Simplified and Unified Analysis of Various Learning Problems by Reduction to Multiple-Instance Learning

Abstract

In statistical learning, many problem formulations have been proposed so far, such as multi-class learning, complementarily labeled learning, multi-label learning, multi-task learning, which provide theoretical models for various real-world tasks. Although they have been extensively studied, the relationship among them has not been fully investigated. In this work, we focus on a particular problem formulation called Multiple-Instance Learning (MIL), and show that various learning problems including all the problems mentioned above with some of new problems can be reduced to MIL with theoretically guaranteed generalization bounds, where the reductions are established under a new reduction scheme we provide as a by-product. The results imply that the MIL-reduction gives a simplified and unified framework for designing and analyzing algorithms for various learning problems. Moreover, we show that the MIL-reduction framework can be kernelized.

1 INTRODUCTION

In this study, we explore how a large class of learning problems can be reduced to the Multiple-Instance Learning (MIL) problem. This is strongly motivated by the results of [Sabato and Tishby, 2012] and [Suehiro et al., 2020]. Suehiro et al. [2020] showed that some local-feature-based learning problems can be reduced to a MIL problem, which gave us an insight that MIL would have a high capability of representing various learning problems. Indeed, the reduced problem is too specific whereas Sabato and Tishby [2012] proposed a much more general formulation of MIL, and thus we believe that a wider class of learning problems can be reduced to MIL.

We provide a MIL-reduction framework and reveal that various learning problems, such as multi-class learning, complementarily labeled learning, multi-label learning, and multi-task learning, can be reduced to MIL. By the reduction, we immediately derive generalization bounds from [Sabato and Tishby, 2012], as well as learning algorithms. That is, our reduction framework greatly *simplifies* the analyses of generalization bounds as compared with the analyses in the previous works [e.g., Lei et al., 2019, Ishida et al., 2017, Yu et al., 2014, Pontil and Maurer, 2013]. Some of the obtained generalization bounds are competitive or incomparable to the existing results. In particular, for multi-label learning, we derive an improved generalization bound, and for complementarily labeled learning, we derive a novel learning algorithm, which is the first polynomial-time algorithm in a certain setting. Moreover, we propose three new learning problems, *multi-label learning with perfectionistic loss*, *top-1 ranking learning* and *top-1 ranking learning with negative feedback*, and we demonstrate that they can be reduced to MIL as well. The results imply that our MIL-reduction gives a *unified framework* for designing and analyzing algorithms for various learning problems.

To provide the MIL-reduction framework, we propose a general reduction scheme among learning problems. Our scheme has two remarkable features as described below. First, our reduction transforms every instance-label pair (x, y) in the given sample of the original learning problem to an instance-label pair (x', y') to form a sample of the reduced learning problem. In contrast, standard reduction schemes employ an instance transformation and an label transformation separately, to construct x' from x and y' from y, respectively. Therefore, our scheme enables us to design reduction algorithms among a wider class of learning problems, e.g., learning-to-rank to classification, and supervised learning to weakly supervised learning. Second, our reduction scheme ensures that the Empirical Risk Minimization (ERM) of the reduced problem implies the ERM of the original one, while the empirical Rademacher complexity of the hypothesis (composed with loss function) classes are preserved through the reduction. This means that we can

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

employ an existing ERM algorithm for the reduced problem to obtain an ERM algorithm for the original problem with a theoretical guaranteed generalization bound, which is immediately derived from a known generalization bound for the reduced problem. We also show that the MIL-reduction framework can be kernelized.

The main contributions are summarized as follows:

- We propose a general reduction scheme based on the ERM, which allows us to derive a generalization risk bound of the original problem immediately.
- We demonstrate that several learning problems, from traditional to new problems, can be reduced to MIL. The results imply that our MIL-reduction gives a simplified and unified framework for the analyses for various learning problems.
- We obtain novel theoretical results for some learning problems.
- We show that the MIL-reduction framework can be kernelized.

Several proofs are shown in supplementary materials.

2 PRELIMINARIES

For an integer u, $[u]$ denotes the set $\{1, \ldots, u\}$. $I(\mathrm{e})$ denotes the indicator function of the event e, that is, $I(\mathrm{e}) = 1$ if e is true and $I(\mathrm{e}) = 0$ otherwise.

A learning problem is represented by a pair $(\mathcal{H}, \ell)$ of a hypothesis class $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Y}\}$ and a loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \to \mathbb{R}$ for some input space $\mathcal{X}$ and output space $\mathcal{Y}$. A learner receives a sample $S = ((x_1, y_1), \ldots, (x_n, y_n))$ where each input-output pair (x_i, y_i) is drawn i.i.d. according to an unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. The goal of the learner is to find, with high probability, a hypothesis $h \in \mathcal{H}$ so that the generalization risk $R_{\mathcal{D}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \ell(x, y, h)$ is small.

For a learning problem $(\mathcal{H}, \ell)$, we define a class of loss functions as $\widehat{\mathcal{H}} = \{(x, y) \mapsto \ell(x, y, h) \mid h \in \mathcal{H}\}$ when the underlying loss function ℓ is clear from the context. We give the definition of the empirical Rademacher complexity, which is used to bound the generalization risk.

Definition 1 (Empirical Rademacher complexity [Bartlett and Mendelson, 2003])**.** *Given a sample $S = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, the empirical Rademacher complexity $\mathfrak{R}_S(\widehat{\mathcal{H}})$ of a class $\widehat{\mathcal{H}}$ w.r.t. S is defined as $\mathfrak{R}_S(\widehat{\mathcal{H}}) = \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\widehat{\mathcal{H}}} \sum_{i=1}^n \sigma_i g(x_i, y_i)\right]$, where $\boldsymbol{\sigma} \in \{-1, 1\}^n$ and each σ_i is an independent uniform random variable in $\{-1, 1\}$.*

Generalization risk bound [Mohri et al., 2018] Let $(\mathcal{H}, \ell)$ be a learning problem and S be a sample of size n drawn according to a distribution $\mathcal{D}$. Then, it holds with probability at least $1 - \delta$ that for all $h \in \mathcal{H}$,

$$R_{\mathcal{D}}(h) \leq \widehat{R}_S(h) + 2\mathfrak{R}_S(\widehat{\mathcal{H}}) + 3\sqrt{\log(2/\delta)/2n},$$

where $\widehat{R}_S(h) = \frac{1}{n}\sum_{i=1}^n \ell(x_i, y_i, h)$ denotes the empirical risk of h for sample S.

3 REDUCTION SCHEME FOR ERM

We introduce a general reduction scheme for empirical risk minimization and provide useful theoretical results.

Definition 2 (ERM-reduction)**.** *A learning problem $(\mathcal{H}, \ell)$ over input-output space $\mathcal{X} \times \mathcal{Y}$ is* ERM-reducible *to another learning problem $(\mathcal{H}', \ell')$ over input-output space $\mathcal{X}' \times \mathcal{Y}'$ if there exist polynomial-time computable functions $\alpha : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}' \times \mathcal{Y}'$ and $\beta : \mathcal{H}' \to \mathcal{H}$ such that for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and for any $h' \in \mathcal{H}'$,*

$$\ell(x, y, h) = \ell'(x', y', h'),$$

where $(x', y') = \alpha(x, y)$ and $h = \beta(h')$.

Here we show the remarkable relationship between the original problem and the reduced problem.

Proposition 1. *Suppose that $(\mathcal{H}, \ell)$ is ERM-reducible to $(\mathcal{H}', \ell')$ with transformations α and β. For any sample $S = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, the following holds:*

(i) (In)equality of the ERMs:

$$\min_{h\in\mathcal{H}} \widehat{R}_S(h) \leq \min_{h\in\mathcal{H}_\beta} \widehat{R}_S(h) = \min_{h'\in\mathcal{H}'} \widehat{R}_{S'}(h'),$$

where $\mathcal{H}_\beta = \{\beta(h') \mid h' \in \mathcal{H}'\}$ and $S' = ((x'_1, y'_1), \ldots, (x'_n, y'_n))$ with $(x'_i, y'_i) = \alpha(x_i, y_i)$ for $i \in [n]$.

(ii) Empirical Rademacher complexity preserving:

$$\mathfrak{R}_S(\widehat{\mathcal{H}}_\beta) = \mathfrak{R}_{S'}(\widehat{\mathcal{H}}').$$

We can design a reduction scheme in a straightforward way as follows. When given a sample S of the original problem, we construct S' of the reduced problem by α and obtain h' by solving the ERM of the reduced problem. Then, we obtain the final hypothesis h by β.

We derive the following generalization risk bound using the propositions on the empirical Rademacher complexity.

Corollary 2. *Let $S = ((x_1, y_1), \ldots, (x_n, y_n))$ be a sample i.i.d. drawn according to unknown distribution $\mathcal{D}$ in an original problem $(\mathcal{H}, \ell)$. If $(\mathcal{H}, \ell)$ is ERM-reducible to $(\mathcal{H}', \ell')$, for $S' = (\alpha(x_1, y_1), \ldots, \alpha(x_n, y_n))$ and $h = \beta(h')$, the following generalization risk bound holds with a probability at least $1 - \delta$ for all $h \in \mathcal{H}_\beta$:*

$$R_{\mathcal{D}}(h) \leq \widehat{R}_{S'}(h') + 2\mathfrak{R}_{S'}(\widehat{\mathcal{H}}') + 3\sqrt{\log(2/\delta)/2n}.$$

That is, we can guarantee the generalization bound of the original problem because of the preservation of the empirical Rademacher complexity.

4 MIL-REDUCTION FRAMEWORK

This section is the highlight of this paper. We define the ERM-reducibility to MIL and show the reducible condition. Moreover, we show that some theoretical analyses can be simplified. We use some symbols with prime (e.g., $\mathcal{X}'$) to indicate that the MIL is the reduced problem.

