by the result, we propose using cardinality constraints to limit the number of relevant attributes in each bag. Given that there are hundreds of instances per bag, cardinality constraints can help control the model complexity. While the probabilistic machinery for implementing cardinality constraints is similar to the approach in You *et al.* [41], the graphical model and inference methods differ. *The modeling difference:* instance-level labels include unknown attribute/cluster labels (see Fig. 1), whereas in the MIML setting in [41] instance-level labels are taken directly from bag-level labels (e.g., the bag-level label $\{2, 5\}$ implies that relevant instances must have labels of either 2 or 5). *The inference difference:* (i) to promote the cluster diversity, we introduce entropy regularization to the original log-likelihood objective, (ii) to reduce the computational burden of large histopathology images, we apply a stochastic gradient descent approach. Both of which are not included in [41].

III. Problem Formulation and Proposed Model

In this section, we formulate the problem of cancer type classification and describe our proposed AbSMIL approach.

A. Problem Formulation

We consider a collection of B bags and their labels, denoted by $\{\boldsymbol{X}_b, Y_b\}_{b=1}^{B}$. The bag level label $Y_b \in \{0, 1\}$ represents each of the two cancer types. The bth bag contains n_b instances $\boldsymbol{X}_b = \{\boldsymbol{x}_{bi}\}_{i=1}^{n_b}$ where $\boldsymbol{x}_{bi} \in \mathbb{R}^d$ is a feature vector for the ith instance. Our goal is to predict the bag label based on the set of feature vectors for its instances. Moreover, we would like to learn a robust model that are capable of explaining the labeling decision. To that end, we consider the following attribute-based assumption on the data.

Relevant/irrelevant instances and attributes assumption. We assume that each instance in a bag can be either relevant or irrelevant. Relevant instances provide useful information towards a specific class. Further, there may be more than one types of relevant instances for the same class, which we capture as instance attributes (clusters). In Fig. 1, for example, the positive (or negative) class always has two relevant attributes: 1 and 2 (or 3 and 4). Formally, we assume the attribute z_{bi}, corresponding to the ith instance in the bth bag, belongs to the set $\{0, 1, \ldots, \mathbb{C}\}$, where 0 is reserved for irrelevant instances and the non-zero integers represent $\mathbb{C}$ attributes for relevant instances. For the bth bag, we denote the set of hidden attributes for all instances by $\boldsymbol{z}_b = [z_{b1}, z_{b2}, \ldots, z_{bn_b}]^T$, and the binary vector indicating the presence/absence of relevant attributes by $\boldsymbol{y}_b = [y_{b1}, y_{b2}, \ldots, y_{b\mathbb{C}}]^T \in \{0, 1\}^{\mathbb{C}}$.

No mixed-class attributes assumption. Let us call attributes that provide sufficient information for predicting the positive bag as positive attributes (PAs) and similarly, for negative bags as negative attributes (NAs). Since the goal is to find distinct attributes that discriminate the two classes, we assume there is no shared relevant attribute between positive bags and negative bags. To be more specific, we consider the first half of the attribute set $C^+ = \{1, \ldots, \frac{\mathbb{C}}{2}\}$ as PAs and the second half $C^- = \{\frac{\mathbb{C}}{2} + 1, \ldots, \mathbb{C}\}$ as NAs ($\mathbb{C}$ is even in our assumption). Although we focus on a balanced number of PAs and NAs, the proposed model is not limited to this symmetry when extending to other settings.

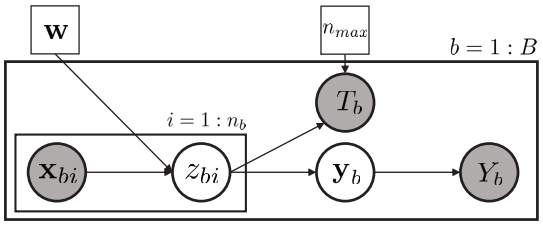

Fig. 2. Graphical model for the proposed AbSMIL model. Observed variables are shaded.

Relevance cardinality constraints. In many applications, the domain knowledge may provide practical information regarding the number of relevant instances. Such heuristics can be exploited with constraints on the maximum number of relevant instances per bag, i.e.,

$$\sum_{i=1}^{n_b} I_{z_{bi} \neq 0} \leq n_{\max}, \qquad \text{for } b = 1, \ldots, B,$$

where I_σ denotes the indicator function taking the value 1 if σ is true and 0 otherwise. Unlike the standard MIL assumption, the relevance cardinality constraints require both positive and negative bags to have a bounded number of relevant instances. In our discriminative model, $n_{\max}$ is a tuning parameter that can be optimized to improve the classification performance. Although those make our model more amenable to cancer type recognition, we note that they are not equivalent to sparsity-promoting prior in generative models such as Laplacian [42] or spike-and-slab [43].

Potential extension to multinomial classification. As the no mixed-class attributes assumption and relevance cardinality constraint apply for every bag, we can easily extend our model to the problem of multinomial MIL classification by considering more categories of attributes. For three-cancer-type classification, by way of illustration, we can consider three groups of labels $\{1, \ldots, \frac{\mathbb{C}}{3}\}$, $\{\frac{\mathbb{C}}{3} + 1, \ldots, \frac{2\mathbb{C}}{3}\}$ and $\{\frac{2\mathbb{C}}{3} + 1, \ldots, \mathbb{C}\}$.

B. Attribute-Based Symmetric MIL Model

The graphical representation of the proposed model is illustrated in Fig. 2. We assume that instances are independent given all the feature vectors in the bag. We follow the discriminative approach in [39] to model the relationship between the attribute of an instance z_{bi} and its feature vector $\boldsymbol{x}_{bi}$ by a multinomial logistic regression function

$$P_{bic}(\boldsymbol{w}) = P(z_{bi} = c \mid \boldsymbol{x}_{bi}, \boldsymbol{w}) = \frac{e^{\boldsymbol{w}_c^T \boldsymbol{x}_{bi}}}{\sum_{c=0}^{\mathbb{C}} e^{\boldsymbol{w}_c^T \boldsymbol{x}_{bi}}}, \qquad (1)$$

where $\boldsymbol{w}_c \in \mathbb{R}^d$ is the weight for the cth attribute and $\boldsymbol{w} = [\boldsymbol{w}_0, \boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{\mathbb{C}}]$. For the cth attribute in the bth bag, a presence or absence indicator y_{bc} is computed following the standard assumption [44]

$$y_{bc} = 1 - \prod_{i=1}^{n_b} I_{z_{bi} \neq c}, \qquad \text{for } c = 1, \ldots, \mathbb{C}.$$

To encode the cardinality constraint $\sum_{i=1}^{n_b} I_{z_{bi}\neq 0} \leq n_{\max}$ for the bth bag, we introduce an binary observation variable

$$T_b = \Big(I_{\sum_{i=1}^{n_b} I_{z_{bi}\neq 0} \geq 1}\Big)\Big(I_{\sum_{i=1}^{n_b} I_{z_{bi}\neq 0} \leq n_{\max}}\Big).$$

By letting $T_b = 1$, we enforce the constraint on the number of relevant instances per bag during training stage. The bag level label Y_b is computed based on the presence of positive and negative attributes

$$Y_b|\boldsymbol{y}_b = \begin{cases} 0, \text{ for } \big(\bigcap_{c=1}^{\frac{\mathbb{C}}{2}} \{y_{bc} = 0\}\big) \bigcap \big(\bigcup_{c=\frac{\mathbb{C}}{2}+1}^{\mathbb{C}} \{y_{bc} = 1\}\big), \\ 1, \text{ for } \big(\bigcup_{c=1}^{\frac{\mathbb{C}}{2}} \{y_{bc} = 1\}\big) \bigcap \big(\bigcap_{c=\frac{\mathbb{C}}{2}+1}^{\mathbb{C}} \{y_{bc} = 0\}\big), \\ 2, \text{ otherwise.} \end{cases}$$

For the completeness, the model may allow $Y_b = 2$; however, the considered dataset contains only positive bags $Y_b = 1$ and negative bags $Y_b = 0$. Thus, we can ignore the case $Y_b = 2$ in our derivation. To summarize, our model includes the observation $\{Y_b, T_b, \boldsymbol{X}_b\}_{b=1}^B$, unknown classifier parameter $\boldsymbol{w}$, hidden variables $\{\boldsymbol{y}_b, \boldsymbol{z}_b\}_{b=1}^B$ and a tuning parameter $n_{\max}$.

IV. Inference

This section provides details of our graphical model, including the derivation of the incomplete log-likelihood, the expectation maximization (EM) approach to estimate the model parameters and the prediction of both instance and bag level labels. To facilitate efficient inference, we further present the dynamic programming method for the expectation step in combination with online learning for the maximization step.

