Correlation-Induced Label Prior for Semi-Supervised Multi-Label Learning

Biao Liu [1 2] **Ning Xu** [1 2] **Xiangyu Fang** [1 2] **Xin Geng** [1 2]

Abstract

Semi-supervised multi-label learning (SSMLL) aims to address the challenge of limited labeled data availability in multi-label learning (MLL) by leveraging unlabeled data to improve the model's performance. Due to the difficulty of estimating the reliable label correlation on minimal multi-labeled data, previous SSMLL methods fail to unlash the power of the correlation among multiple labels to improve the performance of the predictive model in SSMLL. To deal with this problem, we propose a novel SSMLL method named PCLP where the correlation-induced label prior is inferred to enhance the pseudo-labeling instead of directly estimating the correlation among labels. Specifically, we construct the correlated label prior probability distribution using structural causal model (SCM), constraining the correlations of generated pseudo-labels to conform to the prior, which can be integrated into a variational label enhancement framework, optimized by both labeled and unlabeled instances in a unified manner. Theoretically, we demonstrate the accuracy of the generated pseudo-labels and guarantee the learning consistency of the proposed method. Comprehensive experiments on several benchmark datasets have validated the superiority of the proposed method. Source code is available at https://github.com/palm-biaoliu/pclp.

1. Introduction

Multi-label learning (MLL) has emerged as an effective paradigm for handling instances associated with multiple semantic labels, which is common in many real-world applications. Over the past decade, with the strong representation learning ability of deep neural networks, MLL has been successfully applied to a variety of real-world applications such as image annotation (Lanchantin et al., 2021), text classification (Liu et al., 2017), and facial expression recognition (Chen et al., 2020).

In the context of MLL, the complexity of the label space makes the acquisition of labeled data both time-consuming and costly, particularly for large-scale datasets that require expert labeling. This challenge has given rise to the development of Semi-Supervised Multi-Label Learning (SSMLL), which aims to utilize the information in unlabeled data to improve model generalization performance when only limited labeled instances are available for training. Existing SSMLL methods primarily seek inspiration from approaches within single-label semi-supervised learning (SSL), which has made great progress in recent years (Guo & Li, 2022; Wang et al., 2023). For instance, Zhang et al. employed a co-training strategy, where pairwise ranking predictions on unlabeled data are communicated between classifiers for model refinement (Zhan & Zhang, 2017). Similarly, Wang et al. introduced a dual-classifier framework to align the feature distribution in a latent space while generating pseudo-labels for unlabeled instances (Wang et al., 2020). More recently, a class-specific threshold strategy is designed based on the estimated class prior probabilities, which is then used to assign pseudo-labels (Xie et al., 2023).

The complex co-occurrence or mutual exclusion relationships, (Zhang & Zhou, 2013) in multiple labels, i.e., label correlation, is a critical property in MLL. However, the previous methods only directly adopt single-label semi-supervised framework for each class or just utilize an estimated unreliable label correlation matrix for propagating relationships between labels at the output layer. As directly estimating the reliable label correlation on the available minimal labeled instances in SSMLL is difficult, previous SSMLL methods fail to unlash the power of the correlation among multiple labels to improve the performance of the predictive model in SSMLL. In response to this challenge, we consider implicitly constructing label correlation through the correlated label prior probability distribution, which can be inferred with both labeled and unlabeled data, thereby achieving more reliable label correlation.

[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China [2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China. Correspondence to: Ning Xu <xning@seu.edu.cn>, Xin Geng <xgeng@seu.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

In this paper, we propose a novel method named PCLP, i.e., Pseudo-labeling with Correlation-induced Label Prior for SSMLL. Specifically, we construct the correlated label prior probability distribution using Structural Causal Model (SCM) (Yu et al., 2019), which can be seamlessly integrated into a variational label enhancement framework (Xu et al., 2021a; 2020b; 2023), optimized by both labeled and unlabeled data in a unified manner. Simultaneously, the prior can be employed to guide the pseudo-labeling process, ensuring that the correlation of the generated pseudo-labels conforms to the prior. Finally, the predictive model is trained on both labeled instances and unlabeled instances assigned with the pseudo-labels. Additionally, we theoretically demonstrate that the generated pseudo-labels will be more accurate when considering the label correlation inherent in the label prior distribution. Furthermore, the generalization error bound of the proposed method is derived to guarantee the learning consistency (Mohri et al., 2018). The main contributions of this paper are summarized as follows:

- Practically, we propose a novel pseudo-label generation method named PCLP for SSMLL. The proposed method leverages both labeled and unlabeled instances to establish the correlated label prior probability distribution and generates pseudo-labels for unlabeled instances guided by this prior, which ensures that the correlation of the generated pseudo-labels conforms to the learned prior.
- Theoretically, we demonstrate the enhanced accuracy of pseudo-labels when taking into account the label correlation inherent in the label prior distribution. Furthermore, the generalization error bound is derived to guarantee learning consistency.

We conduct extensive experiments on several benchmark datasets to demonstrate the effectiveness of our method. The results show that our method outperforms the state-of-the-art methods.

2. Related Work

Multi-label learning: Multi-label learning is a supervised machine learning technique where an instance is associated with multiple labels simultaneously. A key research focus has been modeling and exploiting the correlations between different labels to improve multi-label classification performance. Early works mainly focused on first-order label correlation by adapting binary classification algorithms like support vector machines to the multi-label setting. These treat each label as an independent binary classification problem without considering relationships between labels (Boutell et al., 2004; Read et al., 2011). Second-order correlation methods aim to capture pairwise relationships between labels (Elisseeff & Weston, 2001; Fürnkranz et al., 2008). However, considering only pairs of labels limits modeling higher-order correlations. More recent works have explored high-order correlations that take into account complex relationships between multiple labels. Examples are graph convolutional networks that operate on a label correlation graph (Chen et al., 2019), and recurrent neural networks like LSTMs that can learn arbitrary label dependencies (Yazici et al., 2020). Beyond correlations, some recent works have explored using label-specific features tailored to each label to improve performance (Yu & Zhang, 2022; Hang & Zhang, 2022). Building upon these advancements, there is an increasing focus on weak supervision scenarios due to the the complexities of the label space in MLL, such as noisy multi-label learning (Li et al., 2022; Chen et al., 2024), partial multi-label learning (Wang et al., 2019; Xu et al., 2020a; Xie & Huang, 2021), multi-label learning with missing labels (Kim et al., 2022; 2023) and single-positive multi-label learning (Cole et al., 2021; Xu et al., 2022; Xie et al., 2022; Liu et al., 2023).