4.1 PROBLEM FORMULATION OF MIL

Let $\mathcal{Z} \subseteq \mathbb{R}^{d'}$ be the instance space. $\mathcal{X}' \subseteq 2^{\mathcal{Z}}$ is an input space and a *bag* $x' \in \mathcal{X}'$ is a finite set of instances chosen from $\mathcal{Z}$. Let $\mathcal{Y}' = \{-1, 1\}$ be an output space. Following the formulation by [Sabato and Tishby, 2012], we define, for the rest of the paper, a MIL problem as a pair $(\mathcal{H}', \ell')$ of a hypothesis class $\mathcal{H}'$ and a loss function ℓ' of the form:

$$\mathcal{H}' = \{h' : x' \mapsto \Psi_p(\{f_2(g(z)) \mid z \in x'\}) \mid g \in \mathcal{G}\}, \quad (1)$$

$$\ell' : (x', y', h') \mapsto f_1(y'h'(x')), \quad (2)$$

where $\mathcal{G} \subseteq \{g : \mathcal{Z} \to \mathbb{R}\}$, $f_1 : \mathbb{R} \to [0, 1]$ is an a-Lipschitz function, $f_2 : \mathbb{R} \to [-1, 1]$ is a b-Lipschitz function, and $\Psi_p : 2^{[-1,1]} \to [-1, 1]$ is a p-norm like function, which is defined for any $p \in [1, \infty)$ as

$$\Psi_p(V) = \left(\frac{1}{m} \sum_{i=1}^{m} (v_i + 1)^p \right)^{1/p} - 1$$

for every finite set $V = \{v_1, v_2, \ldots, v_m\} \subseteq [-1, 1]$. We define Ψ_∞ as $\lim_{p\to\infty} \Psi_p$. Note that Ψ_p is 1-Lipschitz for any p [see, Sabato and Tishby, 2012]. Ψ_p is a user-defined function and behaves as an aggregation of some bag information. Typical Ψ_p are the $\max$ operator ($p = \infty$) and average ($p = 1$).

The only difference in the hypothesis of [Sabato and Tishby, 2012] is f_2. f_2 appears redundant (because $f_2 \circ g$ can be replaced by a single function) but plays an important role in the reduction (the examples are shown in Section 5).

4.2 MIL-REDUCIBILITY

Now we give the definition of ERM-reducibility in a straightforward way.

Definition 3 (MIL-reducibility)**.** *A learning problem $(\mathcal{H}, \ell)$ is said to be MIL-reducible if there exists a MIL problem $(\mathcal{H}', \ell')$ such that $(\mathcal{H}, \ell)$ is ERM-reducible to $(\mathcal{H}', \ell')$.*

4.3 RADEMACHER COMPLEXITY BOUND

We show the empirical Rademacher complexity bound for the MIL-reducible problems using our reduction scheme. As aforementioned, the main advantage of our reduction scheme is to allow us to apply the empirical Rademacher complexity bound of the reduced problem to the original problems. In this paper, we utilize the bound provided by Sabato and Tishby [2012].

Theorem 3 (An application of [Sabato and Tishby, 2012])**.** *Let $(\mathcal{H}', \ell')$ be a MIL problem defined in Eq.*(1) *and* (2)*. Let $S' = ((x'_1, y'_1), \ldots, (x'_n, y'_n))$ be a sample with average bag size $r_{S'}$. Let $\widehat{\mathcal{G}} = \{f_2 \circ g \mid g \in \mathcal{G}\}$. If there exist $C, \rho \geq 0$ such that for all sufficiently large n,*

$$\mathfrak{R}_{S'}(\widehat{\mathcal{G}}) \leq \frac{C \ln^{\rho}(n)}{\sqrt{n}},$$

then

$$\mathfrak{R}_{S'}(\widehat{\mathcal{H}}') = O\left(\frac{\log\left(a^2 n^2 r_{S'}\right) \left(\frac{aC}{\rho+1} \ln^{\rho+1}(a^2 n)\right)}{\sqrt{n}} \right),$$

where $\widehat{\mathcal{H}}' = \{\hat{h}' : x' \mapsto f_1(y'h'(x')) \mid h' \in \mathcal{H}'\}$.

As mentioned in [Sabato and Tishby, 2012], we obtain the following bound when $\mathcal{G}$ is a set of linear functions.

Corollary 4. *Let $\mathcal{G} = \{g : z \mapsto \langle w', z\rangle \mid w' \in \mathbb{R}^{d'}, \|w'\| \leq C_1\}$ and assume that $\|z\| \leq C_2$. Then, the following bound holds:*

$$\mathfrak{R}_{S'}(\widehat{\mathcal{H}}) = O\left(\frac{\log\left(a^2 n^2 r_{S'}\right) \left(abC_1C_2 \ln(a^2 n)\right)}{\sqrt{n}} \right).$$

The above bound is easily derived from the result of $\mathfrak{R}_{S'}$ [see the proof of Theorem 20 of Sabato and Tishby, 2012]) and $\mathfrak{R}_{S'}(\widehat{\mathcal{G}}) \leq b\mathfrak{R}_{S'}(\mathcal{G}) \leq {}^{bC_1C_2}/\sqrt{n} = {}^{bC_1C_2 \ln^0(n)}/\sqrt{n}$ [see, e.g., Theorem 5.8 and 5.10 of Mohri et al., 2018].

Using Theorem 3 and Corollary 2, we obtain a generalization risk bound for MIL-reducible problems.

4.4 LEARNING ALGORITHM

For the reduced MIL problems that satisfy some conditions, we can immediately design a learning algorithm according to the condition. Suppose that $\mathcal{G}$ is a set of linear functions:

$$\mathcal{G} = \{g : z \mapsto \langle w', z\rangle \mid w' \in \mathbb{R}^{d'}, \|w'\| \leq C_1\}. \quad (3)$$

Let $S' = ((x'_1, y'_1), \ldots, (x'_n, y'_n))$. The ERM of MIL is formulated as follows:

$$\min_{\|w'\| \leq C_1} \lambda \|w'\|^2 + \sum_{i=1}^{n} f_1\left(y'_i \Psi_p\left(\{f_2\left(\langle w', z\rangle \mid z \in x'_i\right)\}\right)\right). \quad (4)$$

For the optimization problem (4), we show that the following propositions hold.

Proposition 5. *If $y'_i = -1$ for any $i \in [n]$ for sample S', $f_1(c)$ is convex and nonincreasing* [1]*, and f_2 is a nondecreasing convex function, and $\mathcal{G}$ is given as Eq.*(3)*, then the ERM of $(\mathcal{H}', \ell')$ is a convex programming problem.*

Proposition 6. *If f_1 is a nonincreasing convex* [1] *and $f_1(c)$ is a homogeneous function of degree 1 for $c \in [-1, 1]$*[2]*, and f_2 is a nondecreasing convex function, and $\mathcal{G}$ is given as Eq.*(3)*, then ERM of $(\mathcal{H}', \ell')$ is a DC programming problem.*

Generally, it is difficult to find a global minimum for a DC programming problem; however, it is known that we can find a solution with ϵ-approximation of local optima [see, e.g., Le Thi and Dinh, 2018]. We introduce a standard DC algorithm to solve (4) in Algorithm 1 in Sec. D.

The propositions indicate that, if $(\mathcal{H}, \ell)$ is MIL-reducible to $(\mathcal{H}', \ell')$ and satisfies either of the above conditions, then the solution $h \in \mathcal{H}_\beta$ in the original problem can be obtained by a unified learning algorithm.

5 MIL-REDUCIBLE EXAMPLES

In this section, we demonstrate that various learning problems can be reduced to MIL by the proposed reduction framework. The results imply that our MIL-reduction gives a unified framework for designing and analyzing learning algorithms for various learning problems[3].

5.1 THE EXISTING PROBLEMS

5.1.1 Multi-class learning problem

Problem setting: Let $\mathcal{X} \subseteq \mathbb{R}^d$ be an instance space, and $\mathcal{Y} = [k]$ be an output space. The learner receives the set of labeled instances $S = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$, where each instance is drawn i.i.d. according to some unknown distribution $\mathcal{D}$. The learner predicts the label of x using the hypothesis $h \in \mathcal{H} = \{x \mapsto \arg\max_{y \in [k]} \langle w_y, x \rangle \mid \forall j \in [k], w_j \in \mathbb{R}^d\}$. Let $\ell : (x, y, h) \mapsto \Gamma(\langle w_y, x \rangle - \max_{j \in \mathcal{Y} \setminus y} \langle w_j, x \rangle)$ be a loss function, where $\Gamma : \mathbb{R} \to [0, 1]$ is a convex, nonincreasing and a-Lipschitz function. The generalization risk and empirical risk of h are defined as:

$$R_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}} \ell(x, y, h), \widehat{R}_S(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i, h).$$

[1]More precisely, the extended-value extension f_1 also must be nonincreasing (See details in [Boyd and Vandenberghe, 2004]).

[2]For example, hinge-loss function $f(c) : c \mapsto \{0, 1 - c\}$ satisfies this condition.

[3]The reduction of Multi-task learning and top-1 ranking learning negative feedback are shown in Sec.G and J owing to space limitations.