A. Regularized Maximum Likelihood

Since the bags are independent, the normalized negative incomplete log-likelihood is given by

$$\mathbb{L}_{incm}(\boldsymbol{w}) = -\frac{1}{B}\sum_{b=1}^{B}\Big(\log P(Y_b, T_b \mid \boldsymbol{X}_b, \boldsymbol{w}) + \log P(\boldsymbol{X}_b)\Big)$$

where we assume that $P(\boldsymbol{X}_b)$ is a constant w.r.t. $\boldsymbol{w}$. To ensure the attributes are distinctly different, we introduce entropy regularization that quantifies the cluster diversity (see [45], [46]):

$$\mathbb{H}_b(\boldsymbol{w}) = \sum_{i=1}^{n_b}\Big(-\sum_{c=0}^{\mathbb{C}} P_{bic}(\boldsymbol{w}) \log P_{bic}(\boldsymbol{w})\Big), \tag{2}$$

and a quadratic penalty to control the complexity of the model and avoid over-fitting. Now let us denote $\mathbb{L}_b(\boldsymbol{w}) = -\log P(Y_b, T_b \mid \boldsymbol{X}_b, \boldsymbol{w})$, the objective in our approach can be formulated as

$$\mathbb{L}(\boldsymbol{w}) = \frac{1}{B}\sum_{b=1}^{B}\Big(\mathbb{L}_b(\boldsymbol{w}) + \lambda_e \mathbb{H}_b(\boldsymbol{w})\Big) + \frac{\lambda_q \|\boldsymbol{w}\|^2}{2}, \tag{3}$$

where λ_q, λ_e are parameters of the quadratic regularizer and the entropy minimizer, respectively. Following the principle of maximum likelihood estimation (MLE), our goal is to minimize $\mathbb{L}(\boldsymbol{w})$. However, this optimization problem is challenging as the probability $P(Y_b, T_b \mid \boldsymbol{X}_b, \boldsymbol{w})$ in the objective function is not trivial to compute. For instance, to obtain $P(Y_b = 1, T_b \mid \boldsymbol{X}_b, \boldsymbol{w})$, we marginalize the joint probability model of all the model variables over the hidden variables:

$$P(Y_b = 1, T_b \mid \boldsymbol{X}_b, \boldsymbol{w}) = \sum_{\boldsymbol{z}_b, \boldsymbol{y}_b} P(Y_b = 1, T_b, \boldsymbol{z}_b, \boldsymbol{y}_b \mid \boldsymbol{X}_b, \boldsymbol{w}).$$

Using the graphical model in Fig. 2, we can expand the joint probability as

$$\begin{aligned} &P(Y_b = 1, T_b, \boldsymbol{z}_b, \boldsymbol{y}_b \mid \boldsymbol{X}_b, \boldsymbol{w}) \\ &= P(Y_b = 1 \mid \boldsymbol{y}_b) P(\boldsymbol{y}_b \mid \boldsymbol{z}_b) P(T_b \mid \boldsymbol{z}_b) P(\boldsymbol{z}_b \mid \boldsymbol{X}_b, \boldsymbol{w}). \end{aligned} \tag{4}$$

Finally, by substituting (4) into the marginalization, we obtain

$$\begin{aligned} &P(Y_b = 1, T_b \mid \boldsymbol{X}_b, \boldsymbol{w}) \\ &= \sum_{\boldsymbol{z}_b, \boldsymbol{y}_b} P(Y_b = 1 \mid \boldsymbol{y}_b) P(\boldsymbol{y}_b \mid \boldsymbol{z}_b) P(T_b \mid \boldsymbol{z}_b) P(\boldsymbol{z}_b \mid \boldsymbol{X}_b, \boldsymbol{w}) \\ &= \sum_{\boldsymbol{z}_b} I_{\sum_{i=1}^{n_b} I_{z_{bi}\neq 0} \geq 1} I_{\sum_{i=1}^{n_b} I_{z_{bi}\neq 0} \leq n_{\max}} \prod_{i=1}^{n_b} P(z_{bi} \mid \boldsymbol{x}_{bi}, \boldsymbol{w}) \end{aligned}$$

where the summation in the intermediate step is over $\boldsymbol{z}_b \in \{0, 1, \ldots, \mathbb{C}\}^{n_b}$ and $\boldsymbol{y}_b \in \{0, 1\}^{\mathbb{C}}$, while the latter summation is over $\boldsymbol{z}_b \in \{0, 1, \ldots, \frac{\mathbb{C}}{2}\}^{n_b}$. Notice that this summation includes a total of approximately $\sum_{k=1}^{n_{\max}} \binom{n_b}{k} \big(\frac{\mathbb{C}}{2}\big)^k$ terms, and hence, is computationally expensive or even intractable when $n_{\max}$ is large. Therefore, we consider an EM approach as the alternative approach to efficiently minimizing $\mathbb{L}(\boldsymbol{w})$.

B. Estimation of Model Parameters

Let us begin by identifying the negative complete log-likelihood as

$$\mathbb{L}_{cm}(\boldsymbol{w}) = -\frac{1}{B}\sum_{b=1}^{B} \log P(Y_b, T_b, \boldsymbol{y}_b, \boldsymbol{z}_b \mid \boldsymbol{X}_b, \boldsymbol{w}) + K,$$

where K corresponds to the term $\frac{1}{B}\sum_{b=1}^{B} \log P(\boldsymbol{X}_b)$ independent of the parameter vector $\boldsymbol{w}$. Substituting the model dependence structure from (4) into $\mathbb{L}_{cm}$ and absorbing terms that do not depend on $\boldsymbol{w}$ into the constant yields

$$\mathbb{L}_{cm}(\boldsymbol{w}) = \tfrac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n_b}\Big(-\sum_{c=0}^{\mathbb{C}} I_{z_{bi}=c} \boldsymbol{w}_c^T \boldsymbol{x}_{bi} + \log\big(\sum_{c=0}^{\mathbb{C}} e^{\boldsymbol{w}_c^T \boldsymbol{x}_{bi}}\big)\Big)$$

up to a constant. The EM algorithm seeks to find the minimum of $\mathbb{L}(\boldsymbol{w})$ by iteratively applying the following two steps:

E-step. The surrogate function is obtained by taking the expectation with respect to the current conditional distribution

of the latent data $\{\boldsymbol{y}, \boldsymbol{z}\}$ given the observed data $\{Y, T, \boldsymbol{X}\}$ and the current estimates of the parameters $\boldsymbol{w}^{(k)}$

$$E_{\boldsymbol{y},\boldsymbol{z}|Y,T,\boldsymbol{X},\boldsymbol{w}^{(k)}}\Big[\mathbb{L}_{cm}(\boldsymbol{w})\Big] = \frac{1}{B}\sum_{b=1}^{B}\mathbb{J}_b(\boldsymbol{w}, \boldsymbol{w}^{(k)})$$

where $\mathbb{J}_b(\boldsymbol{w}, \boldsymbol{w}^{(k)})$ is defined as

$$\sum_{i=1}^{n_b}\Bigg(-\sum_{c=0}^{\mathbb{C}} P_{bic}^{post}(\boldsymbol{w}^{(k)}) \cdot \boldsymbol{w}_c^T \boldsymbol{x}_{bi} + \log\Big(\sum_{c=0}^{\mathbb{C}} e^{\boldsymbol{w}_c^T \boldsymbol{x}_{bi}}\Big)\Bigg) \quad (5)$$

and $P_{bic}^{post}(\boldsymbol{w}^{(k)}) = P(z_{bi} = c \mid Y_b, T_b, \boldsymbol{x}_{bi}, \boldsymbol{w}^{(k)})$ denotes the posterior probability. Thus, our surrogate function with regularization is given by

$$\mathbb{Q}(\boldsymbol{w}, \boldsymbol{w}^{(k)}) = \tfrac{1}{B}\sum_{b=1}^{B}\Bigg(\mathbb{J}_b(\boldsymbol{w}, \boldsymbol{w}^{(k)}) + \lambda_e \mathbb{H}_b(\boldsymbol{w}) + \frac{\lambda_q \|\boldsymbol{w}\|^2}{2}\Bigg). \quad (6)$$

In order to compute $\mathbb{Q}(\boldsymbol{w}, \boldsymbol{w}^{(k)})$, it is necessary to compute the posterior probability $P_{bic}^{post}(\boldsymbol{w}^{(k)})$. Using the conditional rule, this probability can be then determined by

$$P_{bic}^{post}(\boldsymbol{w}^{(k)}) = \frac{P(z_{bi} = c, Y_b, T_b \mid \boldsymbol{x}_{bi}, \boldsymbol{w}^{(k)})}{\sum_{t=0}^{\mathbb{C}} P(z_{bi} = t, Y_b, T_b \mid \boldsymbol{x}_{bi}, \boldsymbol{w}^{(k)})}. \quad (7)$$

We provide a detailed calculation of (7) in Section IV-C.

M-step. Since we consider the negative log-likelihood as the objective, this step is essentially to minimize the surrogate $\mathbb{Q}$ by solving

$$\boldsymbol{w}^{(k+1)} = \operatorname*{argmin}_{\boldsymbol{w}} \mathbb{Q}(\boldsymbol{w}, \boldsymbol{w}^{(k)}).$$

Generally, minimizing $\mathbb{Q}(\boldsymbol{w}, \boldsymbol{w}')$ is a non-trivial optimization problem. Alternatively, we use the Generalized EM approach to facilitate a descent approach by taking steps along the gradient of $\mathbb{Q}(\boldsymbol{w}, \boldsymbol{w}')$. By Fisher's identity, this gradient coincides with the gradient of the objective function

$$\frac{\partial \mathbb{L}(\boldsymbol{w})}{\partial \boldsymbol{w}}\Big|_{\boldsymbol{w}=\boldsymbol{w}'} = \frac{\partial \mathbb{Q}(\boldsymbol{w}, \boldsymbol{w}')}{\partial \boldsymbol{w}}\Big|_{\boldsymbol{w}=\boldsymbol{w}'} = \tfrac{1}{B}\textstyle\sum_{b=1}^{B} \frac{\partial \mathbb{Q}_b(\boldsymbol{w}, \boldsymbol{w}')}{\partial \boldsymbol{w}}\Big|_{\boldsymbol{w}=\boldsymbol{w}'}.$$