Semi-supervised learning: The semi-supervised learning (SSL) methods primarily rely on pseudo-labeling and consistency regularization. Pseudo-Labeling utilizes pseudo-labels derived from model predictions, often paired with a confidence-based thresholding that retains unlabeled instances only when the classifier is sufficiently confident (Sohn et al., 2020; Xie et al., 2020; Zhang et al., 2021; Xu et al., 2021b). For instance, Zhang et al. designed varying thresholds for different categories, lowering thresholds for hard-to-learn categories to alleviate class imbalance (Zhang et al., 2021). Xu et al. progressively increased thresholds during training to enhance the utilization of unlabeled data (Xu et al., 2021b). Consistency regularization, another common technique in SSL, enforces the model to produce consistent predictions for different perturbations of the same instance. Techniques for generating random perturbations include data augmentation (French et al., 2018), stochastic regularization (Sajjadi et al., 2016; Laine & Aila, 2017), and adversarial perturbations (Miyato et al., 2018). More recently, it has been shown that combining these two approaches, pseudo-labeling and consistency regularization, can lead to enhanced performance (Berthelot et al., 2020; Sohn et al., 2020; Zheng et al., 2022; Yang et al., 2022; Chen et al., 2022).

Semi-supervised multi-label learning: Recent work has developed semi-supervised approaches to exploit unlabeled data for multi-label learning. Song et al. proposed a method for graph data that uses label embedding and label smoothing to capture label correlation (Song et al., 2021). Shi et al. introduced a deep sequential generative model that treats labels as latent variables to reconstruct data and leverage crowdsourced labels (Shi et al., 2020). Wang et al. proposed a dual relation approach to align distributions with dual classifiers and capture correlations (Wang et al., 2020). More

recently, Xie et al. developed a class-aware pseudo-labeling method that assigns positive and negative pseudo-labels for each class based on labeled class distributions (Xie et al., 2023). In contrast, there are many works that train linear models to solve SSMLL problems (Zhan & Zhang, 2017; Zhao & Guo, 2015; Tan et al., 2017) but they are limited in their ability to scale and capture complex nonlinear relationships as the dataset size grows.

3. Preliminaries

Semi-supervised multi-label learning (SSMLL) aims to utilize unlabeled instances to enhance the predictive performance of the model, especially when labeled data is limited. Let $\mathcal{X} = \mathbb{R}^q$ be the instance space and $\mathcal{Y} = \{0, 1\}^c$ represent the label space with c classes. Given a labeled dataset with n instances $\mathcal{D}_L = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)|1 \leq i \leq n\}$ and an unlabeled dataset with m instances $\mathcal{D}_U = \{\boldsymbol{x}_i|n+1 \leq i \leq n+m\}$, where $\boldsymbol{x}_i \in \mathcal{X}$ is a q -dimensional instance and $\boldsymbol{y}_i \in \mathcal{Y}$ is its corresponding labels, SSMLL seeks to utilize the structure and distribution in $\mathcal{D}_U$ to assist the learning from $\mathcal{D}_L$. Here, $\boldsymbol{y}_i = [y_i^1, y_i^2, \ldots, y_i^c]$ where $y_i^j = 1$ indicates that the j-th label is a relevant label associated with $\boldsymbol{x}_i$ and $y_i^j = 0$ indicates that the j-th label is irrelevant to $\boldsymbol{x}_i$. The goal is to find a multi-label classifier in the hypothesis space $f \in \mathcal{F} : \mathcal{X} \mapsto \mathcal{Y}$ that minimizes the following semi-supervised classification risk:

$$\begin{aligned}\mathcal{R}_S(f) = &\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p(\boldsymbol{x},\boldsymbol{y})}\left[\mathcal{L}(f(\boldsymbol{x}), \boldsymbol{y})\right] \\ &+ \alpha\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\mathcal{L}_u(f(\boldsymbol{x}))\right],\end{aligned} \tag{1}$$

where $\mathcal{L} : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}^+$ is the multi-label loss function that measures the accuracy of the model in fitting the labeled data, $\mathcal{L}_u$ represents a regularization term that captures the structure or distribution in the unlabeled data and α is a tradeoff parameter.

In the pseudo-label-based approach for SSMLL, the regularization term $\mathcal{L}_u$ typically involves assigning pseudo-labels to the unlabeled data, which are then used to guide the training of the model. It can be mathematically expressed as:

$$\mathcal{L}_u(f(\boldsymbol{x})) = \mathcal{L}(f(\boldsymbol{x}), \boldsymbol{d}), \tag{2}$$

where $\boldsymbol{d} = [d_1, \ldots, d_c] = S(\boldsymbol{x})$ denotes the pseudo-label generated for the unlabeled instance $\boldsymbol{x}$ by the pseudo-label generator $S : \mathcal{X} \to [0, 1]^c$.

4. The PCLP Method

To fully utilize both labeled and unlabeled instances for joint learning of label correlation, we explicitly model the correlated label prior probability distribution using a Structural Causal Model (SCM) (Yu et al., 2019). This prior can be seamlessly integrated into a variational inference framework (Kingma & Welling, 2014), optimized by both labeled and unlabeled data in a unified manner. Furthermore, this prior can be employed to constrain the generated pseudo-labels to adhere to the label correlation, resulting in more accurate pseudo-labels.

4.1. The Objective Function

We consider the generative process of a sample, which is determined partly by a low-dimensional attribute feature. In a multi-label context, this can be viewed as a pseudo-label $\boldsymbol{d} = [d_1, d_2, \ldots, d_c]$, representing the probability corresponding to each category. Additionally, a low-dimensional latent feature $\boldsymbol{z} = [z_1, z_2, \ldots, z_k]$ with k dimensions is introduced to incorporate randomness into the generation process. To recover the soft pseudo-label $\boldsymbol{d}$, we incorporate the soft pseudo-label $\boldsymbol{d}$ and the low-dimensional feature space $\boldsymbol{z}$ as latent variables in a variational label enhancement framework (Xu et al., 2021a; 2020b; 2023) for inference.

The inference phase induces the conditional distribution $p_{S,E}(\boldsymbol{d}, \boldsymbol{z}|\boldsymbol{x}) = p_S(\boldsymbol{d}|\boldsymbol{x})p_E(\boldsymbol{z}|\boldsymbol{x})$ and the joint distribution $p_{S,E}(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z}) = p(\boldsymbol{x})p_{S,E}(\boldsymbol{d}, \boldsymbol{z}|\boldsymbol{x})$, where S is a pseudo-label generator as defined in section 3 and E is a low-dimensional feature encoder $E : \mathcal{X} \to \mathbb{R}^k$ to infer $\boldsymbol{z} = E(\boldsymbol{x})$. During the generation process, the generator establishes the conditional distribution $p_G(\boldsymbol{x}|\boldsymbol{d}, \boldsymbol{z})$ and the joint distribution $p_G(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z}) = p(\boldsymbol{d}, \boldsymbol{z})p_G(\boldsymbol{x}|\boldsymbol{d}, \boldsymbol{z})$, where G is a generator $G : [0, 1]^c \times \mathbb{R}^k \to \mathcal{X}$ to reconstruct samples $\boldsymbol{x} = G(\boldsymbol{d}, \boldsymbol{z})$. The objective is designed to minimize the discrepancy between the encoded and generated distributions:

$$\begin{aligned}\mathcal{L}_{\text{gen}}(S, E, G) = \mathrm{KL}\Big[&p_{S,E}(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z}) \,\|\, \\ &p_G(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z})\Big],\end{aligned} \tag{3}$$

where KL is the Kullback-Leibler (KL) divergence. The objective is shown to be equivalent to maximizing the evidence lower bound (ELBO) (Kingma & Welling, 2014):

$$\begin{aligned}\mathcal{L}_{\text{gen}}(S, E, G) = \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\Big[&\mathrm{KL}\left[p_{S,E}(\boldsymbol{d}, \boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{d}, \boldsymbol{z})\right] \\ &+ \mathbb{E}_{\boldsymbol{d},\boldsymbol{z}\sim p_{S,E}(\boldsymbol{d},\boldsymbol{z}|\boldsymbol{x})} \log p_G(\boldsymbol{x}|\boldsymbol{d}, \boldsymbol{z})\Big].\end{aligned} \tag{4}$$