We obtain the following by using MIL-reduction framework:

Theorem 7. *Multi-class learning problem is MIL-reducible.*

Proof. For any (x, y), we define

$$\eta_{(x,y)} = (\mathbf{0}, \ldots, \mathbf{0}, \underbrace{x}_{y-\text{th block}}, \mathbf{0}, \ldots, \mathbf{0}),$$

where $\mathbf{0}$ is a d-dimensional vector, the elements of which are all 0. On the MIL-reduction framework, suppose that $p = \infty$; $f_1(c) = \Gamma(2cC_1C_2)$, $f_2(c) = c/2C_1C_2$ (shifting function to $[-1, +1]$); $\alpha(x, y) = (x'_{(x,y)}, y')$ where $x'_{(x,y)} = \{\eta_{(x,j)} - \eta_{(x,y)} \mid \forall j \in \mathcal{Y} \setminus y\}$; $y' = -1$; for any $z \in \mathbb{R}^{kd}$, $\mathcal{G} = \{g : z \mapsto \langle (w'_1, \ldots, w'_k), z \rangle \mid w'_j \in \mathbb{R}^d, \forall j \in [k], \|W'\| \leq C_1\}$ where $W' = (w'_1, \ldots, w'_k)$ and $\|W'\| = \sqrt{\sum_{j=1}^{k} \|w'_j\|^2}$; $\beta(h') : x \mapsto \arg\max_{j \in [k]} \langle (w'_1, \ldots w'_k), x \rangle$. Then, for any (x, y) and $h \in \mathcal{H}$,

$$\begin{aligned}
\ell'(x', y', h') =& f_1\left(y' \Psi_p\left(\{f_2\left(g(z) \mid z \in x'_{(x,y)}\right\}\right)\right) \\
=& \Gamma\left(-\Psi_\infty\left(\{g(z) \mid z \in x'_{(x,y)}\right\}\right)\right) \\
=& \Gamma\left(-\left(\max_{j \in \mathcal{Y} \setminus y} \left(\langle w_j, x \rangle - \langle w_y, x \rangle\right)\right)\right) \\
=& \ell(x, y, h)
\end{aligned}$$

□

The empirical Rademacher complexity is immediately derived as follows by observing the reduction process.

Corollary 8. *We assume that $\|x_i\| \leq C_2$ for any $i \in [n]$. In the reduced MIL problem from multi-class learning problem, the empirical Rademacher complexity of $\widehat{\mathcal{H}}'$ is given as:*

$$\mathfrak{R}_{S'}(\widehat{\mathcal{H}}') = O\left(\frac{\log\left(\hat{a}^2 2n^2(k-1)\right)\left(2\hat{a}\ln(\hat{a}^2 n)\right)}{\sqrt{n}}\right),$$

where $\hat{a} = 2aC_1C_2$ and we assume $\|w'\| \leq C_1$ in the reduced MIL.

We used the fact that the bag size is $(k-1)$ for all x'_i (i.e., $r_{S'} = k - 1$) and $\mathfrak{R}(\widehat{\mathcal{G}}) \leq 2/\sqrt{n}$ by setting $f_2(c) = c/C_1C_2$. Using Corollary 2, we can obtain the generalization risk bound for the multi-class learning.

The learning algorithm is obtained by the following result.

Corollary 9. *The reduced ERM of the MIL from multi-class learning is a convex programming problem.*

The proof of Theorem 7 shows that f_2 is nondecreasing convex and $y'_i = -1$ for all $i \in [n]$. Therefore, by Proposition 5, if we consider Γ that is a nonicreasing and convex function, the ERM of the reduced MIL problem is a convex programming problem and solved in polynomial time.

5.1.2 Complementarily labeled learning problem

Complementarily labeled learning was proposed by Ishida et al. [2017]. In this problem, some training instances are complementarily labeled (e.g., instance x_i is NOT y_i). We essentially follow the problem setting and some assumptions provided by Ishida et al. [2017].

Problem setting: Let $\mathcal{X} \subseteq \mathbb{R}^d$ be an instance space and $\mathcal{Y} = [k]$ be an output space. Let $\mathcal{D}$ be an unknown distribution over $\mathcal{X} \times \mathcal{Y}$. We assume that the learner receives a sample S drawn i.i.d. according to the distribution $\mathcal{D}'$ which provides the true label with unknown probability θ and the complementary label with unknown probability $1-\theta$. Moreover, we assume that the complementary label is chosen with a uniform probability (i.e., all complementary labels are equally chosen with the probability $1/(k-1)$). [4] More formally, we assume that the sample is given as $S = ((x_1, y_1, \gamma_1) \ldots, (x_n, y_n, \gamma_n))$ which is drawn i.i.d. according to the distribution $\mathcal{D}'$ over $\mathcal{D} \times \{\mathrm{False}, \mathrm{True}\}$, where $\gamma_i = \mathrm{True}$ means that y_i is the true label and $\gamma_i = \mathrm{False}$ means that y_i is the complementary label (i.e., it indicates that x_i is NOT y_i). For any $(x, y) \sim \mathcal{D}$, $\mathcal{D}'(x, y, \mathrm{True}) = \theta$ and $\mathcal{D}'(x, \bar{y}, \mathrm{False}) = \frac{1-\theta}{k-1}$ for any $\bar{y} \neq y$ (i.e., the complementary label is chosen with a uniform probability). The other basic settings are the same as those for the aforementioned multi-class learning. The learner predicts the label of x using the hypothesis $h \in \mathcal{H} = \{x \mapsto \arg\max_{y \in [k]} \langle w_y, x \rangle \mid \forall j \in [k], w_j \in \mathbb{R}^d\}$. The final goal of the learner is to find $h \in \mathcal{H}$ with a small multi-class classification risk:

$$R_{\mathcal{D}}^{\mathrm{MC}}(h) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} I\left(y \neq h(x)\right).$$

However, it is difficult to minimize the empirical multi-class classification risk directly using the complementarily labeled data. Therefore, we consider the following risk[5].

$$R_{\mathcal{D}'}^{\mathrm{LC}}(h) = \mathop{\mathbb{E}}_{(x,y,\gamma)\sim\mathcal{D}'} \left[I\left(\gamma = (y \neq h(x))\right)\right].$$

This risk implies that when $\gamma = \mathrm{True}$, the learner does not incur a risk if it predicts the true label. When $\gamma = \mathrm{False}$, the learner does not incur a risk if it predicts an assigned nontrue label. Thus, the risk measure is defined using the pair $(y, \gamma) \in (\mathcal{Y} \times \{\mathrm{False}, \mathrm{True}\})$. We can show that achieving a small $R_{\mathcal{D}'}^{\mathrm{LC}}(h)$ is consistent with achieving small $R_{\mathcal{D}}^{\mathrm{MC}}(h)$ as follows:

Lemma 1. *For any $h \in \mathcal{H}$, $R_{\mathcal{D}}^{\mathrm{MC}}(h) = \frac{k-1}{\theta(k-2)+1} R_{\mathcal{D}'}^{\mathrm{LC}}(h)$ holds.*

Thus, minimizing $R_{\mathcal{D}'}^{\mathrm{LC}}(h)$ is a reasonable way to achieve a high multi-class classification accuracy.

[4] This assumption was proposed by Ishida et al. [2017] as a reasonable scenario in some practical tasks (e.g., crowdsourcing).

[5] Ishida et al. [2017] used a different surrogate risk. However, they and we have a common goal: to minimize $R_{\mathcal{D}}^{\mathrm{MC}}(h)$

Generally, there is no loss function $\ell((x, \gamma), y, h)$ which is a convex upper bound on the zero-one loss $I\left(\gamma = (y \neq h(x))\right)$ over the domain w. This is because if $I(\gamma = \mathrm{True}) = 1$ then $\max$ is convex w.r.t. w; however, if $I(\gamma = \mathrm{True}) = -1$ then $-\max = \min$ is concave w.r.t. w. Therefore, we consider the convex upper bounded loss only on the risk for complementarily labeled data (i.e., the concave risk for the normally labeled data) using $\Gamma : \mathbb{R} \to [0, 1]$ as $\Gamma\left(\max_{j \in \mathcal{Y} \setminus y} \langle (w_j - w_y), x \rangle\right)$. We then define the nonconvex risk $\ell(x, (\gamma, y), h) = \Gamma\left(I(\gamma = \mathrm{True}) \times \left(\max_{j \in \mathcal{Y} \setminus y} \langle (w_j - w_y), x \rangle\right)\right)$. The empirical risk is formulated as:

$$\widehat{R}_S^{\mathrm{LC}}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(x_i, (\gamma_i, y_i), h\right).$$

The following is obtained by MIL-reduction framework.

Theorem 10. *Complementarily labeled learning is MIL-reducible.*

The difference from the reduction in multi-class learning is that only y' takes $\{-1, 1\}$. y' behaves as a *switch* that turns the loss of complementarily or normally labeled data.

The empirical Rademacher complexity is bounded as:

Corollary 11. *We assume that $\|x_i\| \leq C_2$ for any $i \in [n]$. In the reduced MIL problem from complementarily labeled learning, the empirical Rademacher complexity of $\widehat{\mathcal{H}}'$ is given by:*

$$\mathfrak{R}_{S'}(\widehat{\mathcal{H}}') = O\left(\frac{\log\left(\hat{a}^2 n^2 (k-1)\right)\left(2\hat{a}\ln(\hat{a}^2 n)\right)}{\sqrt{n}}\right),$$

where $\hat{a} = 2aC_1C_2$ and we assume $\|w'\| \leq C_1$ in the reduced MIL problem.

We use the same argument as in Corollary 8. Using Corollary 2 and Lemma 1, we obtain the generalization bound for the complementarily labeled learning.

The learning algorithm is derived by the following result:

Corollary 12. *The reduced ERM of the MIL from complementarily labeled learning is a DC programming problem. If the sample contains only complementarily labeled data, the learning problem is a convex programming problem.*

Generally, $y' \in \{-1, 1\}$ in complementarily labeled learning. Using the proof of Theorem 10 and by Proposition 6, if we consider $\Gamma(c)$ which is a nondecreasing and homogeneous function of degree 1 for $c \in [-1, 1]$ such as hinge-loss function, we can solve the problem by DC algorithm as shown in Algorithm 1. Note that, if the sample contains only complementarily labeled data (i.e., $\forall i \in [n]$, $y_i = -1$), it becomes a convex programming problem.