The gradient $\frac{\partial \mathbb{Q}_b(\boldsymbol{w}, \boldsymbol{w}')}{\partial \boldsymbol{w}_c}$ can be computed as follows

$$\frac{\partial \mathbb{Q}_b(\boldsymbol{w}, \boldsymbol{w}')}{\partial \boldsymbol{w}_c} = \sum_{i=1}^{n_b}\Big(P_{bic}(\boldsymbol{w}) - P_{bic}^{post}(\boldsymbol{w}')\Big)\boldsymbol{x}_{bi} + \lambda_e \sum_{i=1}^{n_b} P_{bic}(\boldsymbol{w})\Big(\sum_{t=0}^{\mathbb{C}} P_{bit}(\boldsymbol{w})(\boldsymbol{w}_t - \boldsymbol{w}_c)^T \boldsymbol{x}_{bi}\Big)\boldsymbol{x}_{bi} + \lambda_q \boldsymbol{w}_c. \quad (8)$$

The full gradient involves enumerating all the bags. To reduce the computation per iteration, we further propose a single-bag-based stochastic gradient descent approach in the flavor of the Pegasos algorithm [47], [48]. At the kth iteration, a random bag b_k is chosen and a single-bag-based gradient is computed as follows

$$\boldsymbol{w}^{(k+1)} = \Pi_\tau\Bigg(\boldsymbol{w}^{(k)} - \eta_k \frac{\partial \mathbb{Q}_{b_k}(\boldsymbol{w}, \boldsymbol{w}^{(k)})}{\partial \boldsymbol{w}}\Big|_{\boldsymbol{w}=\boldsymbol{w}^{(k)}}\Bigg),$$

where $\eta_k = \frac{1}{k\lambda_q}$, $\tau = \sqrt{\frac{2(\lambda_e+1)n_b}{\lambda_q}\log(\mathbb{C}+1)}$, and $\Pi_\tau(\boldsymbol{v}) = \min\Big\{1, \frac{\tau}{\|\boldsymbol{v}\|}\Big\}\boldsymbol{v}$. Detailed derivation of the stochastic gradient update is provided in Appendix I.

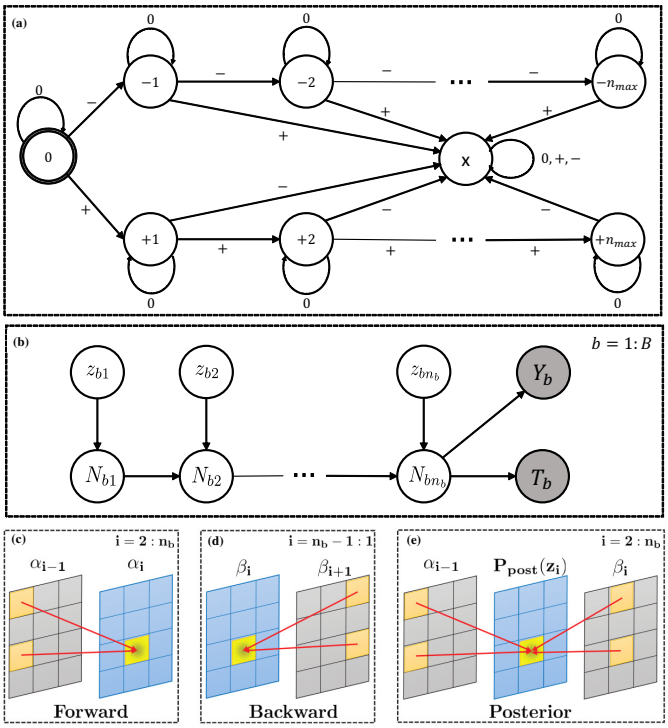

Fig. 3. (a) The number of positively- (negatively-) labeled instances in the first i instances, N_{bi}, as a finite state machine. (b) A reformulation of Fig. 2 as a chain on N_{bi}[2], and (c)-(e) recursive calculation of forward, backward messages, and posterior probability.

C. Proposed Dynamic Programming for E-step

The brute-force calculation of $P(z_{bi} = c, Y_b, T_b \mid \boldsymbol{x}_{bi}, \boldsymbol{w}^{(k)})$ in (7) requires marginalization over all other instance attributes (i.e., z_{bj} for $j = 1, \ldots, n_b$ and $j \neq i$), which is exponential in the number of instances per bag $\big(O(\mathbb{C}^{n_b-1})\big)$. Traditional approaches to address such intractable problems often resort to approximation techniques such as approximate inference [49], variational approximation [50], and black-box alpha [51]. In a recent line of work [39], [40], [52], Pham *et al.* proposed a dynamic programming approach for exact and efficient computation of the posterior. Motivated by the result, we follow a similar idea of converting the V-structure to the chain structure for efficient inference. Specifically, we define a new latent variable N_{bi} as the "signed" number of relevant attributes in the first i instances of the bth bag. The representation of N_{bi} is given by the finite state machine in Fig. 3(a). Each instance in a bag is represented by one of the input symbols $\{0, +, -\}$ (corresponding to an irrelevant instance, a relevant instance in the positive class, or a relevant instance in the negative class, respectively). The number of relevant instances is represented by the set of states $\{0, \pm1, \pm2, \ldots, \pm n_{max}\}$, where the sign indicates whether instances are belong to the positive class or the negative class. For the completeness, we introduce the error state $\{\mathrm{x}\}$ for the case there is a mix of positive and negative attributes in the bag. Practically, this state will not be reached due to our no mixed-class attributes assumption. Thanks to the introduction of N_{bi}, the posterior probability can be calculated

[2]For brevity we omit the $\boldsymbol{x}_{bi}$'s and $\boldsymbol{w}$.

efficiently using the forward and backward message passing on the chain structure (see Fig. 3(b)-(e)). To further simplify our forward-backward message passing derivation, let us denote

$$P_{bi}^0 = P_{bi0}(\boldsymbol{w}),\ P_{bi}^+ = \sum_{c\in C^+} P_{bic}(\boldsymbol{w}),\ P_{bi}^- = \sum_{c\in C^-} P_{bic}(\boldsymbol{w}) \quad (9)$$

where the parameter $\boldsymbol{w}$ is omitted for simplicity. Below is the details of our dynamic programming algorithm for computing the posterior.

Step 1. Forward message passing.
This step computes the forward messages defined as

$$\alpha_{bi}(l) \triangleq P(N_{bi} = l \mid \boldsymbol{X}_b, \boldsymbol{w}) \text{ for } l = 0, \pm 1, \ldots, \pm n_{\max}.$$

The first message is initialized by

$$\alpha_{b1}(l) = I_{l=0}P_{b1}^0 + I_{l=1}P_{b1}^+ + I_{l=-1}P_{b1}^-. \quad (10)$$

The update equation for subsequent messages is given by

$$\begin{aligned}\alpha_{bi}(l) = P_{bi}^0\alpha_{b(i-1)}(l) + I_{l>0}P_{bi}^+\alpha_{b(i-1)}(l-1) \\ + I_{l<0}P_{bi}^-\alpha_{b(i-1)}(l+1),\end{aligned} \quad (11)$$

for $i = 2, \ldots, n_b$.

Step 2. Backward message passing.
This step computes the backward messages defined as

$$\beta_{bi}(l) \triangleq P(Y_b, T_b \mid N_{bi} = l, \boldsymbol{X}_b, \boldsymbol{w}).$$

The first backward message is initialized by

$$\beta_{bn_b}(l) = I_{Y_b=1}I_{0<l\le n_{\max}} + I_{Y_b=0}I_{-n_{\max}\le l<0}. \quad (12)$$

The update equation for subsequent messages is given by

$$\begin{aligned}\beta_{bi}(l) =& P_{b(i+1)}^0\beta_{i+1}(l) + I_{0\le l<n_{\max}}P_{b(i+1)}^+\beta_{i+1}(l+1) \\ &+ I_{0\ge l>-n_{\max}}P_{b(i+1)}^-\beta_{i+1}(l-1),\end{aligned} \quad (13)$$

for $i = n_b - 1, \ldots, 1$.

Step 3. Joint probability calculation.
This step computes the joint probability defined as

$$P_{bic}^{joint}(\boldsymbol{w}) \triangleq P(z_{bi} = c, Y_b, T_b \mid \boldsymbol{x}_{bi}, \boldsymbol{w}),\ \text{for } c = 0, 1, \ldots, \mathbb{C}.$$

First, initialize $P_{b1c}^{joint}(\boldsymbol{w})$ by

$$\big(I_{c=0}\beta_1(0) + I_{c\in C^+}\beta_1(1) + I_{c\in C^-}\beta_1(-1)\big)P_{b1c}(\boldsymbol{w}). \quad (14)$$

Next, for $i = 2, \ldots, n_b$, perform the update for $P_{bic}^{joint}(\boldsymbol{w})$ by

$$\begin{aligned}\Big(I_{c=0}\sum_{l=-n_{\max}}^{+n_{\max}}\beta_i(l)\alpha_{i-1}(l) + I_{c\in C^+}\sum_{l=0}^{n_{\max}-1}\beta_i(l+1)\alpha_{i-1}(l) \\ + I_{c\in C^-}\sum_{l=-n_{\max}+1}^{0}\beta_i(l-1)\alpha_{i-1}(l)\Big)P_{bic}(\boldsymbol{w}).\end{aligned} \quad (15)$$

The detailed derivation for the forward messages, backward messages, and joint probability calculation are given in Appendix II, III, and IV, respectively. We summarize the proposed AbSMIL approach in Algorithm 1.