The term of KL divergence can be decomposed further:

$$\begin{aligned}\mathcal{L}_{\text{gen}}(S, E, G) = \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\Big[&\mathrm{KL}\left[p_S(\boldsymbol{d}|\boldsymbol{x})\|p(\boldsymbol{d})\right] \\ &+ \mathrm{KL}\left[p_E(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z})\right] \\ &+ \mathbb{E}_{\boldsymbol{d},\boldsymbol{z}\sim p_{S,E}(\boldsymbol{d},\boldsymbol{z}|\boldsymbol{x})} \log p_G(\boldsymbol{x}|\boldsymbol{d}, \boldsymbol{z})\Big].\end{aligned} \tag{5}$$

The decomposition allows the model to distinguish the inherent attributes of the data determined by $\boldsymbol{d}$, as well as the variable features introduced by $\boldsymbol{z}$, thereby facilitating a

more precise and flexible representation and generation of data.

To further promote the label enhancement, we introduce a supervised loss term $\mathcal{L}_{\text{sup}}(S)$ to train the pseudo-label generator S on the labeled data. The overall objective function is:

$$\mathcal{L}(S, E, G) = \mathcal{L}_{\text{gen}}(S, E, G) + \lambda \mathcal{L}_{\text{sup}}(S), \tag{6}$$

where $\mathcal{L}_{\text{sup}}(S) = \sum_{j=1}^{c} y^j \log d_j + (1 - y^j) \log(1 - d_j)$ is the binary cross-entropy loss.

4.2. Correlation-Induced Label Prior

In Eq. (5), the term $\mathbb{E}_{\boldsymbol{d},\boldsymbol{z}\sim p_{S,E}(\boldsymbol{d},\boldsymbol{z}|\boldsymbol{x})} \log p_G(\boldsymbol{x}|\boldsymbol{d}, \boldsymbol{z})$ represents the expected log-likelihood of the generated samples given the latent variables. This term ensures that the generated samples are closely aligned with the actual data distribution. The other terms, involving the KL divergence, align the distribution of the latent variables with a prior probability distribution. In previous works (Yao et al., 2021; Xu et al., 2022), these prior probabilities are often assumed to be all factorized, i.e., $p(\boldsymbol{d}) = \Pi_{j=1}^{c} p(d_j)$ and $p(\boldsymbol{z}) = \Pi_{j=1}^{k} p(z_j)$, following the typical factorized Gaussian distribution. This assumption is appropriate for $\boldsymbol{z}$, which is introduced to add diversity by independent noise. However, it is inappropriate for pseudo-labels $\boldsymbol{d}$ in scenarios of MLL, where there is inherent correlation among the labels.

To address this limitation, we propose to use a causal model to capture the complex dependencies between labels, where the label correlation can be interpreted as causal relationships (Zhang & Zhang, 2010). Specifically, we employ the general nonlinear Structural Causal Model (SCM) (Yu et al., 2019) to represent the joint prior distribution over pseudo-labels. It is defined as:

$$\boldsymbol{d} = g((I - A^\top)^{-1} h(\boldsymbol{\epsilon})) := F_\beta(\boldsymbol{\epsilon}), \tag{7}$$

where A is the weighted adjacency matrix of the directed acyclic graph (DAG) upon the c classes, $\boldsymbol{\epsilon}$ is the exogenous variables following $\mathcal{N}(\mathbf{0}, I)$, g and h are element-wise nonlinear transformations, and $\beta = (g, h, A)$ is the learnable parameters of SCM denoted by F, with the parameter space $\mathcal{B}$. When g is invertible, Eq. (7) can be rewritten as:

$$g^{-1}(\boldsymbol{d}) = A^\top g^{-1}(\boldsymbol{d}) + h(\boldsymbol{\epsilon}), \tag{8}$$

where each node in the graph represents a label variable and the edges encode the causal dependencies between labels. Eq. (8) decomposes the generation of labels $\boldsymbol{d}$ into two components. The first term in the right-hand side represents the influence of label-correlation-based causality on the generation of labels, and the second term introduces randomness to account for influences from external factors, such as the influence of brightness or shooting angle from features of an image. The nonlinear transformations g and h enhance the flexibility of SCM to model complex relationships. Together, Eq. (8) enables SCM to capture the complex joint label prior distribution through structured causal dependencies between labels and random exogenous influences and allows generating pseudo-labels respecting the inter-label relationships.

The remaining step is to ascertain the dependencies among labels to construct the connections of DAG. In multi-label learning, the joint class distribution can be decomposed by the chain rule of probability (Wang et al., 2016). Specifically, the joint label prior probability $p(\boldsymbol{y})$ can be expressed as:

$$p(\boldsymbol{y}) = p(y_1) \prod_{j=2}^{c} p(y_j | y_1, \ldots, y_{j-1}). \tag{9}$$

By leveraging this probabilistic structure, we can construct the weighted adjacency matrix A, i.e., each label will have an edge connected to all its preceding labels.

Then, the label-correlation-prior is integrated into the variational label enhancement framework to guide the generation of pseudo-labels by capturing the dependencies among the multiple labels. Specifically, compared with the objective function Eq. (3), $p_G(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z})$ becomes $p_{G,F}(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z}) = p_F(\boldsymbol{d}, \boldsymbol{z}) p_G(\boldsymbol{x}|\boldsymbol{d}, \boldsymbol{z})$, where $p_F(\boldsymbol{d}, \boldsymbol{z}) = p_F(\boldsymbol{d}) p(\boldsymbol{z})$, F

Algorithm 1 The PCLP Algorithm

Input: The SSMLL training set $\mathcal{D}_L$ and $\mathcal{D}_U$, initial models $S_\psi, E_\phi, G_\theta, F_\beta, D_\eta$, the predictive model f, the number of iteration I and the number of epoch T.