5.1.3 Multi-label learning problem

Problem setting Let $\mathcal{X} \subseteq \mathbb{R}^d$ be an instance space and $\mathcal{Y} \in \{-1, 1\}^k$ be an output space, and $\mathcal{D}$ be an unknown distribution over $\mathcal{X}$. Unlike the standard multi-class learning setting introduced in Section 5.1.1, each instance may have multiple labels (e.g., in text-categorization tasks, some texts have multiple topics such as IT and business). y^j denotes the j-th element of y_i. The learner receives a labeled sample $S = (x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$ which is drawn i.i.d. according to the distribution $\mathcal{D}$. The learner predicts whether x belongs to class $j \in [k]$ or not using the hypothesis $h \in \mathcal{H} = \{(x, j) \mapsto \mathrm{sign}(\langle w_j, x\rangle) \mid \forall w_j \in \mathbb{R}^d\}$. Let $\ell : (x, y, h) \mapsto \frac{1}{k}\sum_{j=1}^{k} \Gamma(-y^j \langle w_j, x\rangle)$ where $\Gamma : \mathbb{R} \to [0, 1]$ is a convex, nondecreasing and b-Lipschitz function[6]. The generalization and empirical risk of h are defined as:

$$R_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [\ell(x,y,h)], \widehat{R}_S(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i, h).$$

Reduction to MIL

Theorem 13. *Multi-label learning is MIL-reducible.*

Proof. On the MIL-reduction framework, suppose that $p = 1$; $f_1 : f_1(a) = -a$ for $a \in \mathbb{R}$; f_2 is Γ; $\alpha(x, y) = (x'_{(x,y)}, y')$ where $x'_{(x,y)} = \{(-y^1 x, 1), \ldots, (-y^k x, k)\}$; $y' = -1$; $\mathcal{G} = \{g : (z, j) \mapsto \langle w'_j, z\rangle \mid w'_j \in \mathbb{R}^d, \forall j \in [k], \|W'\| \leq C_1\}$ where $W' = (w'_1, \ldots, w'_k)$; $\beta(h') : (x, j) \mapsto \mathrm{sign}(\langle w'_j, x\rangle)$. For any (x, y) and $h \in \mathcal{H}$, we have that

$$\begin{aligned}
\ell'(x', y', h') =& f_1\left(y'\Psi_p\left(\left\{f_2\left(g(z)\right) \mid z \in x'_{(x,y)}\right\}\right)\right) \\
=& \frac{1}{|x'_{(x,y)}|} \sum_{(y^j x, j) \in x'_{(x,y)}} \Gamma\left(-\langle w_j, y^j x\rangle\right) \\
=& \ell(x, y, h)
\end{aligned}$$

□

The empirical Rademacher complexity is bounded as:

Corollary 14. *We assume that $\|x_i\| \leq C_2$ for any $i \in [n]$. In the reduced MIL problem, the empirical Rademacher complexity of $\widehat{\mathcal{H}}'$ is given as follows:*

$$\mathfrak{R}_{S'}(\widehat{\mathcal{H}}') = O\left(\frac{\log\left(2n^2 k\right)\left(bC_1C_2\ln(n)\right)}{\sqrt{n}}\right),$$

where $\|w'\| \leq C_1$ in the reduced MIL.

[6]Note that we use the negative score $-y^j \langle w_j, x\rangle$ to employ a nondecreasing Γ.

We used the fact that the size of each bag is k. Using Corollary 2, we obtain the generalization risk bound for the multi-label learning.

The learning algorithm is obtained by the following result.

Corollary 15. *The reduced ERM of the MIL from multi-label learning is a convex programming problem.*

The proof of Theorem 13 shows that, f_1 is nonincreasing and convex, and $y'_i = -1$ for all $i \in [n]$. Therefore, by Proposition 5, if we consider Γ that is nondecreasing and convex, the reduced problem is a convex programming problem and it is solved in polynomial time.

5.2 APPLICATION TO THE NEW PROBLEMS

5.2.1 Multi-label learning with perfectionistic loss

Problem setting: In a standard multi-label learning (see Sec.5.1.3), we consider the average prediction error (loss) with the classes. On the other hand, we consider a *perfectionistic* error in multi-label learning problem. More formally, we consider the following loss in a multi-label learning:

$$\ell : (x, y, h) \mapsto \max_{j\in[k]} \Gamma(-y^j \langle w_j, x\rangle),$$

where $\Gamma : \mathbb{R} \to [0, 1]$ is a convex, nondecreasing and b-Lipschitz function. This loss means that the learner incurs the risk unless the learner perfectly predict the correct labels. The generalization and empirical risks of h are given as $R_{\mathcal{D}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(x, y, h)]$, $\widehat{R}_S(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i, h)$, respectively.

Using MIL-reduction framework, we obtain the following:

Theorem 16. *Multi-label learning with perfectionistic loss is MIL-reducible.*

This can be derived by the same argument with multi-label learning except for $p = \infty$ (see Sec.H).

The empirical Rademacher complexity is bounded as:

Corollary 17. *We assume that $\|x_i\| \leq C_2$ for any $i \in [n]$. In the reduced MIL problem, the empirical Rademacher complexity of $\widehat{\mathcal{H}}'$ is given as follows:*

$$\mathfrak{R}_{S'}(\widehat{\mathcal{H}}') = O\left(\frac{\log\left(2n^2 k\right)\left(bC_1C_2\ln(n)\right)}{\sqrt{n}}\right),$$

where we assume $\|w'\| \leq C_1$.

Interestingly, we can have the same generalization risk bound with the standard multi-label learning.

The learning algorithm is derived by the following result.

Corollary 18. *The reduced ERM of the MIL from multi-label learning with perfectionistic loss is a convex programming problem.*

This is easily obtained by observing the reduction process shown in Sec.H and using Prpoposition 5.

A naive approach for the multi-label learning with perfectionistic loss is to reduce to multi-class learning. That is, we consider all combinations of the multi-label as multi-classes and solve 2^k-class learning problem with high computational cost. However, by the above corollary, multi-label learning with perfectionistic loss can be solved efficiently.

5.2.2 Top-1 ranking learning

Learning to rank is a fundamental problem, and many applications, such as recommendation systems, exist. We consider the following natural scenario in a recommendation problem; a learner has a set that contains several items, and it wishes to recommend an item to a target user from the set.

Problem setting: Let $\mathcal{X} \subseteq \mathbb{R}^d$ be an instance space, and $s : \mathcal{X} \to \mathbb{R}$ be a target scoring function. Set A is a finite set of instances selected from $\mathcal{X}$. The learner receives the sequence of the sets of items and the chosen item $S = (A_1, x_1^*), \ldots, (A_n, x_n^*)$, where each $x_i^* \in A_i$ is the highest-valued item determined by the target function s. k denotes the average size of the item sets in S, that is, $k = \frac{1}{n}\sum_{i=1}^n |A_i|$. Each sample set of items is drawn i.i.d. from $\mathcal{X}$ according to an unknown distribution $\mathcal{D}$ over $2^{\mathcal{X}}$. Assume that the learner predicts the item from the item set using the hypothesis $h \in \mathcal{H} = \{A \mapsto \arg\max_{x\in A}\langle w, x\rangle \mid w \in \mathbb{R}^d\}$.[7] Let $\ell(A, x^*, h)$ is a convex upper bound on the zero-one loss function $I(y \neq \hat{y})$. Equivalently, we consider the zero-one loss $I(\langle w, x^*\rangle - \max_{x\in A\setminus x^*}\langle w, x\rangle \leq 0)$ and its convex upper bounded loss $\ell : (A, x^*, h) \mapsto \Gamma(\langle w, x^*\rangle - \max_{x\in A\setminus x^*}\langle w, x\rangle)$ where $\Gamma : \mathbb{R} \to [0, 1]$ is a convex, nonincreasing and a Lipschitz function. The goal of the learner is to find $h \in \mathcal{H}$ with a small misranking risk w.r.t. the target s. Thus, the generalization and empirical risks are formulated as follows:

$$R_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{A\sim\mathcal{D}}\left[\ell\left(A, x^*, h\right)\right], \widehat{R}_S(h) = \frac{1}{n}\sum_{i=1}^n \ell\left(A, x_i^*, h\right),$$

where $x^* = \arg\max_{x\in A} s(x)$.

We obtain the following by using MIL-reduction framework:

Theorem 19. *Top-1 ranking learning is MIL-reducible.*

The reducible condition is satisfied when we set $\alpha(A, x^*) = (x', y')$ where $x' = \{x - x^* \mid x \in A\setminus x^*\}$ $y'_i = -1$ for all $i \in [n]$. The details of the reduction process is in Sec.I.

The empirical Rademacher complexity bound is as follows:

[7] We consider an $\arg\max$ with a fixed tie-breaking rule.

Corollary 20. *We assume that $\|x\| \leq C_2$ for any $x \in A_i, \forall i \in [n]$. In the reduced MIL problem, the empirical Rademacher complexity of $\widehat{\mathcal{H}}'$ is given as follows:*

$$\mathfrak{R}_{S'}(\widehat{\mathcal{H}}') = O\left(\frac{\log\left(\hat{a}^2 n^2 (k-1)\right)\left(\hat{a}\ln(2\hat{a}^2 n)\right)}{\sqrt{n}}\right),$$

where $\hat{a} = 2aC_1C_2$ and we assume $\|w'\| \leq C_1$.

The generalization bound can be derived by applying $r_{S'} = k-1$ and using the fact that $\|z\| \leq 2C_2$ for any $z \in x'_i, \forall i \in [n]$ in the reduced MIL. By using Corollary 2, we can obtain the generalization risk bound for the Top-1 ranking learning.

The learning algorithm is designed by the following result:

Corollary 21. *The reduced ERM of MIL from top-1 ranking learning is a convex programming problem.*

The corollary can be esasily derived from the reduction process detailed in I.

Important extension: We can consider *top-1 ranking learning with negative feedback* which is an extension of top-1 ranking learning. We show the details in Sec.J. Remarkably, the ERM of the reduced MIL problem is a DC programming problem.