Algorithm 1 Attribute-based Symmetric Multiple Instance Learning (AbSMIL)

Input: Training data $\{\boldsymbol{X}_b, Y_b\}_{b=1}^B$, cardinality constraint $n_{\max}$, positive constants λ_q and λ_e, initial weight $\boldsymbol{w}^{(0)}$
Output: $\{\boldsymbol{w}^{(k)}\}$

$k = 0$
repeat
  Select a random bag b
  // **E-step:**
  Compute prior probability $P_{bic}(\boldsymbol{w}^{(k)})$ using (1)
  Compute prior probability P_{bi}^0, P_{bi}^+ and P_{bi}^- using (9)
  Compute forward message $\alpha_{bi}(l)$ for $i = 1, \ldots, n_b$ and $l = 0, \pm 1, \ldots, \pm n_{\max}$ using (10) and (11)
  Compute backward message $\beta_{bi}(l)$ for $i = n_b, \ldots, 1$ and $l = 0, \pm 1, \ldots, \pm n_{\max}$ using (12) and (13)
  Compute joint probability $P_{bic}^{joint}(\boldsymbol{w}^{(k)})$ for $i = 1, \ldots, n_b$ and $c = 0, 1, \ldots, \mathbb{C}$ using (14) and (15)
  Compute posterior probability $P_{bic}^{post}(\boldsymbol{w}^{(k)})$ for $i = 1, \ldots, n_b$ and $c = 0, 1, \ldots, \mathbb{C}$ using (7)
  // **M-step:**
  $\tau = \sqrt{\frac{2(\lambda_e+1)n_b}{\lambda_q}\log(\mathbb{C}+1)}$
  Compute $\frac{\partial \mathbb{Q}_b(\boldsymbol{w}, \boldsymbol{w}^{(k)})}{\partial \boldsymbol{w}_c}$ for $c = 0, 1, \ldots, \mathbb{C}$ using (8)
  $\boldsymbol{w}^{(k+1)} = \Pi_\tau\Big(\boldsymbol{w}^{(k)} - \frac{1}{k\lambda_q}\frac{\partial \mathbb{Q}_b(\boldsymbol{w}, \boldsymbol{w}^{(k)})}{\partial \boldsymbol{w}^{(k)}}\Big)$
  $k = k + 1$
until stopping criteria is met

D. Prediction

After learning the weight vector $\boldsymbol{w}$, the instance level label for new test data can be predicted as follows:

$$\hat{z}_{bi} = \text{argmax}_{0\le c\le \mathbb{C}} P(z_{bi} = c \mid \boldsymbol{x}_{bi}, \boldsymbol{w}).$$

Note that this prediction is made without knowing the bag label. To this end, the bag level label can be predicted as

$$\hat{Y}_b = \text{argmax}_{m\in\{0,1\}} P(Y_b = m, T_b = 1 \mid \boldsymbol{X}_b, \boldsymbol{w})$$

where $P(Y_b = m, T_b = 1 \mid \boldsymbol{X}_b, \boldsymbol{w})$ is given by

$$\begin{cases} \sum_{l=1}^{n_{\max}} \alpha_{bn_b}(l) & \text{for } m = 1, \\ \sum_{l=-n_{\max}}^{-1} \alpha_{bn_b}(l) & \text{for } m = 0. \end{cases}$$

E. Complexity Analysis

To compute the posterior probability $P_{bic}^{post}(\boldsymbol{w}')$, we need to obtain the forward and backward messages over all instances of the bag (n_b) and all possible numbers of relevant instances ($2n_{\max} + 1$). Given the label Y_b of the bag, the overall complexity of the E-step is $O(n_b n_{\max})$. Our proposed dynamic programming approach offers an efficient computation that is *linear* with the number of instances per bag n_b when the number of relevant instances is constrained to be small. On the other hand, the M-step requires $O(n_b\mathbb{C}d)$ to compute each single bag-based stochastic gradient. Thus, the total complexity per iteration is $O\big(n_b(\mathbb{C}d + n_{\max})\big)$. When each instance is a high-dimensional vector, i.e., $d \gg n_{\max}$, the

M-step dominates other steps and the overall complexity per iteration of our algorithm is $O(\mathbb{C}n_b d)$. In terms of memory, the dominant factor stems from forward and backward messages. In order to store all possible messages, the space complexity per bag is $O(n_b n_{\max})$, which is often smaller than that of the instances per bag ($O(dn_b)$) in practice.

Non-linear extension. If there is a large number of attributes, a kernel extension can be used as an alternative to the linear model. By introducing non-linear kernel functions, data is transformed to capture the clustering nature of multiple attributes without assigning additional clusters. In practice, $\mathbb{C} = 2$ is often sufficient for binary classification. To implement a radial basis function (RBF) kernel, one can follow the approach of [53], replacing $\boldsymbol{x}$ with

$$\boldsymbol{\phi}(\boldsymbol{x}) = \frac{1}{\sqrt{k}}\left[\cos(\boldsymbol{g}_1^T\boldsymbol{x}), \sin(\boldsymbol{g}_1^T\boldsymbol{x}), \ldots, \cos(\boldsymbol{g}_k^T\boldsymbol{x}), \sin(\boldsymbol{g}_k^T\boldsymbol{x})\right]^T \tag{16}$$

where $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{0}_d, \sigma_g^2 \boldsymbol{I}_d)$. We demonstrate the effectiveness of the non-linear extension through the experiment in Section V-D.

V. Experiments

In this section, we evaluate the performance of the proposed algorithm (AbSMIL) in terms of runtime and classification accuracy. We compare its performance with state-of-the-art MIL frameworks on a number of datasets, including 4 different histopathology datasets.

A. Datasets

We consider 2 sets of experiments to evaluate the runtime (see V-C) and classification performance (see V-D). Firstly, in order to evaluate the runtime performance of AbSMIL, we create a synthetic MIL-based dataset from the MNIST dataset as follows. We select all the images of digits 0, 1, 2 and manually label the right/left-tilted 0s and 1s as the positive/negative relevant instances, and 2s as the irrelevant instances. Then, for each positive (negative) bag, we randomly select n_{rel} relevant instances from the images of right-tilted (left-tilted) digits and $n_b - n_{rel}$ irrelevant instances from the images of 2. The total number of bags is $B = 200$, with a balanced number between positive and negative bags. The goal is to learn the orientation (left-tilted versus right-tilted) of 0 and 1 while the orientation of 2 is ignored.

Secondly, in order to evaluate the classification performance of AbSMIL, we use seven benchmark datasets in a wide range of applicability. **(i)** The first group includes three datasets popularly used in studies of MIL: Tiger, Fox, and Elephant datasets (see [12], [27], [28], [54], [55]). There are 200 bags in which 100 positive bags associated with the target animal images and 100 negative bags associated with other kinds of animal. For each of these datasets, 140 images are used for training and 60 images for test. The maximum number of instance per bag is 13. More details of these datasets can be found in [27]. **(ii)** The second group contains histopathology images of mammalian organs, provided by the Animal Diagnostics Lab (ADL) at Pennsylvania State University. Each of the three datasets (kidney, lung and spleen) contains 300 images of size 4000×3000 from either inflammatory or healthy tissues. A healthy tissue image largely consists of healthy patches while an inflammatory tissues have a dominant portion of diseased patches. 250 samples are used for training and the remaining ones are used for testing. There are 130 instances per bag in each dataset. More details of ADL datasets can be found in [11]. **(iii)** The last dataset, referred as the TCGA dataset [56], contains WSIs of brain cancer from The Cancer Genome Atlas (TCGA), provided by the National Institute of Health. This dataset contains 96 histopathology samples of two types of glioma: 48 samples for astrocytoma and 48 samples for oligodendroglioma. In each WSI (as one bag), cancerous regions occupy only a small portion of various shapes and color shading, and this portion is usually surrounded by benign cells, making the TCGA dataset inherently harder than the ADL datasets [11]. Noticeably, the problem of classifying cancer types in the TCGA dataset fits well the symmetric setting of AbSMIL. To obtain instances from each WSI, we first remove some redundant parts such as glass or folding areas, then randomly select a set of $100 \times 100 \times 3$ patches in the image (with potential overlaps). The instances are obtained by featurizing these patches using 84 features including histogram of oriented gradients, histogram of gray images and SFTA-texture features [57]. The maximum number of instances per bag is 1258. A split of 76 training versus 20 test images is considered in our experiment. In all of these datasets, a balanced number of positive/negative bags is used in both training and testing stages.

B. Baselines

In our experiment, we compare the proposed method with a variety of popular approaches, including mi-SVM [27], MIL-Boost [33], miGraph [28], MCIL [26], DFDL [11], ORLR [39] and MIML-NC [40]. The first four methods are MIL-based approaches that utilize the standard asymmetric assumption. In particular, mi-SVM assumes that there is at least one pattern from every positive bag in the positive halfspace, while all patterns belonging to negative bags are in the negative halfspace. In [28], miGraph implicitly constructs a graph that model each bag and the relations among the instances within the bag. Both of these methods were previously used on the Tiger, Fox, Elephant datasets. In histopathological image classification, MCIL and DFDL are the state-of-the-art methods. While MCIL is designed for MIL setting and can be used directly in our experiment, DFDL learns a dictionary bases using manually extracted regions in the WSIs (i.e., annotating at instance level). In order to adapt DFDL to the MIL setting, we resort to assuming that all instances in positive bags are positive and all instances in negative bags are negative (similar to [11]). MIL-Boost can be seen as a special case of MCIL where there is one cluster in positive bags. The last two methods, ORLR and MIML-NC, are generally designed for the multiple instance multiple label learning (MIML) setting. As discussed in Section II, MIML setting is similar to our symmetric MIL assumption where both positive and negative bag contains instances from multiple clusters. However, while

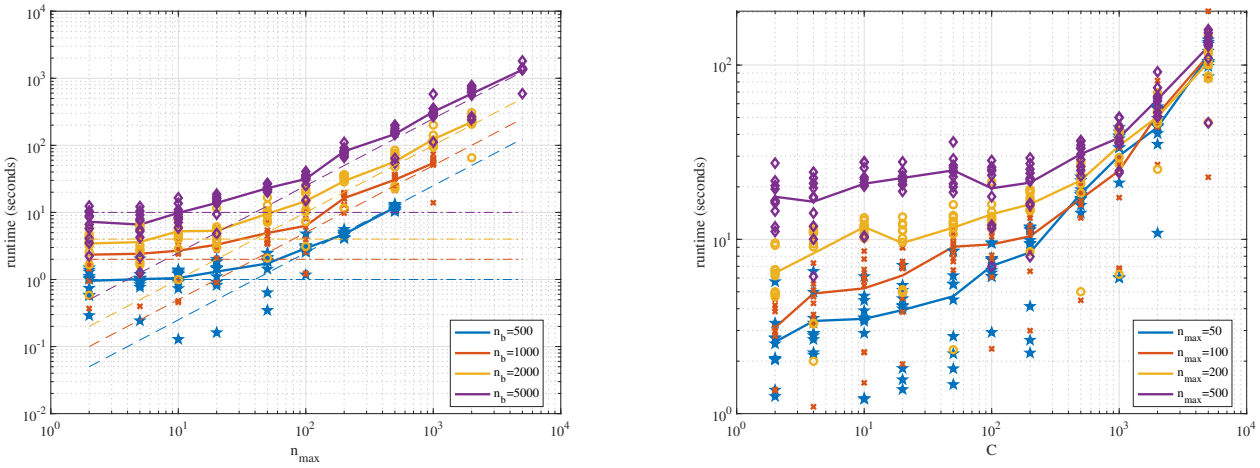

Fig. 4. Running time of AbSMIL as a function of the cardinality constraint $\boldsymbol{n_{\max}}$ (a) and the number of attributes $\mathbb{C}$ (b) in log-log scale. Each curve corresponds to a different setting in terms of $\boldsymbol{n_b}$ (a) and in terms of $\boldsymbol{n_{\max}}$ (b). The runtime values in each curve are calculated by averaging the runtime across $\mathbf{10}$ different runs (indicated by the markers). In (a), the dash-dotted lines and the dashed lines are added to demonstrate the asymptotic behavior of analytical complexity $\boldsymbol{O\big(n_b(\mathbb{C}d + n_{\max})\big)}$ when $\boldsymbol{n_{\max}}$ is small and when $\boldsymbol{n_{\max}}$ is large, respectively.

the MIML-based models are designed for a more general setting where the bag label is the union of all instance labels, the proposed AbSMIL method is designed for the specific setting of HIC where bag labels are binary based on the relevant instances.

C. Runtime Evaluation

In the following, we present the experiment to evaluate and compare the runtime performance of AbSMIL with the aforementioned algorithms.

Setting. In the first set of simulations, we test the impact of the cardinality constraint $n_{\max}$ on the computational performance of AbSMIL by varying n_b through the set of values $\{500, 1000, 2000, 5000\}$ and $n_{\max}$ through the set of values $\{2, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000\}$ (such that $n_{\max} \leq n_b$). Since among the aforementioned values we include fairly large values of n_b and $n_{\max}$, we consider the following to reduce the runtime. First, the original $784 = 28 \times 28$ image vector is trimmed down to a 30-dimensional vector by selecting only the first $d = 30$ elements. Next, the number of attributes is set to $\mathbb{C} = 2$ and the parameters λ_q and λ_e are fixed to 10^{-6} and 10^{-2}, respectively. We also adjust the number of epochs inversely proportional to $n_{\max}$ so that each setting takes approximately the same amount of time. Finally, we report the average running time per epoch for each setting. In the second set of simulations, we test the impact of the number of attributes $\mathbb{C}$ on the computational performance of AbSMIL by varying $n_{\max}$ through the set of values $\{50, 100, 200, 500\}$ and $\mathbb{C}$ through the set of values $\{2, 4, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000\}$. The same setting for $\mathbb{C}$, d, λ_q and λ_e from the first simulation is used in this simulation and the average running time per epoch is then reported. To reduce the runtime variance, each setting is repeated 10 times on the same computer.

Results and Analysis. We plot our results from the simulations in Fig. 4. The left plot shows the average running time of AbSMIL as a function of $n_{\max}$. It can be seen that the running time exhibits two different modes with respect to the change in $n_{\max}$. Recall that our overall complexity per bag is $O\big(n_b(\mathbb{C}d + n_{\max})\big)$. When $n_{\max}$ is small, the dominant term in the sum is $\mathbb{C}d$ (here, $\mathbb{C}d = 60$) and we observe a flat region at the beginning of the four curves for different values of n_b. When $n_{\max}$ becomes larger, linear behavior (i.e., runtime $\propto n_{\max}$) is observed as $n_{\max}$ gains dominance over $\mathbb{C}d$. Note that in practice, the implementation of AbSMIL for the special case when $n_{\max} = n_b$ is less costly than reported and this case is equivalent to AbSMIL with no constraint. In the right plot, a similar behavior is observed: the running time remains stable when $\mathbb{C}$ is small and a linearly increasing runtime is observed when $\mathbb{C}$ is large. Note that for different values of $n_{\max}$, the running time converges since the $\mathbb{C}d$ becomes dominant, thereby making the different values of $n_{\max}$ negligible. Similar to the number of weak classifiers T in MCIL, the number of attributes $\mathbb{C}$ in our algorithm also contributes as a linear term in the computational complexity. We also notice variation in the running time due to the differences among the computational cluster nodes used in this experiment.

D. Real-world Datasets

This subsection presents the experiment to evaluate and compare the accuracy performance of AbSMIL with the baseline algorithms on different real-world datasets.

Settings. Since instance labels are unavailable for the aforementioned real-world datasets, all the results are evaluated using bag level prediction with 10-fold cross validation [61]. The value of $n_{\max}$ is selected in the set $\{5\%, 10\%, 20\%, 30\%, 40\%, 50\%, 100\%\}$ of n_b in Tiger, Fox, and Elephant datasets; $\{5\%, 10\%, 20\%, 50\%, 75\%\}$ of n_b in Kidney, Lung, Spleen datasets; and $\{1\%, 2\%, 5\%, 10\%, 20\%\}$

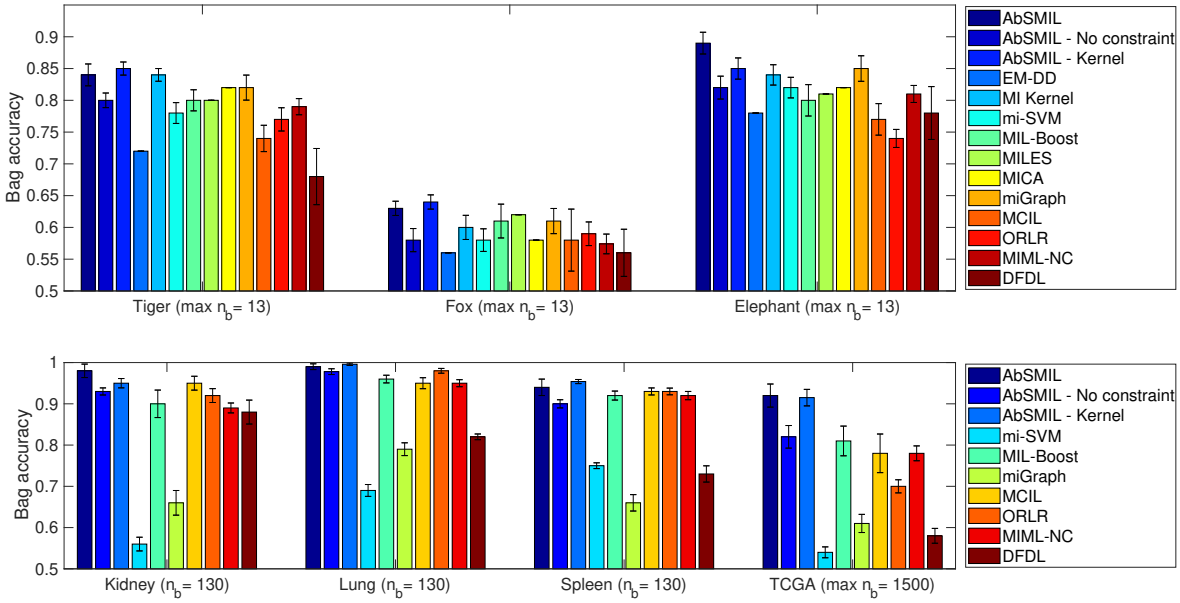

Fig. 5. Bag level accuracy various algorithms in seven real-world datasets. The proposed AbSMIL method shows a competitive performance with other state-of-the-art methods for various datasets in our experiment. Especially for the case of TCGA dataset, the outstanding result of AbSMIL indicates that the symmetric MIL assumption and the cardinality constraint are more suitable to such settings. The results with EM-DD [31], MI Kernel [58], MILES [59], and MICA [60] are observed from [60] and [55].

of n_b in TCGA dataset. Note that $\boldsymbol{x}_{bi}$'s are assumed to be normalized such that the mean of each entry over the data is zero and the variance of each entry is 1. In AbSMIL, λ_q and λ_e are searched in a grid of $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$. For each setting, AbSMIL is initialized ten times, and the model that yields the lowest training negative log-likelihood is chosen to report the performance. We also report the results for two variants of the proposed method: AbSMIL - No constraint and AbSMIL - Kernel. AbSMIL - No constraint is essentially AbSMIL without using cardinality constraint. AbSMIL - Kernel is mentioned in (16), and its hyperparameter σ_g^2 is selected in $\{0.999, 0.998, 0.995, 0.99, 0.98, 0.95, 0.9, 0.8, 0.5\}$. For mi-SVM, we change the loss constant $C \in \{10^{-2}, 10^{-1}, 1, 10, 10^2\}$ and choose a linear kernel function. In MIL-Boost and MCIL, we search over the generalized mean (GM), the log-sum-exponential (LSE) softmax function with parameter named $r \in \{15, 20, 25\}$ and the number of weak classifiers named $T \in \{150, 200, 250\}$. In miGraph, we tune the number of classes $C \in \{50, 100, 150, 200, 500\}$, RBF kernel $\gamma \in \{1, 5, 10, 15, 20, 25, 50\}$ and threshold used in computing the weight of each instance $\epsilon \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. ORLR and MIML-NC are tuning-free methods. For DFDL, we tune the regularization parameter $\rho \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, sparsity level parameter $\lambda \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$, and the number of dictionary bases $k \in \{100, 200, 500\}$. These tuning values are extracted from the corresponding aforementioned publications.