for $t = 1$ **to** T **do**
  for $k = 1$ **to** I **do**
    Fetch random mini-batch $\mathcal{B} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_b\}$ from $\mathcal{D}_L \cup \mathcal{D}_U$;
    Generate $\boldsymbol{z}_i$ from $\mathcal{N}(\mathbf{0}, I)$, $1 \le i \le b$;
    Generate $\boldsymbol{d}_i$ from the causal model $\boldsymbol{d}_i = F_\beta(\boldsymbol{\epsilon}_i)$, where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, I)$, $1 \le i \le b$;
    Update η by minimizing Eq. (16) with $\mathcal{B}$;
    Fetch random mini-batch $\mathcal{B}_U = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{b_u}\}$ from $\mathcal{D}_U$ and $\mathcal{B}_L = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_{b_l}, \boldsymbol{y}_{b_l})\}$ from $\mathcal{D}_L$;
    Generate $\boldsymbol{z}_i$ and $\boldsymbol{d}_i$ for $\mathcal{B}_U$ as above;
    Update ψ by Eq. (12) with $\mathcal{B}_U$;
    Update ϕ by Eq. (13) with $\mathcal{B}_U$;
    Update θ by Eq. (14) with $\mathcal{B}_U$;
    Update β by Eq. (15) with $\mathcal{B}_U$;
    Update ψ by minimizing $\mathcal{L}_{\text{sup}}$ with $\mathcal{B}_L$.
  end for
end for
Train the predictive model f by minimizing Eq. (1) and Eq. (2) with the pseudo-labels generated by S_ψ.

Output: The predictive model f.

denotes the SCM model with learnable parameters $\beta = (g, h, A)$, and $p_F(\boldsymbol{d})$ is the distribution of $F_\beta(\boldsymbol{\epsilon})$ and $p(\boldsymbol{z})$ is a standard Gaussian $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, I)$. Then, the objective function becomes:

$$\mathcal{L}_{\text{gen}}(S, E, G, F) = \mathrm{KL}[p_{S,E}(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z}) \| p_{G,F}(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z})], \tag{10}$$

By incorporating $p_F(\boldsymbol{d})$, the pseudo-label generator S is guided to generate pseudo-labels that are consistent with the learned joint label prior distribution.

Then, the overall objective function becomes:

$$\mathcal{L}(S, E, G, F) = \mathcal{L}_{\text{gen}}(S, E, G, F) + \lambda \mathcal{L}_{\text{sup}}(S). \tag{11}$$

Next, we discuss the method employed to optimize the objective function. We represent the pseudo-label generator S, the encoder E, and the sample generator G as neural networks, parameterized by S_ψ, E_ϕ, and G_θ respectively. Aa demonstrated in (Karras et al., 2021; Mescheder et al., 2017), implicit distributions, where the randomness is fed into the input or intermediate layers of the neural network, are more flexible than explicit distributions in terms of expressiveness. We inject Gaussian noises into each convolutional layer of the generator G to generate samples. Then, the label-correlation-prior $p_F(\boldsymbol{d})$ and implicit distribution $p_G(\boldsymbol{x}|\boldsymbol{d}, \boldsymbol{z})$ make Eq. (11) lose an analytic form.

We adopt a GAN method to adversarially estimate the gradient of Eq. (11) as in the literature (Shen et al., 2020; 2022). Specifically, we introduce a discriminator D to present the adversarial gradient of the objective function. Let $D^*(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z}) = \log \frac{p_{S,E}(\boldsymbol{x},\boldsymbol{d},\boldsymbol{z})}{p_{G,F}(\boldsymbol{x},\boldsymbol{d},\boldsymbol{z})}$ be the discriminator.

Then, the gradient with respect to the pseudo-label generator S_ψ is derived:

$$\nabla_\psi \mathcal{L}_{\text{gen}} = \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\Big[\nabla_{\boldsymbol{d}} D^*(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z})^\top|_{\boldsymbol{d}=S_\phi(\boldsymbol{x})} \nabla_\psi S_\psi(\boldsymbol{x})\Big], \tag{12}$$

where the first gradient component $\nabla_{\boldsymbol{d}} D^*(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z})$ give the feedback that how distinguishable the pseudo-labels $\boldsymbol{d}$ make $p_{S,E}(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z})$ from $p_{G,F}(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z})$, the second component $\nabla_\psi S_\psi(\boldsymbol{x})$ guides the update of towards generating more accurate pseudo-labels.

For the encoder E_ϕ, the gradient is:

$$\nabla_\phi \mathcal{L}_{\text{gen}} = \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\Big[\nabla_{\boldsymbol{z}} D^*(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z})^\top|_{\boldsymbol{z}=E_\phi(\boldsymbol{x})} \nabla_\phi E_\phi(\boldsymbol{x})\Big], \tag{13}$$

which shares a structural similarity with that of S_ψ.

For the sample generator G_θ, the gradient is computed as:

$$\nabla_\theta \mathcal{L}_{\text{gen}} = -\mathbb{E}_{\boldsymbol{d},\boldsymbol{z}\sim p_{F_\beta}(\boldsymbol{d},\boldsymbol{z})}\Big[s(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z}) \nabla_{\boldsymbol{x}} D^*(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z})^\top|_{\boldsymbol{x}=G_\theta(\boldsymbol{d},\boldsymbol{z})} \nabla_\theta G_\theta(\boldsymbol{d}, \boldsymbol{z})\Big], \tag{14}$$

which is slightly different from the the previous gradients due to the inclusion of a scaling factor $s(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z}) = e^{D^\star(\boldsymbol{x},\boldsymbol{d},\boldsymbol{z})}$, acting as a weight for the gradient to enhance the stability of gradient estimation (Shen et al., 2020).

Lastly, for the SCM prior F_β, the gradient is given by:

$$\nabla_\beta \mathcal{L}_{\text{gen}} = -\mathbb{E}_{\boldsymbol{\epsilon},\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},I)} \Big[s(\boldsymbol{x},\boldsymbol{d},\boldsymbol{z})\big(\nabla_{\boldsymbol{x}} D^*(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z})^\top \nabla_\beta G\left(F_\beta(\boldsymbol{\epsilon}), \boldsymbol{z}\right) + \nabla_{\boldsymbol{d}} D^*(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z})^\top \nabla_\beta F_\beta(\boldsymbol{\epsilon})\big)|_{\boldsymbol{d}=F_\beta(\boldsymbol{\epsilon})}^{\boldsymbol{x}=G_\theta(F_\beta(\boldsymbol{\epsilon}),\boldsymbol{z})}\Big]. \tag{15}$$

Since $D^\star$ depends on the unknown densities, which makes the gradient in Eq. (12-15) intractable, we estimate hte gradients by training a discriminator D_η via the empirical logistic regression:

$$\min_{D'_\eta} \frac{1}{N_d}\Bigg[\sum_{i:w_i=1} \log\left(1 + e^{-D'_\eta(\boldsymbol{x}_i,\boldsymbol{d}_i,\boldsymbol{z}_i)}\right) + \sum_{i:w_i=0} \log\left(1 + e^{D'_\eta(\boldsymbol{x}_i,\boldsymbol{d}_i,\boldsymbol{z}_i)}\right)\Bigg], \tag{16}$$

where the weights $w_i = 1$ if $(\boldsymbol{x}_i, \boldsymbol{d}_i, \boldsymbol{z}_i) \sim p_{S,E}(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z})$ and $w_i = 0$ if $(\boldsymbol{x}_i, \boldsymbol{d}_i, \boldsymbol{z}_i) \sim p_{G,F}(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z})$ with $i = 1, \ldots, N_d$, N_d is the total number of sampled instances for training the discriminator.