6 KERNELIZED EXTENSION

Although we consider a linear function set as $\mathcal{G}$; in practice, a nonlinear kernel is required for various learning tasks. A straightforward method is to employ a kernel-approximation technique [see, e.g., Sec.6.6 in Mohri et al., 2018], which constructs feature vectors $\Phi(x) \in \mathbb{R}^D$ with the theoretical guarantee that $\langle\Phi(x_1), \Phi(x_2)\rangle \approx K(x_1, x_2)$ for a user-determined dimension D. However, we can use only a limited number of kernels via the approximation technique. Therefore, we show the kernelized version of the reduction.

6.1 SETTINGS

We assume that an original problem is defined by $\mathcal{H}, \ell, \mathcal{X}, \mathcal{Y}$, and $\Phi : \mathcal{X} \to \mathbb{H}$, where $\mathbb{H}$ is a reproducing kernel Hilbert space associated to $K(x_1, x_2) = \langle\Phi(x_1), \Phi(x_2)\rangle$. Aside from the computability, we can virtually consider the sample as $S = ((\Phi(x_1), y_1), \ldots, (\Phi(x_n), y_n))$. The ERM-reducible condition is that there exist $(x', y') = \alpha(\Phi(x), y)$, $h = \beta(h')$ and ℓ' that satisfies $\ell(\Phi(x), y, h) = \ell'(x'_i, y'_i, h')$ for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

Let $S' = ((x'_1, y'_1), \ldots, (x'_n, y'_n))$ and let $\mathcal{G} = \{g : z \mapsto \langle w', z\rangle \mid w' \in \mathbb{H}'\}$. We assume that $(\mathcal{H}, \ell)$ is MIL-reducible to $\mathcal{H}', \ell'$. The ERM of the reduced MIL is formulated as:

$$\min_{w'\in\mathbb{H}'} \lambda\|w'\|_{\mathbb{H}'} + \mathcal{L}_{w'}, \tag{5}$$

where $\mathcal{L}_{w'} = \sum_i^n f_1\left(y'_i \Psi_p\left(\{f_2\left(\langle w', z\rangle \mid z \in x'_i\right)\}\right)\right)$.

6.2 COMPUTABILITY

We show that the representer theorem holds for the optimization problem (5).

Theorem 22 (Representer theorem)**.** *An optimal solution of the ERM problem (5) has the form* $\tilde{w}' = \sum_{z \in P_{S'}} \mu_z z$, *where* $P_{S'} = \bigcup_{i=1}^{n} x'_i$.

Thus, the ERM problem (5) is equivalently formulated as:

$$\min_{\boldsymbol{\mu} \in \mathbb{R}^{|P_{S'}|}} \lambda \sum_{z, \hat{z} \in P_{S'}} \mu_z \mu_{\hat{z}} \langle z, \hat{z} \rangle + \mathcal{L}_{\boldsymbol{\mu}},$$

where $\mathcal{L}_{\boldsymbol{\mu}} = \sum_{i=1}^{n} f_1(y_i \Psi_p(\{f_2(\sum_{z \in P_{S'}} \mu_z \langle z, \hat{z} \rangle) \mid \hat{z} \in x'_i\}))$.

Therefore, if $\langle z_1, z_2 \rangle$ is polynomial-time computable for any $z_1, z_2 \in x'$ using the original kernel function K as an oracle, the ERM of the MIL can be solved similar to linear case according to the condition in Proposition 5 and 6 (DC algorithm for the kernel version is in Sec. L). For all MIL-reducible problems introduced in the paper, $\langle z_1, z_2 \rangle$ is polynomial-time computable using K (see details in Sec.M). Moreover, we can construct β in polynomial time.

7 DISCUSSION AND CONCLUSION

7.1 RELATED WORK

Other reduction techniques: Several machine-learning reduction schemes exist [see, e.g., Beygelzimer et al., 2015], and we found general reduction schemes, such as [Pitt and Warmuth, 1990, Beygelzimer et al., 2005]. A major difference between the proposed scheme and existing approaches is that we focus on the reduction of ERM. Various applications of machine learning reductions, such as reduction from multi-class learning to binary classification [James and Hastie, 1998, Ramaswamy et al., 2014], and from ranking to binary classification [Balcan et al., 2008, Ailon and Mohri, 2010, Agarwal, 2014], exist. To the best of our knowledge, the reduction to MIL has not yet been discussed.

Multi-Class Learning: Recently, Lei et al. [2019] achieved $\log(k)$-dependent generalization bound. The proposed generalization bound is competitive with the bound. However, our derivation is highly simpler than the analysis of [Lei et al., 2019] because the reduction allows us to apply the existing MIL bound of [Sabato and Tishby, 2012].

Complementarily-labeled learning: Ishida et al. [2017] provided the generalization risk bound in the case in which the training sample contains only complementarily labeled instances (i.e., $\theta = 0$). The proposed generalization bound is incomparable to the bound (see details in Sec.N). Ishida et al. [2017] selected nonconvex loss functions and optimized the empirical risks using a gradient-based algorithm in practice. However, there is no guarantee of the optimality of the solution. We show that the learning problem can be solved by DC algorithm and guarantee the local optima. Moreover, in the special case that sample contains only complementarily labeled data, the learning problem becomes convex programming and we can obtain global optima. To the best of our knowledge, the provided learning algorithm is a first polynomial-time algorithm in the special case.

Multi-label learning: Various approaches and generalization analyses have been provided [Yu et al., 2014, Bhatia et al., 2015, Xu et al., 2016a,b]. However, to the best of our knowledge, this paper is the first to propose a $\log(k)$-dependent generalization bound for the linear (or nonlinear kernel) hypothesis class, where k is the number of classes.

Multi-task learning: A similar generalization bound was reported by [Pontil and Maurer, 2013]. Their results suggest the advantage of regularizing the weights $w_1, \ldots, w_T$ over T tasks. However, our result is derived from an entirely different argument from [Pontil and Maurer, 2013] and the derivation is highly simplified.

Top-1 ranking learning: Top-1 ranking measure was originally discussed in [Hidasi and Karatzoglou, 2018]. However, the basic problem setting is different from ours. They assumed that the recommender has i.i.d. positive and negative items as the sample. Moreover, they did not propose a general form of the problem and theoretical analysis.

MIL: MIL was originally proposed by Dietterich et al. [1997], which is known as weakly supervised learning and there have been proposed many real applications [Gärtner et al., 2002, Andrews et al., 2003, Zhang et al., 2013, Doran and Ray, 2014, Carbonneau et al., 2018]. Moreover, the generalization bound and learning algorithm are analyzed from the theoretical perspective [Sabato and Tishby, 2012, Doran, 2015, Suehiro et al., 2020]. Suehiro et al. [2020] found that a local-feature-based time-series classification problem can be reduced to a MIL problem. Our results first show that various learning problems can be reduced to MIL.

7.2 CONCLUSION AND FUTURE WORK

We revealed that various learning problems can be reduced to a MIL problem by our ERM-based reduction scheme. The results imply that our MIL-reduction gives a simplified and unified framework for the analyses for various learning problems. Moreover, we obtained novel theoretical results for some learning problems. As future work, we consider other reduction frameworks based on the proposed ERM-reduction. Moreover, we explore the relaxation of the reducible condition. In this paper, two learning algorithms are introduced and the demonstrated MIL-reducible problems are solved by either of them. However, there may be some MIL-reducible problems that can be solved by other learning algorithms or cannot be solved in polynomial time.

References

Shivani Agarwal. Surrogate regret bounds for bipartite ranking via strongly proper losses. *The Journal of Machine Learning Research*, 15(1):1653–1674, 2014.

Nir Ailon and Mehryar Mohri. Preference-based learning to rank. *Machine Learning*, 80(2-3):189–211, 2010.

Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems*, pages 577–584, 2003.

Maria-Florina Balcan, Nikhil Bansal, Alina Beygelzimer, Don Coppersmith, John Langford, and Gregory B Sorkin. Robust reductions from ranking to classification. *Machine learning*, 72(1-2):139–153, 2008.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003.

Alina Beygelzimer, Varsha Dani, Tom Hayes, John Langford, and Bianca Zadrozny. Error limiting reductions between classification tasks. In *International Conference on machine Learning*, pages 49–56, 2005.

Alina Beygelzimer, Hal Daumé, John Langford, and Paul Mineiro. Learning reductions that really work. *Proceedings of the IEEE*, 104(1):136–147, 2015.

Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. Sparse local embeddings for extreme multi-label classification. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Stephan Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329 – 353, 2018.

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

Gary Doran. *Multiple Instance Learning from Distributions*. PhD thesis, Case WesternReserve University, 2015.

Gary Doran and Soumya Ray. A theoretical and empirical analysis of support vector machine methods for multiple-instance classification. *Machine Learning*, 97(1-2):79–102, 2014.

Thomas Gärtner, Peter A. Flach, Adam Kowalczyk, and Alex J. Smola. Multi-instance kernels. In *International Conference on Machine Learning*, pages 179–186, 2002.

Balázs Hidasi and Alexandros Karatzoglou. Recurrent neural networks with top-k gains for session-based recommendations. In *International Conference on Information and Knowledge Management*, pages 843–852, 2018.

Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. Learning from complementary labels. In *Advances in neural information processing systems*, pages 5639–5649, 2017.

Gareth James and Trevor Hastie. The error coding method and picts. *Journal of Computational and Graphical statistics*, 7(3):377–387, 1998.

Hoai An Le Thi and Tao Pham Dinh. Dc programming and dca: thirty years of developments. *Mathematical Programming*, 169(1):5–68, 2018.

Y. Lei, Ü. Dogan, D. Zhou, and M. Kloft. Data-dependent generalization bounds for multi-class classification. *IEEE Transactions on Information Theory*, 65(5):2995–3021, 2019.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Leonard Pitt and Manfred K Warmuth. Prediction-preserving reducibility. *Journal of Computer and System Sciences*, 41(3):430–467, 1990.

Massimiliano Pontil and Andreas Maurer. Excess risk bounds for multitask learning with trace norm regularization. In *Conference on Learning Theory*, pages 55–76. PMLR, 2013.

Harish G Ramaswamy, Balaji Srinivasan Babu, Shivani Agarwal, and Robert C Williamson. On the consistency of output code based learning algorithms for multiclass learning problems. In *Conference on Learning Theory*, pages 885–902, 2014.

Sivan Sabato and Naftali Tishby. Multi-instance learning with any hypothesis class. *Journal of Machine Learning Research*, 13(1):2999–3039, October 2012. ISSN 1532-4435.

Daiki Suehiro, Kohei Hatano, Eiji Takimoto, Shuji Yamamoto, Kenichi Bannai, and Akiko Takeda. Theory and algorithms for shapelet-based multiple-instance learning. *Neural Computation*, 32(8):1580–1613, 2020.

Chang Xu, Tongliang Liu, Dacheng Tao, and Chao Xu. Local rademacher complexity for multi-label learning. *IEEE Transactions on Image Processing*, 25(3):1495–1507, 2016a.

Chang Xu, Dacheng Tao, and Chao Xu. Robust extreme multi-label learning. In *ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1275–1284, 2016b.

Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit Dhillon. Large-scale multi-label learning with missing labels. In *International Conference on Machine Learning*, pages 593–601, 2014.

Dan Zhang, Jingrui He, Luo Si, and Richard Lawrence. MILEAGE: Multiple instance learning with global embedding. In *International Conference on Machine Learning*, pages 82–90, 2013.

A PROOF OF THEOREM 3

Proof. In [Sabato and Tishby, 2012, Theorem 20 in], using the fact that ψ_p is 1-Lipschitz for all p and $\mathfrak{R}_S$ which is shown in the proof of [Sabato and Tishby, 2012, Theorem 20 in], we can obtain the target theorem. □

B PROOF OF PROPOSITION 5

Proof of Proposition 5. First we have that $\hat{f} = f_2 \circ g$ is a convex function of w' because f_2 is a nondecreasing convex and $\langle w', z \rangle$ is a convex function (see, e.g., Eq. (3.11) in Boyd and Vandenberghe [2004]). Subsequently, we show that $\Psi_p \circ \hat{f}$ is a convex function. Without loss of generality, we can consider Ψ_p as a function $\mathbb{R}^m \to \mathbb{R}$ where m is the size of the set x'. Ψ_p is a nondecreasing function in each argument and $\hat{f}$ is convex and thus $\Psi_p \circ \hat{h}$ is convex. Finally, because $-\Psi_p(\{f_2(\langle w', z \rangle) \mid z \in x'\})$ is concave and f_1 is nonincreasing convex, $f_1(-\Psi_p(\{f_2(\langle w', z \rangle) \mid z \in x'\})$ is convex Boyd and Vandenberghe [2004]. □

C PROOF OF PROPOSITION 6

Proof. Because $f_1(c)$ is a homogeneous function of degree 1 for $c \in [-1, 1]$, we have $f_1(-\Psi_p(\{f_2(\langle w', z \rangle) \mid z \in x'\})) = -f_1(\Psi_p(\{f_2(\langle w', z \rangle) \mid z \in x'\}))$. As we proved in Proof of Proposition 5, $f_1(-\Psi_p(\{f_2(\langle w', z \rangle) \mid z \in x'\}))$ is convex. Moreover, we have $f_1(\Psi_p(\{f_2(\langle w', z \rangle) \mid z \in x'\})) = -f_1(-\Psi_p(\{f_2(\langle w', z \rangle) \mid z \in x'\}))$ and thus $f_1(\Psi_p(\{f_2(\langle w', z \rangle) \mid z \in x'\}))$ is concave. Therefore, we have that $f_1(\Psi_p(\{f_2(\langle w', z \rangle) \mid z \in x'\})) + f_1(-\Psi_p(\{f_2(\langle w', z \rangle) \mid z \in x'\}))$ is a DC function. □

D DC ALGORITHM FOR THE REDUCED MIL PROBLEM

The algorithm is shown in Algorithm 1

Algorithm 1 MIL optimization via DC Algorithm

Inputs:
S', λ
Initialize:
$w'_0 \in \mathbb{R}^{d'}$
for $t = 1, \ldots,$ (until convergence) **do**
Compute the subgradient:

$$s_t \in \nabla_{w'} \left(\sum_{i:y_i=-1} f_1 \left(\Psi_p \left(\{ f_2 \left(\langle w', z \rangle \right) \mid z \in x'_i \} \right) \right) \right)$$

at w'_{t-1}.
Solve the following subproblem:

$$w'_t \leftarrow \arg \min_{w':\|w'\| \leq C_1} \lambda \|w'\|^2 + \sum_{i:y_i=+1} f_1 \left(\Psi_p \left(\{ f_2 \left(\langle w', z \rangle \right) \mid z \in x'_i \} \right) \right) - s_t^\top w' \tag{6}$$

end for
return w_t

The subproblem (6) is a convex programming problem that can be solved in polynomial time.

E PROOF OF LEMMA 1

Proof. Based on the assumption of $\mathcal{D}'$, the expected risk $R^{\mathrm{LC}}_{\mathcal{D}'}(h)$ is represented using $\mathcal{D}$, k, and θ as follows:

$$R^{\mathrm{LC}}_{\mathcal{D}'}(h) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\theta I\left((y \neq h(x))\right) + (1-\theta)\sum_{\bar{y}\neq y}\frac{1}{k-1} I\left(\bar{y} = h(x)\right).\right]$$

Let $\rho_1 = I\left(y \neq h(x)\right)$ in $R^{\mathrm{MC}}_{\mathcal{D}}(h)$ and let $\rho_2 = \theta I\left((y \neq h(x))\right) + (1-\theta)\sum_{\bar{y}\neq y}\frac{1}{k-1} I\left((\bar{y} = h(x))\right)$ in $R^{\mathrm{LC}}_{\mathcal{D}'}(h)$. We consider two cases of h for any $h \in \mathcal{H}$ as follows: For a fixed (x, y), (i) If $h(x) = y$: $\rho_1 = 0$ and $\rho_2 = 0$, and thus there is no gap. (ii) If $h(x) \neq y$:, the first term of ρ_2 is θ and the second term is equal to $(1-\theta)/(k-1)$, because there exists a unique $\hat{y} : \hat{y} \neq y$ that satisfies $\hat{y} = h(x)$. Therefore, ρ_2 is equal to $\theta + \frac{1-\theta}{k-1}$. In this case, $\rho_1 = 1$. Thus, we have the bound $\frac{k-1}{\theta(k-2)+1} R^{\mathrm{LC}}_{\mathcal{D}'}(h) = R^{\mathrm{MC}}_{\mathcal{D}}(h)$. □

F PROOF OF THEOREM 10

Proof. We use $\eta_{(x,y)}$ defined in (5.1.1). On the MIL-reduction framework, suppose that $p = \infty$; $f_1(c) = \Gamma(2cC_1C_2)$; $f_2(c) = c/2C_1C_2$ (shifting function to $[-1, +1]$); $\alpha(x, (\gamma, y)) = (x'_{(x,y)}, y')$ where $x'_{(x,y)} = \{\eta_{(x,j)} - \eta_{(x,y)} \mid \forall j \in \mathcal{Y}\backslash y\}$; $y' = I(\gamma = \mathrm{True})$; for any $z \in \mathbb{R}^{kd}$, $\mathcal{G} = \{g : z \mapsto \langle (w'_1, \ldots, w'_k), z\rangle \mid w'_j \in \mathbb{R}^d, \forall j \in [k], \|W'\| \leq C_1\}$ where $W' = (w'_1, \ldots, w'_k)$ and $\|W'\| = \sqrt{\sum_{j=1}^k \|w'_j\|^2}$; $\beta(h') : x \mapsto \arg\max_{j\in[k]}\langle (w'_1, \ldots w'_k), x\rangle$. Then, for any (x, y) and $h \in \mathcal{H}$,

$$\begin{aligned}
\ell'(x', y', h') =& f_1\left(y'\Psi_p\left(\{f_2\left(g(z) \mid z \in x'_{(x,y)}\right\}\right)\right)\\
=& \Gamma\left(I(\gamma = \mathrm{True}) \times \Psi_\infty\left(\{g(z) \mid z \in x'_{(x,y)}\}\right)\right)\\
=& \Gamma\left(I(\gamma = \mathrm{True}) \times \left(\max_{j\in\mathcal{Y}\backslash y}\left(\langle w_j, x\rangle - \langle w_y, x\rangle\right)\right)\right)\\
=& \ell(x, (\gamma, y), h).
\end{aligned}$$

□

G REDUCTION FROM MULTI-TASK LEARNING PROBLEM

In multi-task learning, the learner finds a common rule in multiple-tasks, which correctly predicts the outputs of the instances. For example, in the multi-classification-task problem, there are three different binary classification tasks for image data, cat or dog, car or train, and apple or tomato.

Problem setting Let $\mathcal{X} \subseteq \mathbb{R}^d$ be an input space and $\mathcal{Y} \in \{-1, 1\}$ be an output space. We assume that the learner has T different tasks with different data distributions. The learner receives T sets of samples $S = S_1, \ldots, S_T$ where $S_t = ((x^t_1, y^t_1), \ldots, (x^t_n, y^t_n))$ is drawn i.i.d. according to unknown distribution $\mathcal{D}_t$. (x^t, y^t) denote an instance and its label, respectively. Let $\mathcal{H} = \{h : (x^t) \mapsto \mathrm{sign}(\langle w_t, x^t\rangle) \mid w_t \in \mathbb{R}^d\rangle\}$ be a hypothesis class. Let $\ell : ((x^1, \ldots, x^T), (y^1, \ldots, y^T), h) \mapsto \frac{1}{T}\sum_{t=1}^T \Gamma(-y^t\langle w_t, x^t\rangle)$ where $\Gamma : \mathbb{R} \to [0, 1]$ is a convex, nondecreasing and b-Lipschitz function. The generalization risk and empirical risk are formulated as:

$$\mathop{\mathbb{E}}_{t}[R_{\mathcal{D}_t}(h)] = \frac{1}{T}\sum_{t=1}^T \mathop{\mathbb{E}}_{(x^t,y^t)\sim\mathcal{D}_t}\left[\Gamma(-y^t\langle w_t, x^t\rangle)\right],$$

$$\widehat{R}_S(h) = \frac{1}{T}\sum_{t=1}^T \frac{1}{n}\sum_{i=1}^n \Gamma(-y^t_i\langle w_t, x^t_i\rangle) = \frac{1}{n}\sum_{i=1}^n \ell\left((x^1_i, \ldots, x^T_i), (y^1_i, \ldots, y^T_i), h\right).$$

Reduction to MIL

Theorem 23. *Multi-task learning is MIL-reducible.*

Proof. For simplicity, we denote $(x^1, \ldots, x^T)$ by $\mathbf{x}$ and denote $(y^1, \ldots, y^T)$ by $\mathbf{y}$. On the MIL-reduction framework, suppose that $p = 1$; $f_1 : f_1(a) = -a$; f_2 is Γ; $\alpha(\mathbf{x}, \mathbf{y}) = (x'_{(\mathbf{x},\mathbf{y}))}, y')$ where $x'_{(\mathbf{x},\mathbf{y})} = \{(y^1 x^1, 1), \ldots, (y^T x^T, T)\}$; $y' = -1$; $\mathcal{G} = \{g : (z, t) \mapsto \langle w'_t, z\rangle \mid \forall j \in [T], w'_t \in \mathbb{R}^d \text{ and } \|W'\| \leq C_1\}$ where $W' = (w'_1, \ldots, w'_T)$; $\beta(h') : (x^t) \mapsto \mathrm{sign}(\langle w'_t, x^t\rangle)$. For any $((x^1, \ldots, x^T), (y^1, \ldots, y^T))$ and $h \in \mathcal{H}$, we have that

$$\begin{aligned}
\ell'(x', y', h') =& f_1\left(y'\Psi_p\left(\left\{f_2\left(g(z)\right) \mid z \in x'_{(\mathbf{x},\mathbf{y})}\right\}\right)\right) \\
=& \frac{1}{|x'_{(\mathbf{x},\mathbf{y})}|} \sum_{(x,t)\in x'_{(\mathbf{x},\mathbf{y})}} \Gamma\left(-\langle w_t, y^t x^t\rangle\right) \\
=& \ell((x^1, \ldots, x^T), (y^1, \ldots, y^T), h)
\end{aligned}$$

□

ERM algorithm

Corollary 24. *The reduced ERM of the MIL from multi-task learning is a convex programming problem.*

As shown in the proof of Theorem 23, f_1 is nonincreasing and $y'_i = -1$ for all $i \in [n]$. Thus, by Proposition 5, if we consider Γ that is nondecreasing and convex, the reduced MIL problem is a convex programming problem and solved in polynomial time.

Generalization bound

Corollary 25. *We assume that $\|x_i^t\| \leq C_2$ for any $i \in [n]$ and $t \in [T]$. In the reduced problem, the empirical Rademacher complexity of $\widehat{\mathcal{H}}'$ is given as follows:*

$$\mathfrak{R}_{S'}(\widehat{\mathcal{H}}') = O\left(\frac{\log\left(2n^2T\right)\left(bC_1C_2\ln(n)\right)}{\sqrt{n}}\right),$$

where we assume $\|w'\| \leq C_1$.

We can derive the above from the same argument from the proof of Theorem 23. Using Corollary 2, we can obtain the generalization risk bound for the multi-task learning problem.

H PROOF OF THEOREM 16

Proof. On the MIL-reduction framework, suppose that $p = \infty$; $f_1 : f_1(a) = -a$ for $a \in \mathbb{R}$; f_2 is Γ; $\alpha(x, y) = (x'_{(x,y)}, y')$ where $x'_{(x,y)} = \{(-y^1 x, 1), \ldots, (-y^k x, k)\}$; $y' = -1$; $\mathcal{G} = \{g : (z, j) \mapsto \langle w'_j, z\rangle \mid w'_j \in \mathbb{R}^d, \forall j \in [k], \|W'\| \leq 1\}$ where $W' = (w'_1, \ldots, w'_k)$; $W' = (w'_1, \ldots, w'_k)$; $\beta(h') : (x, j) \mapsto \langle w'_j, x\rangle$. For any (x, y) and $h \in \mathcal{H}$, we have that

$$\begin{aligned}
\ell'(x', y', h') =& f_1\left(y'\Psi_p\left(\left\{f_2\left(g(z)\right) \mid z \in x'_{(x,y)}\right\}\right)\right) \\
=& \max_{(y^j x, j)\in x'_{(x,y)}} \Gamma\left(-\langle w_j, y^j x\rangle\right) \\
=& \ell(x, y, h)
\end{aligned}$$

□

I PROOF OF THEOREM 19

Proof. On the MIL-reduction framework, suppose that $p = \infty$; $f_1(c) = \Gamma(2cC_1C_2)$; $f_2(c) = c/2C_1C_2$; $\alpha(A, x^*) = (x', y')$ where $x' = \{x - x^* \mid x \in A \backslash x^*\}$; $y' = -1$; $\mathcal{G} = \{g : z \mapsto \langle w', z\rangle \mid \|w'\| \leq C_1\}$; $\beta(h') : A \mapsto \arg\max_{x \in A} \langle w', x\rangle$. For any (A, x^*) and $h \in \mathcal{H}$, the following holds:

$$
\begin{aligned}
\ell'(x', y', h') =& f_1\left(y'\Psi_p\left(\{f_2\left(g(z) \mid z \in x'_{(x,y)}\right\}\right)\right) \\
=& \Gamma\left(-\Psi_\infty\left(\{g(z) \mid z \in x'_{(x,y)}\}\right)\right) \\
=& \Gamma\left(-\left(\max_{j \in A \backslash x^*}\left(\langle w, x\rangle - \langle w, x^*\rangle\right)\right)\right) \\
=& \ell(A, x^*, h)
\end{aligned}
$$

□

J TOP-1 RANKING LEARNING WITH NEGATIVE FEEDBACK

As an extension of the Top-1 rank learning problem, we consider the following scenario. In practice, some item sets do not include the user-preferred item. Therefore, we assume that the item sets are partitioned into two types: the item sets that include the most preferred item and those that do not include the preferred item. For the second type of item set, we assume that we can receive information on non-preferred items as negative feedback from the user.

More formally, we assume that the target user has a scoring function s and a parameter $\gamma_i \in \{-1, +1\}$, where γ takes $+1$ for an item set that includes the preferred item and takes -1 otherwise. The learner receives the sequence of the sets of items and the chosen item with positive or negative information $S = (A_1, (x_1^*, \gamma_1)), \ldots, (A_n, (x_n^*, \gamma_n)$. $\gamma_i = +1$ indicates that item set A_i includes the preferred item, and $\gamma_i = -1$ indicates that the item set A_i does not include the preferred item. For the item set A_i with $\gamma = +1$, $x_i^* = \max_{x \in A_i} s(x)$. Conversely, for the item set A_i with $\gamma = -1$, $x_i^* \in \{A' = A \backslash x' \mid x' = \max_{x \in A_i} r(x)\}$, that is, if $\gamma = -1$, the user selects an item except for the best-scored item by s. Note that we assume that γ is a known parameter only in the training phase. The other settings are the same as those in Sec. 5.2.2.

A reasonable goal of the learner is to predict the best item from a given set of items even in this setting. Therefore, the learner can recommend the most preferred item if $\gamma = +1$ and can recommend a preferable item if $\gamma = -1$. Similar to top-1 ranking learning, we consider a loss function $\ell : (A, (x^*, \gamma), h) \mapsto \Gamma(\gamma(\langle w, x^*\rangle - \max_{x \in A \backslash x^*} \langle w, x\rangle))$ where $\Gamma : \mathbb{R} \to [0, 1]$ is a convex, nonincreasing and a-Lipschitz function. The generalization risk and empirical risk are formulated as follows:

$$
\begin{aligned}
R_{\mathcal{D}}(h) &= \mathop{\mathbb{E}}_{(A,\gamma) \sim \mathcal{D}}\left[\ell\left(A, (x^*, \gamma), h\right)\right], \\
\widehat{R}_S(h) &= \frac{1}{n}\sum_{i=1}^{n} \ell\left(A, (x_i^*, \gamma_i), h\right),
\end{aligned}
$$

where $x^* = \arg\max_{x \in A} s(x)$.

Reduction to MIL

Theorem 26. *Top-1 ranking learning with negative feedback is MIL-reducible.*

The difference from the top-1 ranking learning is just $y'_i = -\gamma_i$, and thus we can easily prove it.

Proof. On the MIL-reduction framework, suppose that $p = \infty$; $f_1(c) = \Gamma(2cC_1C_2)$; $f_2(c) = c/2C_1C_2$; $\alpha(A, x^*) = (x', y')$ where $x' = \{x - x^* \mid x \in A \backslash x^*\}$; $y' = -\gamma$; $\mathcal{G} = \{g : z \mapsto \langle w', z\rangle \mid \|w'\| \leq 1\}$; $\beta(h') : A \mapsto \arg\max_{x \in A} \langle w', x\rangle$.

For any (A, x^*) and $h \in \mathcal{H}$, the following holds:

$$\begin{aligned}\ell'(x', y', h') =& f_1\left(y'\Psi_p\left(\{f_2\left(g(z) \mid z \in x'_{(x,y)}\}\right)\right)\right)\\ =&\Gamma\left(\gamma\left(\Psi_\infty\left(\{g(z) \mid z \in x'_{(x,y)}\}\right)\right)\right)\\ =&\Gamma\left(\gamma\left(\max_{j\in A\setminus x^*}\left(\langle w, x\rangle - \langle w, x^*\rangle\right)\right)\right)\\ =&\ell(A, x^*, h)\end{aligned}$$

□

Generalization bound

Corollary 27. *We assume that $\|x\| \leq C_2$ for any $x \in A_i \forall i \in [n]$. In the reduced MIL problem, the empirical Rademacher complexity of $\widehat{\mathcal{H}}'$ is given as follows:*

$$\mathfrak{R}_{S'}(\widehat{\mathcal{H}}') = O\left(\frac{\log\left(\hat{a}^2 n^2 (k-1)\right)\left(2\hat{a}\ln(\hat{a}^2 n)\right)}{\sqrt{n}}\right),$$

where $\hat{a} = 2aC_1C_2$ we assume $\|w'\| \leq C_1$.

Using Corollary 2, we can obtain the generalization risk bound for the Top-1 ranking learning with negative feedback.

ERM algorithm

Corollary 28. *The reduced ERM of MIL from top-1 ranking learning with negative feedback is a DC programming problem.*

In top-1 ranking learning, $y' \in \{-1, 1\}$. By the proof of Theorem 26 and by Proposition 6, if we consider a loss function $\Gamma(c)$ as a nondecreasing and homogeneous function of degree 1 for $c \in [-1, 1]$ such as hinge-loss, we can solve the problem by DC algorithm as shown in Algorithm 1.

K PROOF OF THEOREM 22

Proof. For the optimization problem (5), we can apply the standard representer theorem (see, e.g., Theorem 6.11 of Mohri et al. [2018]). We define $\mathbb{H}_1$ as the subspace spanned by $\{\langle z, \cdot\rangle \mid z \in P_{S'}\}$, namely, $\mathbb{H}_1 = \{w \in \mathbb{H} \mid w = \sum_{z\in P_{S'}} \mu_z z, \mu_z \in \mathbb{R}\}$. For any $w \in \mathbb{H}$, we can consider the decomposition $w = w_1 + w_1^\perp$, where $w_1 \in \mathbb{H}_1$, and $w_1^\perp \in \mathbb{H}_1^\perp$ is its orthogonal component. Because $\mathbb{H}_1$ is a subspace of $\mathbb{H}$, $\|w\|_\mathbb{H} = \sqrt{\|w_1\|_\mathbb{H}^2 + \|w_1^\perp\|_\mathbb{H}^2} \geq \|w_1\|_\mathbb{H}$. Moreover, by the definition of $\mathbb{H}_1$, $\langle w, z\rangle = \langle w_1, z\rangle$. Thus, $f_1(y'_i\Psi_p(\{f_2(\langle w, z\rangle) \mid z \in x'_i\})) = f_1(y'_i\Psi_p(\{f_2(\langle w_1, z\rangle) \mid z \in x'_i\}))$ and $\|w_1\|_\mathbb{H} \leq \|w\|_\mathbb{H}$. This implies that the optimal solution is contained in $\mathbb{H}_1$. □

L DC ALGORITHM FOR KERNELIZED EXTENSION

The algorithm is shown in Algorithm 2.

M EXAMPLE:REDUCTION FROM MULTI-CLASS LEARNING WITH KERNEL

M.1 REDUCTION TO MIL WITH KERNEL

Theorem 29. *Multi-class learning with kernel is MIL-reducible.*

Algorithm 2 MIL optimization via DC Algorithm

Inputs:
S', λ
Initialize:
$\boldsymbol{\mu}_0 \in \mathbb{R}^{|P_{S'}|}$
for $t = 1, \ldots,$ (until convergence) **do**
Compute the subgradient:

$$s_t \in \nabla_{\boldsymbol{\mu}} \left(\sum_{i:y_i=-1} f_1 \left(\Psi_p \left(\left\{ f_2 \left(\sum_{v \in P_{S'}} \mu_v K(v,z) \right) \mid z \in x'_i \right\} \right) \right) \right)$$

at $\boldsymbol{\mu}_{t-1}$.
Solve the following subproblem:

$$\begin{aligned} \boldsymbol{\mu}_t \leftarrow \arg \min_{\boldsymbol{\mu} \in \mathbb{R}^{|P_{S'}|}} & \lambda \sum_{v,\hat{v} \in P_{S'}} \mu_v \mu_{\hat{v}} K(v,\hat{v}) \\ & + \sum_{i:y_i=+1} f_1 \left(\Psi_p \left(\left\{ f_2 \left(\sum_{v \in P_{S'}} \mu_z K(z,x) \right) \mid z \in x'_i \right\} \right) \right) \\ & - s_t^\top \boldsymbol{\mu} \end{aligned}$$

end for
return $\boldsymbol{\mu}_t$

Proof. For any (x, y), we define

$$\eta_{(x,y)} = (0_{\mathbb{H}}, \ldots, 0_{\mathbb{H}}, \underbrace{\Phi(x)}_{y-\text{th block}}, 0_{\mathbb{H}}, \ldots, 0_{\mathbb{H}}) \in \mathbb{H}^k, \tag{7}$$

where $0_{\mathbb{H}}$ is a point in $\mathbb{H}$ satisfying $\langle 0_{\mathbb{H}}, v \rangle = 0$ for any $v \in \mathbb{H}$. On the MIL-reduction framework, suppose that $p = \infty$; $f_1(c) = \Gamma(cC_1C_2)$; $f_2(c) = c/C_1C_2$; $\alpha(x,y) = (x'_{(x,y)}, y')$ where $x'_{(x,y)} = \{\eta_{(x,j)} - \eta_{(x,y)} \mid \forall j \in \mathcal{Y} \backslash y\}$; $y' = -1$; $\mathcal{G} = \{g : z \mapsto \langle (w'_1, \ldots, w'_k), z \rangle \mid \forall j \in [k], w'_j \in \mathbb{H}, \|W'\|_{\mathbb{H}^k} \leq C_1\}$ where $W' = (w'_1, \ldots, w'_k)$, $\|W'\|_{\mathbb{H}^k} = \sqrt{\sum_{j=1}^k \|w'_j\|_{\mathbb{H}}^2}$. Then, for any (x, y) and $h \in \mathcal{H}$,

$$\begin{aligned} \ell'(x', y', h') =& f_1 \left(y' \Psi_p \left(\{ f_2 \left(g(z) \mid z \in x'_{(x,y)} \} \right) \right) \right) \\ =& \Gamma \left(-\Psi_\infty \left(\{ g(z) \mid z \in x'_{(x,y)} \} \right) \right) \\ =& \Gamma \left(- \left(\max_{j \in \mathcal{Y} \backslash y} \left(\langle W', \eta_{(x,j)} - \eta_{(x,y)} \rangle \right) \right) \right) \\ =& \Gamma \left(- \left(\max_{j \in \mathcal{Y} \backslash y} \left(\langle w_j, \Phi(x) \rangle - \langle w_y, \Phi(x) \rangle \right) \right) \right) \\ =& \ell(x, y, h) \end{aligned}$$

□

M.2 CONSTRUCTION OF β

By Theorem 22, W' is returned by using μ as

$$W' = \sum_{z \in P_{S'}} \mu_z z.$$

Moreover, w'_j can be represented as:

$$w'_j = \sum_{z[j] \in P_{S',j}} \mu_{z[j]} v[j],$$

where $P_{S',j} = \{z[j] \mid z \in \bigcup_{i=1}^{n} x'_i\}$ and $z[j]$ is j-th block of z. That is, $z[j]$ can be rewritten as $\Phi(\tilde{x}_j)$ for some $\tilde{x}_j$. Note that, because z is based on $\eta_{(x,y)}$ as shown in (7), $z[j]$ is in the Hilbert space $\mathbb{H}$ in the original problem. Based on the relationship between $W' = (w'_1, \ldots, w'_k)$ and $W = (w_1, \ldots, w_k)$, therefore, the hypothesis $h(x)$ in the original problem is obtained by:

$$\begin{aligned} h(x) &= \arg\max_{j \in [k]} \langle w_j, \Phi(x) \rangle \\ &= \arg\max_{j \in [k]} \langle w'_j, \Phi(x) \rangle \\ &= \arg\max_{j \in [k]} \sum_{z[j] \in P_{S',j}} \mu_{z[j]} \langle z[j], \Phi(x) \rangle \\ &= \arg\max_{j \in [k]} \sum_{\tilde{x_j}} \mu_{\tilde{x}_j} K(\tilde{x_j}, x). \end{aligned}$$

We can show that the other learning problems presented in this paper can be kernelized. For the other learning problems introduced in this study, there are two types of the domains of z: the concatenation of the Hilbert vector (complementarily labeled learning problems, multi-label learning, multi-task learning) and difference of the Hilbert vector (top-1 ranking learning). For the difference in the Hilbert vector, that is, for $z = \Phi(x_1) - \Phi(x_2)$ and $\Phi(x)$, $\langle z, \Phi(x) \rangle$ can be computed as:

$$\begin{aligned} &\langle z, \Phi(x) \rangle \\ =&\langle \Phi(x_1) - \Phi(x_2), \Phi(x) \rangle \\ =&K(x_1, x) - K(x_2, x), \end{aligned}$$

and thus $h(x)$ is polynomial-time computed by h'.

N COMPARISON TO THE EXISTING GENERALIZATION BOUND FOR COMPLEMENTARILY LABELED LEARNING

Ishida et al. [2017] stated that, for a linear-hypothesis class, the following bound holds with a probability of at least $1 - \delta$: $R_{\mathcal{D}}^{\mathrm{MC}}(h) \leq \widehat{R}(h) + k(k-1)L\sqrt{R\Lambda/n} + (k-1)\sqrt{8\ln(2/\delta)/n}$. They used the empirical risk $\widehat{R}(h)$ for complementarily labeled instances, which is different from the risk that we defined [see details in Ishida et al., 2017]. According to this difference, the proposed generalization bound is incomparable to the existing bound. However, we can say that if we achieve a small empirical risk close to zero, the proposed risk bound is k times tighter than the existing bound.