Results and Analysis. Fig. 5 demonstrates the bag level accuracy of the aforementioned methods on different real-world datasets. Overall, the proposed AbSMIL method shows a competitive performance with other state-of-the-art methods for various datasets in our experiment. Additionally, we observe a major improvement by adding the cardinality constraint to the AbSMIL model. Let us discuss the detail of each group of datasets below.

Tiger, Fox, and Elephant datasets: Overall, Fig. 5 shows that AbSMIL obtains the highest accuracy on Elephant dataset, while AbSMIL - Kernel outperforms other methods on Fox and Tiger datasets. As pointed out in [27], due to the limited accuracy of the image segmentation, the relatively small number of region descriptors, and the small training set size, Fox dataset yields a harder classification problem than the other two datasets. Regarding result of 'AbSMIL-No constraint', it can be seen that removing the cardinality constraint lessens the performance of AbSMIL significantly (e.g., roughly 7%).

Kidney, Lung, and Spleen datasets: The results reported for AbSMIL are obtained by setting $n_{\max}$ at 20% of the total instances per bag in Kidney and Lung datasets and at 50% in Spleen dataset. Again, it can be seen that AbSMIL and AbSMIL - Kernel slightly outperform other methods in terms of bag level accuracy. A comparison of images of instances from the same attribute obtained by AbSMIL and DFDL on Kidney dataset are shown in Fig. 6. For $\mathbb{C} = 2$, the cancerous tissue recognized by AbSMIL appears as a mixture of *vascular proliferation* (red parts) and *necrosis* (areas with no *nuclei* - appeared as the purple dot in the image). On the other hand, for $\mathbb{C} = 4$, cancerous tissue splits into two distinctly different categories. In comparison, the performance of DFDL appears worse since in this MIL setting, DFDL assumes that all instances in positive bags are positive and all instances in negative bags are negative.

TCGA dataset: The result obtained by AbSMIL in this experiment is at $n_{\max}$ of nearly 5% the total number of instances

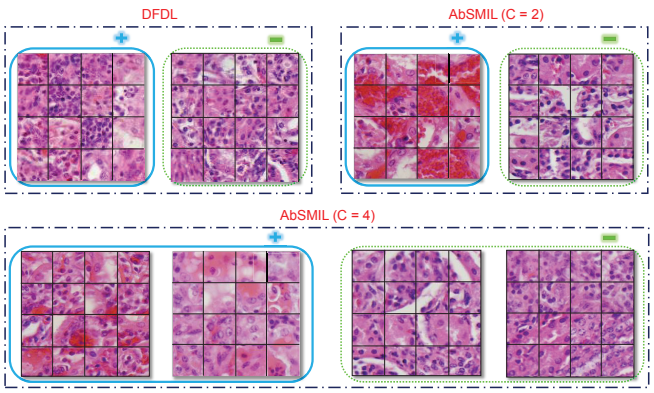

Fig. 6. Groups of instances in the same class predicted by AbSMIL and DFDL. Cancer groups are in solid blue while normal ones in dotted green. Compared to two DFDL bases, AbSMIL exhibits distinctly different tissue types between the two categories with $\mathbb{C} = \mathbf{2}$ and can be clustered to distinguish phenomenon in each category with $\mathbb{C} = \mathbf{4}$.

in a bag. As in Fig. 5, AbSMIL and AbSMIL - Kernel significantly outperform the other frameworks. This improvement is remarkable compared to the previous experiment and matches with our intuition on the TCGA dataset: in each WSI, cancerous regions occupy only a small portion, as opposed to the ADL dataset where inflammatory tissues dominate the entire WSI. Thereby, the cardinality constraint offers an advantage to AbSMIL over other methods in effectively gathering information from patches.

VI. Conclusion

In this paper, we introduced Symmetric MIL, a novel setting for multiple-instance learning where both positive and negative bags contain relevant class-specific instances as well as irrelevant instances that do not contribute to differentiating the classes. We presented a probabilistic model for attribute-based Symmetric MIL that accommodates for the presence of numerous irrelevant instances in the data, and takes into account prior information about the sparsity of the relevant instances. We developed an efficient inference approach that is linear in the number of instances and is suitable for the online learning scenario, updating the model using one bag at a time. We evaluated our framework on the real-world datasets: Tiger, Fox, Elephant, Kidney, Lung, Spleen, and TCGA. We obtained competitive results on all datasets and in particular for TCGA where bags contain mainly irrelevant instances. The results validate the merit of the proposed symmetric MIL framework.

Appendix I
Gradient Derivation

From (6), the gradient of $\mathbb{Q}_b$ w.r.t. $\boldsymbol{w}_c$ can be decomposed into

$$\frac{\partial \mathbb{Q}_b(\boldsymbol{w}, \boldsymbol{w}')}{\partial \boldsymbol{w}_c} = \frac{\partial \mathbb{J}_b(\boldsymbol{w}, \boldsymbol{w}')}{\partial \boldsymbol{w}_c} + \frac{\partial \mathbb{H}_b(\boldsymbol{w})}{\partial \boldsymbol{w}_c} + \lambda_q \boldsymbol{w}_c. \tag{17}$$

On the one hand, differentiating $\mathbb{J}_b$ from (5) yields

$$\frac{\partial \mathbb{J}_b(\boldsymbol{w}, \boldsymbol{w}')}{\partial \boldsymbol{w}_c} = -\sum_{i=1}^{n_b} P_{bic}^{post}(\boldsymbol{w}')\boldsymbol{x}_{bi} + \sum_{i=1}^{n_b} P_{bic}(\boldsymbol{w})\boldsymbol{x}_{bi}.$$

On the other hand, differentiating $\mathbb{H}_b$ from (2) yields

$$\begin{aligned}
\frac{\partial \mathbb{H}_b(\boldsymbol{w})}{\partial \boldsymbol{w}_c} &= \frac{\partial}{\partial \boldsymbol{w}_c}\Big(-\sum_{i=1}^{n_b}\sum_{t=0}^{\mathbb{C}} P_{bit}(\boldsymbol{w}) \log P_{bit}(\boldsymbol{w})\Big) \\
&= -\sum_{i=1}^{n_b}\sum_{t=0}^{\mathbb{C}}\Big(1 + \log P_{bit}(\boldsymbol{w})\Big)P_{bit}(\boldsymbol{w})\Big(I_{t=c} - P_{bic}(\boldsymbol{w})\Big)\boldsymbol{x}_{bi} \\
&= \sum_{i=1}^{n_b} P_{bic}(\boldsymbol{w})\boldsymbol{x}_{bi}\Big(\sum_{t=0}^{\mathbb{C}} P_{bit}(\boldsymbol{w})\big(\log P_{bit}(\boldsymbol{w}) - \log P_{bic}(\boldsymbol{w})\big)\Big) \\
&= \sum_{i=1}^{n_b} P_{bic}(\boldsymbol{w})\Big(\sum_{t=0}^{\mathbb{C}} P_{bit}(\boldsymbol{w})(\boldsymbol{w}_t - \boldsymbol{w}_c)^T\boldsymbol{x}_{bi}\Big)\boldsymbol{x}_{bi}.
\end{aligned}$$

Thus, substituting the results back into (17) yields

$$\begin{aligned}
\frac{\partial \mathbb{Q}_b(\boldsymbol{w}, \boldsymbol{w}')}{\partial \boldsymbol{w}_c} = \sum_{i=1}^{n_b}\Big(P_{bic}(\boldsymbol{w}) - P_{bic}^{post}(\boldsymbol{w}')\Big)\boldsymbol{x}_{bi} + \\
\lambda_e \sum_{i=1}^{n_b} P_{bic}(\boldsymbol{w})\Big(\sum_{t=0}^{\mathbb{C}} P_{bit}(\boldsymbol{w})(\boldsymbol{w}_t - \boldsymbol{w}_c)^T\boldsymbol{x}_{bi}\Big)\boldsymbol{x}_{bi} + \lambda_q \boldsymbol{w}_c.
\end{aligned}$$

The ball radius τ is then computed as follows. Let $\boldsymbol{w}^* = \operatorname{argmin}_{\boldsymbol{w}} \mathbb{Q}_b(\boldsymbol{w}, \boldsymbol{w}')$. Then the following always holds

$$\begin{aligned}
& \mathbb{Q}_b(\boldsymbol{w}^*, \boldsymbol{w}') \le \mathbb{Q}_b(\mathbf{0}, \boldsymbol{w}') \\
\Leftrightarrow\ & \mathbb{J}_b(\boldsymbol{w}^*, \boldsymbol{w}') + \lambda_e \mathbb{H}_b(\boldsymbol{w}^*) + \frac{\lambda_q \|\boldsymbol{w}^*\|^2}{2} \\
& \qquad \le \sum_{i=1}^{n_b} \log(\mathbb{C}+1) + \lambda_e \sum_{i=1}^{n_b} \log(\mathbb{C}+1) \\
\Rightarrow\ & \frac{\lambda_q \|\boldsymbol{w}^*\|^2}{2} \le (\lambda_e + 1) n_b \log(\mathbb{C}+1) \qquad (*) \\
\Leftrightarrow\ & \|\boldsymbol{w}^*\|^2 \le \frac{2(\lambda_e+1) n_b \log(\mathbb{C}+1)}{\lambda_q} \\
\Leftrightarrow\ & \tau = \|\boldsymbol{w}^*\| \le \sqrt{\frac{2(\lambda_e+1) n_b}{\lambda_q} \log(\mathbb{C}+1)}
\end{aligned}$$

where $(*)$ stems from the fact that both $\mathbb{J}_b(\boldsymbol{w}^*, \boldsymbol{w}')$ and $\mathbb{H}_b(\boldsymbol{w}^*)$ are non-negative.

Appendix II
Forward Message Derivation

Initialize $\alpha_i(l)$ for $i = 1$ and $l = -n_{\max}, \ldots, +n_{\max}$:

$$\begin{aligned}
\alpha_1(l) &= P(N_1 = l \mid \boldsymbol{X}_b, \boldsymbol{w}) \\
&= \sum_c P(N_1 = l, z_{b1} = c \mid \boldsymbol{x}_{b1}, \boldsymbol{w}) \\
&= \sum_c P(N_1 = l \mid z_{b1} = c) P_{b1c}(\boldsymbol{w}) \\
&= \sum_c \Big(I_{l=0} I_{c=0} + I_{l=1} I_{c\in C^+} + I_{l=-1} I_{c\in C^-}\Big) P_{b1c}(\boldsymbol{w}) \\
&= I_{l=0} P_{b10}(\boldsymbol{w}) + I_{l=1} \sum_{c\in C^+} P_{b1c}(\boldsymbol{w}) + I_{l=-1} \sum_{c\in C^-} P_{b1c}(\boldsymbol{w}) \\
&= I_{l=0} P_{b1}^0 + I_{l=1} P_{b1}^+ + I_{l=-1} P_{b1}^-.
\end{aligned}$$

Update $\alpha_i(l)$ for $i = 2, \ldots, n_b$ and $l = -n_{\max}, \ldots, +n_{\max}$:

$$\alpha_i(l) = P(N_i = l \mid \boldsymbol{X}_b, \boldsymbol{w})$$

$$= \sum_{c,k} P(N_i = l, N_{i-1} = k, z_{bi} = c \mid \boldsymbol{X}_b, \boldsymbol{w})$$
$$= \sum_{c,k} P(N_i = l \mid N_{i-1} = k, z_{bi} = c) \cdot P(N_{i-1} = k \mid \boldsymbol{X}_b, \boldsymbol{w}) P_{bic}(\boldsymbol{w})$$
$$= \sum_{c,k} \Big(I_{c=0} I_{k=l} + I_{l \geq 1} I_{c \in C^+} I_{k=l-1} + I_{l \leq -1} I_{c \in C^-} I_{k=l+1} \Big) \alpha_{i-1}(k) P_{bic}(\boldsymbol{w})$$
$$= \alpha_{i-1}(l) P_{bi}^0 + I_{l>0} \alpha_{i-1}(l-1) P_{bi}^+ + I_{l<0} \alpha_{i-1}(l+1) P_{bi}^-.$$

Appendix III
Backward Message Derivation

Initialize $\beta_i(l)$ for $i = n_b$ and $l = -n_{\max}, \ldots, +n_{\max}$:

$$\begin{aligned} \beta_{n_b}(l) &= P(Y_b, T_b = 1 \mid N_{n_b} = l, \boldsymbol{X}_b, \boldsymbol{w}) \\ &= P(T_b = 1 \mid N_{n_b} = l) P(Y_b \mid N_{n_b} = l) \\ &= I_{Y_b=1} I_{0<l\leq n_{\max}} + I_{Y_b=0} I_{-n_{\max}\leq l<0}. \end{aligned}$$

Update $\beta_i(l)$ for $i = n_b - 1, \ldots, 1$, $l = -n_{\max}, \ldots, +n_{\max}$:

$$\begin{aligned} \beta_i(l) &= P(Y_b, T_b = 1 \mid N_i = l, \boldsymbol{X}_b, \boldsymbol{w}) \\ &= \sum_{c,k} P(Y_b, T_b = 1, N_{i+1} = k, z_{b(i+1)} = c \mid N_i = l, \boldsymbol{X}_b, \boldsymbol{w}) \\ &= \sum_{c,k} P(Y_b, T_b = 1 \mid N_{i+1} = k, \boldsymbol{X}_b, \boldsymbol{w}) \\ &\quad \cdot P(N_{i+1} = k \mid N_i = l, z_{b(i+1)} = c) P_{b(i+1)c}(\boldsymbol{w}) \\ &= \sum_{c,k} \beta_{i+1}(k) \Big(I_{c=0} I_{k=l} + I_{0\leq l<n_{\max}} I_{c\in C^+} I_{k=l+1} \\ &\quad + I_{0\geq l>-n_{\max}} I_{c\in C^-} I_{k=l-1} \Big) P_{b(i+1)c}(\boldsymbol{w}) \\ &= \beta_{i+1}(l) P^0_{b(i+1)}(\boldsymbol{w}) + I_{0\leq l<n_{\max}} \beta_{i+1}(l+1) P^+_{b(i+1)} \\ &\quad + I_{0\geq l>-n_{\max}} \beta_{i+1}(l-1) P^-_{b(i+1)}. \end{aligned}$$

Appendix IV
Joint Probability Derivation

For valid Y_b and T_b, the state machine will not travel through the invalid state x. So we can safely ignore the invalid state x in our derivation and only consider valid states $0, \pm 1, \ldots, \pm n_{\max}$. Initialize $P_{bic}^{joint}(\boldsymbol{w})$ for $i = 1$:

$$\begin{aligned} P_{b1c}^{joint}(\boldsymbol{w}) &= P(z_{b1} = c, Y_b, T_b = 1 \mid \boldsymbol{X}_b, \boldsymbol{w}) \\ &= \sum_k P(Y_b, T_b = 1, N_1 = k, z_{b1} = c \mid \boldsymbol{X}_b, \boldsymbol{w}) \\ &= \sum_k P(Y_b, T_b = 1 \mid N_1 = k, \boldsymbol{X}_b, \boldsymbol{w}) \\ &\quad \cdot P(N_1 = k \mid z_{b1} = c) P_{b1c}(\boldsymbol{w}) \\ &= \sum_k \beta_1(k) \Big(I_{c=0} I_{k=0} + I_{c\in C^+} I_{k=1} \\ &\quad + I_{c\in C^-} I_{k=-1} \Big) P_{b1c}(\boldsymbol{w}) \\ &= \Big(I_{c=0} \beta_1(0) + I_{c\in C^+} \beta_1(1) + I_{c\in C^-} \beta_1(-1) \Big) P_{b1c}(\boldsymbol{w}). \end{aligned}$$

Update rule for $i = 2, \ldots, n_b - 1$:

$$\begin{aligned} P_{bic}^{joint}(\boldsymbol{w}) &= P(z_{bi} = c, Y_b, T_b = 1 \mid \boldsymbol{X}_b, \boldsymbol{w}) \\ &= \sum_{k,l} P(Y_b, T_b = 1, N_i = k, N_{i-1} = l, z_{bi} = c \mid \boldsymbol{X}_b, \boldsymbol{w}) \\ &= \sum_{k,l} P(Y_b, T_b = 1 \mid N_i = k, \boldsymbol{X}_b, \boldsymbol{w}) \cdot \\ &\quad P(N_i = k \mid N_{i-1} = l, z_{bi} = c) P(N_{i-1} = l \mid \boldsymbol{X}_b, \boldsymbol{w}) P_{bic}(\boldsymbol{w}) \\ &= \sum_{k,l} \beta_i(k) \Big(I_{c=0} I_{k=l} + I_{l<n_{\max}} I_{c\in C^+} I_{k=l+1} \\ &\quad + I_{l>-n_{\max}} I_{c\in C^-} I_{k=l-1} \Big) \alpha_{i-1}(l) P_{bic}(\boldsymbol{w}) \\ &= \Big(I_{c=0} \sum_l \beta_i(l) \alpha_{i-1}(l) + I_{c\in C^+} \sum_{l=0}^{n_{\max}-1} \beta_i(l+1) \alpha_{i-1}(l) \\ &\quad + I_{c\in C^-} \sum_{l=-n_{\max}+1}^{0} \beta_i(l-1) \alpha_{i-1}(l) \Big) P_{bic}(\boldsymbol{w}). \end{aligned}$$

Acknowledgment

The authors would like to thank Dr. Souptik Barua at Rice University for his help with TCGA data organization.

References

[1] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener, "Histopathological image analysis: A review," *IEEE Rev. Biomed. Eng.*, vol. 2, pp. 147–171, 2009.

[2] D. Komura and S. Ishikawa, "Machine learning methods for histopathological image analysis," *Comput. Struct. Biotechnol. J.*, vol. 16, pp. 34–42, 2018.

[3] J. Kong, O. Sertel, H. Shimada, K. L. Boyer, J. H. Saltz, and M. N. Gurcan, "Computer-aided evaluation of neuroblastoma on whole-slide histology images: Classifying grade of neuroblastic differentiation," *Pattern Recognit.*, vol. 42, no. 6, pp. 1080–1092, 2009.

[4] M. M. Dundar, S. Badve, G. Bilgin, V. Raykar, R. Jain, O. Sertel, and M. N. Gurcan, "Computerized classification of intraductal breast lesions using histopathological images," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 7, pp. 1977–1984, 2011.

[5] P.-W. Huang and C.-H. Lee, "Automatic classification for pathological prostate images based on fractal analysis," *IEEE Trans. Med. Imag.*, vol. 28, no. 7, pp. 1037–1050, 2009.

[6] U. Srinivas, H. S. Mousavi, V. Monga, A. Hattel, and B. Jayarao, "Simultaneous sparsity model for histopathological image representation and classification," *IEEE Trans. Med. Imag.*, vol. 33, no. 5, pp. 1163–1179, 2014.

[7] S. Doyle, S. Agner, A. Madabhushi, M. Feldman, and J. Tomaszewski, "Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features," in *Int. Symp. Biomed. Imag.: From Nano to Macro*. IEEE, 2008, pp. 496–499.

[8] V.-T. Ta, O. Lézoray, A. Elmoataz, and S. Schüpp, "Graph-based tools for microscopic cellular image segmentation," *Pattern Recognit.*, vol. 42, no. 6, pp. 1113–1125, 2009.

[9] L. E. Boucheron, "Object- and spatial-level quantitative analysis of multispectral histopathology images for detection and characterization of cancer," Ph.D. dissertation, University of California, Santa Barbara, Mar 2008. [Online]. Available: https://vision.ece.ucsb.edu/abstract/518

[10] F. Zana and J.-C. Klein, "Segmentation of vessel-like patterns using mathematical morphology and curvature evaluation," *IEEE Trans. Image Process.*, vol. 10, no. 7, pp. 1010–1019, 2001.

[11] T. H. Vu, H. S. Mousavi, V. Monga, G. Rao, and U. A. Rao, "Histopathological image classification using discriminative feature-oriented dictionary learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 3, pp. 738–751, 2016.

[12] S. Ray and M. Craven, "Supervised versus multiple instance learning: An empirical comparison," in *Proc. Int. Conf. Mach. Learn.*, 2005, pp. 697–704.

[13] H. Hajimirsadeghi and G. Mori, "Multiple instance real boosting with aggregation functions," in *Int. Conf. Pattern Recognit.* IEEE, 2012, pp. 2706–2710.

[14] X. Wang, B. Wang, X. Bai, W. Liu, and Z. Tu, "Max-margin multiple-instance dictionary learning," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 846–854.

[15] D. Li, J. Wang, X. Zhao, Y. Liu, and D. Wang, "Multiple kernel-based multi-instance learning algorithm for image classification," *J. Vis. Commun. and Image Represent.*, vol. 25, no. 5, pp. 1112–1117, 2014.

[16] J. Wu, Y. Zhao, J.-Y. Zhu, S. Luo, and Z. Tu, "MILCut: A sweeping line multiple instance learning paradigm for interactive image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 256–263.

[17] B. Settles, M. Craven, and S. Ray, "Multiple-instance active learning," in *Adv. Neural Inf. Process. Syst.*, 2008, pp. 1289–1296.

[18] E. Alpaydın, V. Cheplygina, M. Loog, and D. M. Tax, "Single-vs. multiple-instance classification," *Pattern Recognit.*, vol. 48, no. 9, pp. 2831–2838, 2015.

[19] B. Babenko, M.-H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1619–1632, 2011.

[20] H. Kück and N. de Freitas, "Learning about individuals from group statistics," in *Conf. Uncertain. Artif. Intell.* AUAI Press, 2005, p. 332–339.

[21] K. Zhang and H. Song, "Real-time visual tracking via online weighted multiple instance learning," *Pattern Recognit.*, vol. 46, no. 1, pp. 397–411, 2013.

[22] J. Gibson, A. Katsamanis, F. Romero, B. Xiao, P. Georgiou, and S. Narayanan, "Multiple instance learning for behavioral coding," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 81–94, 2017.

[23] G. Quellec, M. Lamard, M. Cozic, G. Coatrieux, and G. Cazuguel, "Multiple-instance learning for anomaly detection in digital mammography," *IEEE Trans. Med. Imag.*, vol. 35, no. 7, pp. 1604–1614, 2016.

[24] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, 2017.

[25] M. M. Dundar, S. Badve, V. C. Raykar, R. K. Jain, O. Sertel, and M. N. Gurcan, "A multiple instance learning approach toward optimal classification of pathology slides," in *Int. Conf. Pattern Recognit.* IEEE, 2010, pp. 2732–2735.

[26] Y. Xu, J.-Y. Zhu, E. Chang, and Z. Tu, "Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* IEEE, 2012, pp. 964–971.

[27] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Adv. Neural Inf. Process. Syst.*, 2003, pp. 577–584.

[28] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, "Multi-instance learning by treating instances as non-iid samples," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 1249–1256.

[29] M. Kandemir, A. Feuchtinger, A. Walch, and F. A. Hamprecht, "Digital Pathology: Multiple instance learning can detect Barrett's cancer," in *Int. Symp. Biomed. Imag.* IEEE, 2014, pp. 1348–1351.

[30] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Adv. Neural Inf. Process. Syst.*, 1998, pp. 570–576.

[31] Q. Zhang and S. A. Goldman, "EM-DD: An improved multiple-instance learning technique," in *Adv. Neural Inf. Process. Syst.*, 2002, pp. 1073–1080.

[32] Y. Chen and J. Z. Wang, "Image categorization by learning and reasoning with regions," *J. Mach. Learn. Res.*, vol. 5, no. Aug, pp. 913–939, 2004.

[33] C. Zhang, J. C. Platt, and P. A. Viola, "Multiple instance boosting for object detection," in *Adv. Neural Inf. Process. Syst.*, 2006, pp. 1417–1424.

[34] 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 *Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*. Springer Berlin Germany, 2009, pp. 15–30.

[35] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, 2010.

[36] T. Deselaers and V. Ferrari, "A conditional random field for multiple-instance learning," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 287–294.

[37] J. Foulds and E. Frank, "A review of multi-instance learning assumptions," *Knowl. Eng. Rev.*, vol. 25, no. 1, pp. 1–25, 2010.

[38] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognit.*, vol. 77, pp. 329–353, 2018.

[39] A. T. Pham, R. Raich, and X. Z. Fern, "Dynamic programming for instance annotation in multi-instance multi-label learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2381–2394, 2017.

[40] A. T. Pham, R. Raich, X. Z. Fern, and J. Pérez Arriaga, "Multi-instance multi-label learning in the presence of novel class instances," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2427–2435.

[41] Z. You, R. Raich, X. Z. Fern, and J. Kim, "Weakly supervised dictionary learning," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2527–2541, 2018.

[42] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Stat. Soc.: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[43] M. R. Andersen, O. Winther, and L. K. Hansen, "Bayesian inference for structured spike and slab priors," in *Adv. Neural Inf. Process. Syst.*, 2014, pp. 1745–1753.

[44] N. Weidmann, E. Frank, and B. Pfahringer, "A two-level learning method for generalized multi-instance problems," in *Proc. Eur. Conf. Mach. Learn.* Springer Berlin Germany, 2003, pp. 468–479.

[45] T. Jaakkola, M. Meila, and T. Jebara, "Maximum entropy discrimination," in *Adv. Neural Inf. Process. Syst.*, 1999, pp. 470–476.

[46] P. Hennig and C. J. Schuler, "Entropy search for information-efficient global optimization," *J. Mach. Learn. Res.*, vol. 13, no. Jun, pp. 1809–1837, 2012.

[47] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal estimated sub-gradient solver for SVM," *Math. Program.*, vol. 127, no. 1, pp. 3–30, 2011.

[48] N. Cesa-Bianchi, S. Shalev-Shwartz, and O. Shamir, "Efficient learning with partially observed attributes," *J. Mach. Learn. Res.*, vol. 12, no. Oct, pp. 2857–2878, 2011.

[49] Z. Wang, K. Mülling, M. P. Deisenroth, H. Ben A., D. Vogt, B. Schölkopf, and J. Peters, "Probabilistic movement modeling for intention inference in human-robot interaction," *Int. J. Rob. Res.*, vol. 32, no. 7, pp. 841–858, 2013.

[50] J. Futoma, M. Sendak, B. Cameron, and K. Heller, "Predicting disease progression with a model for multivariate longitudinal clinical data," in *Mach. Learn. Healthcare Conf.*, 2016, pp. 42–54.

[51] J. M. Hernández-Lobato, Y. Li, M. Rowland, D. Hernández-Lobato, T. Bui, and R. Turner, "Black-box α-divergence minimization," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1511–1520.

[52] A. T. Pham, R. Raich, and X. Z. Fern, "Efficient instance annotation in multi-instance learning," in *Proc. IEEE Workshop Stat. Signal Process.* IEEE, 2014, pp. 137–140.

[53] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Adv. Neural Inf. Process. Syst.*, 2008, pp. 1177–1184.

[54] M. Dundar, B. Krishnapuram, R. Rao, and G. M. Fung, "Multiple instance learning for computer aided diagnosis," in *Adv. Neural Inf. Process. Syst.*, 2007, pp. 425–432.

[55] C. Leistner, A. Saffari, and H. Bischof, "MIForests: Multiple-instance learning with randomized trees," in *Proc. Eur. Conf. Comput. Vis.* Springer Berlin Germany, 2010, pp. 29–42.

[56] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge," *Contemporary Oncology*, vol. 19, no. 1A, p. A68, 2015.

[57] A. F. Costa, G. Humpire-Mamani, and A. J. M. Traina, "An efficient algorithm for fractal analysis of textures," in *SIBGRAPI Conference on Graphics, Patterns and Images*. IEEE, 2012, pp. 39–46.

[58] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola, "Multi-instance kernels," in *Proc. Int. Conf. Mach. Learn.*, 2002, pp. 179–186.

[59] Y. Chen, J. Bi, and J. Z. Wang, "MILES: Multiple-instance learning via embedded instance selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1931–1947, 2006.

[60] O. L. Mangasarian and E. W. Wild, "Multiple instance classification via successive linear programming," *J. Optim. Theory Appl.*, vol. 137, no. 3, pp. 555–568, 2008.

[61] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. Int. Joint Conf. Artif. Intell.*, 1995, pp. 1137–1145.