The overall process of PCLP is shown in Algorithm 1.

5. Theoretical Analysis

In this section, we delve into the effectiveness of using the SCM prior to guide the label generator in producing pseudo-labels. Before delving into the analysis, we first establish a definition of what constitutes a high-quality pseudo-label.

Definition 5.1. Bayesian-informed pseudo-label generator: Given the ground-truth label $\boldsymbol{y}$ of instance $\boldsymbol{x}$, a pseudo-label generator S is said to be a bayesian-informed pseudo-label generator with respect to $\boldsymbol{y}$ if $\forall j = 1 \ldots c$, such that $d_j := [S(\boldsymbol{x})]_j = p(y^j|\boldsymbol{x})$.

Definition 5.1 indicates that a high-quality pseudo-label should closely match the Bayesian posterior probabilities for each class.

Following the given definition, we demonstrate that if the correlations within the joint label prior probability distribution are neglected, the learned label generator will not qualify as a high-quality pseudo-label generator as defined.

Theorem 5.2. *Let $S^\star$ be any bayesian-informed pseudo-label generator with respect to $\boldsymbol{y}$. Let $b^\star = \mathcal{L}_{sup}(S^\star)$, $a = \min_{E,G} \mathcal{L}_{gen}(S^\star, E, G)$, and $b = \min_{\{(S,E,G):\mathcal{L}_{gen}=0\}} \mathcal{L}_{sup}(S)$, where E is any encoder and G is any samples generator. Assume that the joint label prior probability $p(\boldsymbol{d})$ is factorized, i.e., $p(\boldsymbol{d}) = \prod_{j=1}^{c} p(d_j)$. Then we have $a > 0$, and either when $b^\star \geq b$ or $b^\star < b$ and $\lambda < \frac{a}{b-b^\star}$, there exists a solution (S', E', G') so that S' is not a Bayesian-informed pseudo-label generator, and for any E and G, we have $\mathcal{L}(S', E', G') < \mathcal{L}(S^\star, E, G)$.*

The proof can be found in Appendix A.1. Theorem 5.2 reveals that if the joint label prior probability is assumed to be factorized, then it is possible to find a pseudo-label generator S' that does not satisfy Definition 5.1 yet achieves a lower loss than a Bayesian-informed pseudo-label generator $S^\star$, when λ is not large enough. However, in real-world applications, when λ is too large, the model is prone to overfitting on the labeled dataset. Therefore, if the joint label prior probability is factorized, setting this parameter becomes particularly challenging.

Theorem 5.3. *Assume that the underlying distribution $p(\boldsymbol{y})$ belongs to the distribution family $\{p_\beta : \beta \in \mathcal{B}\}$, where $\mathcal{B}$ is the parameter space of β, i.e., there exists $\beta_0 = (g_0, h_0, A_0)$ such that $p(\boldsymbol{y}) = p_{\beta_0}$. And suppose the infinite capacity of S, E and G. Let $(S^\star, E^\star, G^\star, F^\star) \in \arg\min_{S,E,G,F} \mathcal{L}(S, E, G, F)$ be the optimal solution. Then $S^\star$ is a bayesian-informed pseudo-label generator with respect to $\boldsymbol{y}$.*

The proof can be found in Appendix A.2. Theorem 5.3 emphasizes that when the correlations within the joint label prior probability are taken into account, a high-quality pseudo-label generator will be learned.

Additionally, we establish an estimation error bound for Eq. (1) to demonstrate its learning consistency (Mohri et al., 2012). Our goal is to minimize the expective risk:

$$\mathcal{R}(f) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p(\boldsymbol{x},\boldsymbol{y})} \mathcal{L}(f(\boldsymbol{x}), \boldsymbol{y}). \tag{17}$$

However, we have no access to sufficient labeled data to estimate the true distribution $p(\boldsymbol{x}, \boldsymbol{y})$, we can only estimate the risk by minimizing the empirical risk $\widehat{\mathcal{R}}_S(f) = \widehat{\mathcal{R}}_l(f) + \widehat{\mathcal{R}}_u(f)$ on the training set $\mathcal{D}_L$ and $\mathcal{D}_U$, where $\widehat{\mathcal{R}}_l(f)$ and $\widehat{\mathcal{R}}_u(f)$ are the empirical risk of the labeled loss $\mathcal{L}(f(\boldsymbol{x}), \boldsymbol{y})$ and the unlabeled loss $\mathcal{L}_u(f(\boldsymbol{x}))$ in Eq. (1) respectively:

$$\begin{aligned} \widehat{\mathcal{R}}_l(f) &= \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f(\boldsymbol{x}_i), \boldsymbol{y}_i), \\ \widehat{\mathcal{R}}_u(f) &= \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}_u(f(\boldsymbol{x}_i)), \end{aligned} \tag{18}$$

where $\mathcal{L}_u(f(\boldsymbol{x}_i)) = \mathcal{L}(f(\boldsymbol{x}_i), \boldsymbol{d}_i), \boldsymbol{d}_i = S(\boldsymbol{x}_i)$. Let $\hat{f} = \min_{f\in\mathcal{F}} \widehat{\mathcal{R}}_S(f)$ be the empirical risk minimizer and $f^\star = \min_{f\in\mathcal{F}} \mathcal{R}(f)$ be the true risk minimizer. Besides, we define the function space H_y for the label $y \in 1, \ldots, c$ as $\mathcal{H}_y = \{h : \boldsymbol{x} \to f_y(\boldsymbol{x}) | f \in \mathcal{F}\}$. Let $\mathfrak{R}_{n+m}(\mathcal{H}_y)$ be the expected Rademacher complexity (Mohri et al., 2012), then we have the following theorem:

Theorem 5.4. *Assume the loss function $\mathcal{L}$ and $\mathcal{L}_u$ is ρ-Lipschitz with respect to $f(\boldsymbol{x})$ and upper bounded by M. Then, for any $\delta > 0$, with probability at least $1 - \delta$:*

$$\begin{aligned} \mathcal{R}(\hat{f}) - \mathcal{R}(f^\star) \leq 4\sqrt{2}\rho \sum_{y=1}^{c} \mathfrak{R}_{n+m}(\mathcal{H}_y) \\ + M\sqrt{\frac{\log 2/\delta}{2n}} + M\sqrt{\frac{\log 2/\delta}{2m}}. \end{aligned} \tag{19}$$

The proof can be found in Appendix A.3. The term involving the Rademacher complexity captures the capacity of the hypothesis space. Meanwhile, the second term provides a convergence rate that diminishes as the total number of instances increases. This theorem assures that with a proper choice of hypothesis space and enough training data, the empirical risk minimizer $\hat{f}$ will converge to the true risk minimizer $f^\star$ as $n, m \to \infty$.

6. Experiments

6.1. Experimental Configurations

Datasets We evaluate the effectiveness of the proposed method on three large-scale multi-label image classification datasets, including `Pascal-VOC-2012 (VOC)` (Everingham et al., 2010), `MS-COCO-2014 (COCO)` (Lin et al., 2014), and `NUS-WIDE (NUS)` (Chua et al., 2009). Detailed information about these datasets is provided in the appendix. During the training process, a proportion $p \in \{0.05, 0.1, 0.15, 0.2\}$ of samples with complete labels is selected at random, while the rest of the samples are without any supervisory information. Performance is evaluated using Mean Average Precision (mAP).

Baselines Following the experimental setting in previous SSMLL literature (Xie et al., 2023), we compared the proposed method against five groups of approaches to validate the effectiveness:

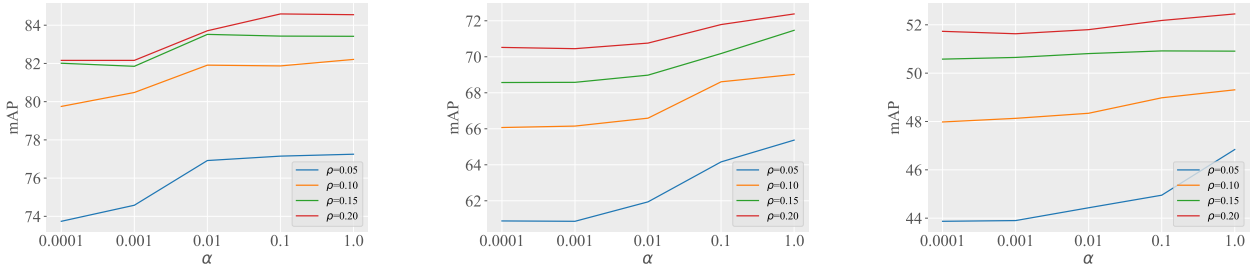
1) Two baseline methods: BCE and ASL (Ridnik et al., 2021) that only use the labeled data for training.

2) Three instance-based pseudo-labeling methods:

   - IAT-1 (Xie et al., 2023) selects the most probable label in model prediction as pseudo-labels one for unlabeled training instance.
   - IAT-K (Xie et al., 2023) selects the top k probable labels as pseudo-labels for each instance.

Table 1: Mean average precision (mAP) of each comparing method on VOC, COCO and NUS. The best performance is highlighted in bold.

Methods	VOC				COCO				NUS			
	$p = 0.05$	$p = 0.10$	$p = 0.15$	$p = 0.20$	$p = 0.05$	$p = 0.10$	$p = 0.15$	$p = 0.20$	$p = 0.05$	$p = 0.10$	$p = 0.15$	$p = 0.20$
BCE	65.40	75.48	77.87	79.00	57.09	62.34	65.55	67.31	40.12	45.04	47.04	48.29
ASL	71.41	77.81	79.12	79.84	57.87	62.95	65.73	67.43	42.04	46.07	48.04	49.55
IAT-1	75.34	80.80	82.93	83.67	59.42	63.52	65.13	66.88	39.27	46.05	47.02	46.69
IAT-K	73.62	80.20	82.17	83.03	59.83	64.02	65.21	67.45	39.18	46.15	46.98	46.70
IAT	71.88	80.18	82.87	83.99	60.76	65.60	65.31	69.29	40.10	46.45	47.39	47.15
LL-R	73.58	79.68	81.79	82.73	60.97	65.25	67.48	68.75	44.12	47.10	48.93	49.54
LL-CT	71.13	79.03	81.43	82.55	58.82	63.31	64.55	67.19	39.24	45.93	48.09	49.89
LL-CP	71.41	79.63	82.25	83.25	57.90	63.93	65.67	68.07	40.72	46.60	48.46	49.61
PLC	74.57	80.78	81.99	83.05	58.65	65.07	67.66	69.09	44.99	48.70	49.99	51.30
ADSH	75.40	80.36	82.76	83.97	60.71	65.36	67.69	69.05	43.95	47.27	49.19	49.99
FREEMATCH	75.07	80.68	82.57	83.62	59.95	64.43	66.76	68.04	43.05	46.61	48.68	49.56
DRML	61.75	70.97	72.97	74.44	53.59	57.02	58.62	59.18	30.57	35.03	37.93	40.01
CAP	75.90	81.83	83.10	84.32	62.88	67.18	68.99	70.43	44.98	47.81	49.04	51.37
PCLP	**77.25**	**82.21**	**83.72**	**84.59**	**64.43**	**69.02**	**70.86**	**71.52**	**46.39**	**48.83**	**50.57**	**52.45**



(a) Sensitivity analysis of α on VOC. (b) Sensitivity analysis of α on COCO. (c) Sensitivity analysis of α on NUS.

Figure 1: Parameter sensitivity analysis of α ranging from $\{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ on VOC, COCO and NUS.

- IAT (Xie et al., 2023) adopts an instance-aware threshold to assign pseudo-labels.

3) Two state-of-the-art multi-label learning with missing labels (MLML) methods:

- LL (Kim et al., 2022) treats unobserved labels as noisy labels and dynamically adjusts the threshold to reject or correct samples with a large loss, in order to prevent the model from memorizing the noisy labels. including three variants LL-R, LL-CT and LL-CP.
- PLC (Xie et al., 2022) designs a label-aware global consistency regularization to recover the pseudo-labels leveraging the manifold structure information learned by contrastive learning.

4) Two state-of-the-art semi-supervised learning methods:

- ADSH (Guo & Li, 2022) involves adaptive thresholding for different classes and optimize the number of pseudo-labels for each class.
- FREEMATCH (Wang et al., 2023) introduce a self-adaptive class fairness regularization penalty to encourage the model for diverse predictions and adjust the confidence threshold in a self-adaptive manner according to the model's learning status to assign pseudo-labels.

5) Two state-of-the-art deep SSMLL methods:

- DRML (Wang et al., 2020) introduces a dual-classifier framework to align the feature distribution in a latent space while generating pseudo-labels for unlabeled instances.
- CAP (Xie et al., 2023) designs a class-specific threshold strategy based on the estimated class prior probabilities to assign pseudo-labels.

Implementation In order to ensure the fairness of comparison, for all compared method, We use ResNet-50 network (He et al., 2016) pre-trained on ImageNet (Russakovsky et al., 2015) for the predictive model f. The architecture of the pseudo-label generator S and the encoder E is a ResNet-50 followed by a 4-layers MLP and SAGAN architecture (Zhang et al., 2019) is used for the discriminator D and the

Table 2: Ablation Study Results.

Dataset	p	VI	SCM	VI	SCM	VI	SCM
		×	×	×	✓	✓	✓
VOC	0.05	75.18	76.38	**77.25**			
	0.10	80.41	81.45	**82.21**			
	0.15	81.53	82.57	**83.72**			
	0.20	83.11	83.84	**84.59**			
COCO	0.05	62.17	63.38	**64.43**			
	0.10	66.81	67.92	**69.02**			
	0.15	68.78	69.87	**70.86**			
	0.20	69.47	70.57	**71.52**			
NUS	0.05	44.61	45.44	**46.39**			
	0.10	46.71	47.72	**48.83**			
	0.15	48.43	49.31	**50.57**			
	0.20	50.37	51.30	**52.45**			

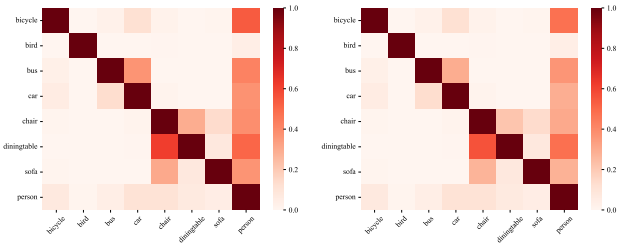
sample generator G. For the tradeoff parameter λ, we fix it as 1 for all datasets. We use Adam optimizer (Loshchilov & Hutter, 2019) with $\beta = (0.9, 0.999)$, RandAugment (Cubuk et al., 2020) and Cutout (DeVries & Taylor, 2017) for data augmentation for all datasets in all experiments. The batch size is 64, the learning rate is 10^{-4}. We implement all experiments by PyTorch on NVIDIA RTX 3090 GPUs.

6.2. Experimental Results

Table 1 summarizes the performance of PCLP and the compared methods in terms of mAP on VOC, COCO and NUS. From the results, we can observe that PCLP achieves state-of-the-art performance on all benchmark datasets. DRML, as an SSMLL approach, does not yield satisfactory results, possibly due to its reliance on a limited number of labeled instances to construct label correlation. This imprecise correlation might mislead the model's predictions, resulting in suboptimal performance. CAP achieves the second-best results in the comparison, proving that the information of the class prior is of great help in solving the SSMLL. However, CAP only models the class priors separately, ignoring the dependencies between class priors, which is the key to why our PCLP can achieve better performance. These comparison results convincingly confirm that by mining the dependencies between class prior dependencies, our method can generate higher quality pseudo-labels to solve the SSMLL problem.

6.3. Sensitivity Analysis

Our algorithm incorporates a tradeoff parameter α in the SSMLL risk in Eq. (1), for which we have conducted a parameter sensitivity analysis. We chose the value of this hyper-parameter from the set $\{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. Figure 1 reports the impact of this weight on the results



(a) Case with full labels. (b) Case with $p = 0.05$.

Figure 2: Visualization of label co-occurrence matrix on dataset VOC.

of the datasets VOC, COCO and NUS with the labeled proportions $p \in \{0.05, 0.1, 0.15, 0.2\}$. Experimental results demonstrate that assigning a larger weight to the regularization term $\mathcal{L}_u$ significantly enhances model performance. This improvement underscores the high quality of the pseudo-labels generated by PCLP, effectively contributing to the advancement of model performance.

6.4. Ablation Study

In this section, we conduct an ablation study to investigate the significance of leveraging the joint label prior distribution learned by SCM in guiding the pseudo-labeling process. To assess the impact of the SCM component, we removed it from the model, relying solely on the original variational inference method (VI) to generate pseudo-labels for training. Furthermore, we experimented with using the raw outputs of the network as pseudo-labels for the unlabeled data. The first column indicates results using only the network's output for pseudo-label generation. The second column shows results when only variational inference is employed for generating pseudo-labels. The third column represents the implementation of a SCM to construct the correlated label prior probability distribution for pseudo-labeling. The results demonstrate the effectiveness of our approach, which integrates the joint label prior distribution into the label enhancement to guide the generation of pseudo-labels. The ablation study confirms that this integration is crucial for aligning the generated pseudo-labels with the learned label prior distribution and significantly contributes to the performance of the predictive model.

6.5. Visualization of Label Correlation

Figure 2a displays the statistical label-occurrence matrix obtained under the condition of full labeling. Figure 2b depicts the label-occurrence matrix obtained from the output of the predictive model on the dataset VOC with $p = 0.05$. A deeper color indicates a higher co-occurrence rate. By comparison, it is evident that our method can learn a label-occurrence matrix similar to that of the fully labeled sce-

nario, even with a limited number of labeled samples, such as between 'bus' and 'car', 'dining table' and 'chair', 'sofa' and 'chair', as well as 'bicycle' and 'person'. These observations demonstrate that our method can effectively capture the correlations between labels and generate high-quality pseudo-labels.

7. Conclusion

In this paper, we study semi-supervised multi-label learning (SSMLL) and propose a novel method called PCLP, which effectively employs both labeled and unlabeled data through a Structural Causal Model integrated with a variational label enhancement process, guaranteeing that the generated pseudo-labels are aligned with the learned correlated prior. The theoretical analysis provided confirms that by considering label correlation in the label prior distribution, our method achieves more precise pseudo-labeling. Additionally, the generalization error bound guarantees the learning consistency. Extensive experiments on three large-scale multi-label image classification datasets demonstrate that PCLP achieves state-of-the-art performance.

Acknowledgments

This research was supported by the National Science Foundation of China (62206050, 62125602, and 62076063), China Postdoctoral Science Foundation (2021M700023), Jiangsu Province Science Foundation for Youths (BK20210220), Young Elite Scientists Sponsorship Program of Jiangsu Association for Science and Technology (TJ-2022-078), the Big Data Computing Center of Southeast University, the Fundamental Research Funds for the Central Universities.

Impact Statement

This research aims to further the methodologies and technologies in the field of Machine Learning. It employs training data that may raise data privacy concerns and exhibit potential biases inherent in the data. We acknowledge the significance of these issues and emphasize the importance of ethical data handling practices and the development of unbiased algorithms to mitigate societal consequences.

References

Berthelot, D., Carlini, N., Cubuk, E. D., Kurakin, A., Sohn, K., Zhang, H., and Raffel, C. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *International Conference on Learning Representations*, Addis Ababa, Ethiopia, 2020.

Boutell, M. R., Luo, J., Shen, X., and Brown, C. M. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

Chen, B., Jiang, J., Wang, X., Wan, P., Wang, J., and Long, M. Debiased self-training for semi-supervised learning. *Advances in Neural Information Processing Systems*, 35: 32424–32437, 2022.

Chen, J.-Y., Li, S.-Y., Huang, S.-J., Chen, S., Wang, L., and Xie, M.-K. Unm: A universal approach for noisy multi-label learning. *IEEE Transactions on Knowledge and Data Engineering*, pp. in press, 2024.

Chen, S., Wang, J., Chen, Y., Shi, Z., Geng, X., and Rui, Y. Label distribution learning on auxiliary label space graphs for facial expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13981–13990, Seattle, WA, 2020.

Chen, Z., Wei, X., Wang, P., and Guo, Y. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5177–5186, Long Beach, CA, 2019.

Chua, T., Tang, J., Hong, R., Li, H., Luo, Z., and Zheng, Y. NUS-WIDE: a real-world web image database from national university of singapore. In *Proceedings of the 8th ACM International Conference on Image and Video Retrieval*, Santorini Island, Greece, 2009.

Cole, E., Aodha, O. M., Lorieul, T., Perona, P., Morris, D., and Jojic, N. Multi-label learning from single positive labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 933–942, virtual, 2021.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops*, pp. 702–703, Seattle, WA, 2020.

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Elisseeff, A. and Weston, J. A kernel method for multi-labelled classification. In *Advances in Neural Information Processing Systems*, pp. 681–687, Vancouver, BC, Canada, 2001.

Everingham, M., Gool, L. V., Williams, C. K. I., Winn, J. M., and Zisserman, A. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

French, G., Mackiewicz, M., and Fisher, M. Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations*, Vancouver, BC, Canada, 2018.

Fürnkranz, J., Hüllermeier, E., Loza Mencía, E., and Brinker, K. Multilabel classification via calibrated label ranking. *Machine Learning*, 73(2):133–153, 2008.

Guo, L.-Z. and Li, Y.-F. Class-imbalanced semi-supervised learning with adaptive thresholding. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 8082–8094, Virtual, 2022.

Hang, J. and Zhang, M. Collaborative learning of label semantics and deep label-specific features for multi-label classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9860–9871, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, 2016.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4217–4228, 2021.

Kim, Y., Kim, J. M., Akata, Z., and Lee, J. Large loss matters in weakly supervised multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14156–14165, New Orleans, LA, 2022.

Kim, Y., Kim, J. M., Jeong, J., Schmid, C., Akata, Z., and Lee, J. Bridging the gap between model explanations in partially annotated multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3408–3417, Vancouver, BC, Canada, 2023.

Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. In *Proceedings of the 5th International Conference on Learning Representations*, Banff, AB, Canada, 2014.

Laine, S. and Aila, T. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, Toulon, France, 2017.

Lanchantin, J., Wang, T., Ordonez, V., and Qi, Y. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16478–16488, Virtual, 2021.

Li, S., Xia, X., Zhang, H., Zhan, Y., Ge, S., and Liu, T. Estimating noise transition matrix with label correlations for noisy multi-label learning. *Advances in Neural Information Processing Systems*, 35:24184–24198, 2022.

Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: common objects in context. In *Proceedings of 13th European Conference on Computer Vision*, volume 8693, pp. 740–755, Zurich, Switzerland, 2014.

Liu, B., Xu, N., Lv, J., and Geng, X. Revisiting pseudo-label for single-positive multi-label learning. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 22249–22265, Honolulu, HI, 2023.

Liu, J., Chang, W., Wu, Y., and Yang, Y. Deep learning for extreme multi-label text classification. In *Proceedings of the 40-th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 115–124, Tokyo, Japan, 2017.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, New Orleans, LA, 2019.

Mescheder, L., Nowozin, S., and Geiger, A. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 2391–2400. PMLR, 2017.

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (8):1979–1993, 2018.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. Adaptive computation and machine learning. MIT Press, 2012. ISBN 978-0-262-01825-8.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Read, J., Pfahringer, B., Holmes, G., and Frank, E. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359, 2011.

Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., and Zelnik-Manor, L. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 82–91, Virtual, 2021.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein,

M., et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115: 211–252, 2015.

Sajjadi, M., Javanmardi, M., and Tasdizen, T. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Shen, X., Zhang, T., and Chen, K. Bidirectional generative modeling using adversarial gradient estimation. *arXiv preprint arXiv:2002.09161*, 2020.

Shen, X., Liu, F., Dong, H., Lian, Q., Chen, Z., and Zhang, T. Weakly supervised disentangled generative causal representation learning. *The Journal of Machine Learning Research*, 23(1):10994–11048, 2022.

Shi, W., Sheng, V. S., Li, X., and Gu, B. Semi-supervised multi-label learning from crowds via deep sequential generative model. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1141–1149, New York, 2020.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C.-L. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020.

Song, Z., Meng, Z., Zhang, Y., and King, I. Semi-supervised multi-label learning for graph-structured data. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, pp. 1723–1733, New York, 2021. Association for Computing Machinery.

Tan, Q., Yu, Y., Yu, G., and Wang, J. Semi-supervised multi-label classification using incomplete label information. *Neurocomputing*, 260:192–202, 2017.

Wang, H., Liu, W., Zhao, Y., Zhang, C., Hu, T., and Chen, G. Discriminative and correlative partial multi-label learning. In *Proceedings of International Joint Conference on Artificial Intelligence*, pp. 3691–3697, Macao, China, 2019.

Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., and Xu, W. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2285–2294, Las Vegas, NV, 2016.

Wang, L., Liu, Y., Qin, C., Sun, G., and Fu, Y. Dual relation semi-supervised multi-label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 6227–6234, New York, 2020.

Wang, Y., Chen, H., Heng, Q., Hou, W., Fan, Y., Wu, Z., Wang, J., Savvides, M., Shinozaki, T., Raj, B., Schiele, B., and Xie, X. Freematch: Self-adaptive thresholding for semi-supervised learning. In *International Conference on Learning Representations*, Kigali, Rwanda, 2023.

Xie, M.-K. and Huang, S.-J. Partial multi-label learning with noisy label identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3676–3687, 2021.

Xie, M.-K., Xiao, J., and Huang, S.-J. Label-aware global consistency for multi-label learning with single positive labels. *Advances in Neural Information Processing Systems*, 35:18430–18441, 2022.

Xie, M.-K., Xiao, J.-H., Niu, G., Sugiyama, M., and Huang, S.-J. Class-distribution-aware pseudo labeling for semi-supervised multi-label learning. In *Advances in Neural Information Processing Systems*, New Orleans, LA, 2023.

Xie, Q., Dai, Z., Hovy, E., Luong, T., and Le, Q. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33: 6256–6268, 2020.

Xu, N., Liu, Y.-P., and Geng, X. Partial multi-label learning with label distribution. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6510–6517, New York, 2020a.

Xu, N., Shu, J., Liu, Y.-P., and Geng, X. Variational label enhancement. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 10597–10606, Virtual, 2020b.

Xu, N., Liu, Y.-P., and Geng, X. Label enhancement for label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1632–1643, 2021a.

Xu, N., Qiao, C., Lv, J., Geng, X., and Zhang, M.-L. One positive label is sufficient: Single-positive multi-label learning with label enhancement. In *Advances in Neural Information Processing Systems*, pp. 21765–21776, New Orleans, LA, 2022.

Xu, N., Shu, J., Zheng, R., Geng, X., Meng, D., and Zhang, M.-L. Variational label enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (5):6537–6551, 2023.

Xu, