Simplified and Unified Analysis of Various Learning Problems by Reduction to Multiple-Instance Learning (Supplementary materials)

Daiki Suehiro[1, 2] **Eiji Takimoto**[3]

[1]Kyushu University, Department of Advanced Information Technology, 744 Motooka, Fukuoka, Japan
[2]RIKEN, Center for Advanced Intelligence Project, Nihonbashi 1-chome Mitsui Building, 15th floor,1-4-1 Nihonbashi, Chuo-ku, Tokyo, Japan
[3]Kyushu University, Department of Informatics,744 Motooka, Fukuoka, Japan

A PROOF OF THEOREM 3

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

B PROOF OF PROPOSITION 5

Proof. 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 of w' (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. The subproblem (A-1) 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:

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

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

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:

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

end for
return w_t

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 scheme, 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 $x'_{(x,y)} = \{\eta_{(x,j)} - \eta_{(x,y)} \mid \forall j \in \mathcal{Y} \backslash y\}$; $y' = I(\gamma = \text{True})$; 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 $\|W'\| = \sqrt{\sum_{j=1}^k \|w'_j\|^2}$; $\beta(h') : x \mapsto \arg\max_{j \in [k]} \langle w'_j, x \rangle$. Then, for any (x, y) and $h \in \mathcal{H}$,

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

□

G 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 $S_t = ((x_1^t, y_1^t), \ldots, (x_n^t, y_n^t))$ is drawn i.i.d. according to unknown distribution $\mathcal{D}_t$. (x^t, y^t) denote an instance and its label, respectively. Let $\mathcal{H} = \{h : (x^t) \mapsto \text{sign}(\langle w_t, x^t \rangle) \mid w_t \in \mathbb{R}^d\rangle\}$ 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:

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

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

Reduction to MIL

Theorem 1. *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 scheme, 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 $W' = (w'_1, \ldots, w'_T)$; $\beta(h') : (x^t) \mapsto \text{sign}(\langle w'_t, x^t \rangle)$. 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 2. *The reduced ERM of the MIL from multi-task learning is a convex programming problem.*

As shown in the proof of Theorem 1, 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 3. *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 1. 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 scheme, 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 scheme, 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 4. *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 scheme, 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:

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

□

Generalization bound

Corollary 5. *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 6. *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 4 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.

Algorithm 2 MIL optimization via DC Algorithm (kernelized)

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

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

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

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

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

M EXAMPLE OF THE REDUCTION OF KERNELIZED LEARNING PROBLEMS: MULTI-CLASS LEARNING

M.1 REDUCTION TO MIL WITH KERNEL

Theorem 7. *Multi-class learning with kernel is MIL-reducible.*

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

$$\eta_{(x,y)} = (0_{\mathbb{H}}, \ldots, 0_{\mathbb{H}}, \underbrace{\Phi(x)}_{y-\text{th block}}, 0_{\mathbb{H}}, \ldots, 0_{\mathbb{H}}) \in \mathbb{H}^k, \tag{A-2}$$

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 scheme, 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 (A-2), $z[j]$ 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:

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

M.3 REDUCTION OF OTHER KERNELIZED LEARNING PROBLEMS

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 computed by h' in polynomial time.

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 $1 - \delta$: $R_{\mathcal{D}}^{\mathrm{MC}}(h) \leq \widehat{R}(h) + ak(k-1)\frac{C_1 C_2}{\sqrt{n}} + (k-1)\sqrt{8\ln(2/\delta)/n}$. 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.

O ARTIFICIAL DATASETS ON COMPLEMENTARILY LABELED LEARNING

We prepared three datasets, artificial1, artificial2, and artificial3. Each dataset has 1000 training and 1000 test instances. The number of dimension d is 50. They have 5, 10, and 25 classes, respectively. The feature values of each data is determined by the following rule: If the data belongs to class j, $\{\frac{(j-1)d}{k} + 1, \ldots, \frac{jd}{k}\}$-th features have the values drawn according to $\mathcal{N}(2, 1)$ and other features have the values drawn according to $\mathcal{N}(0, 1)$.

References

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

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

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

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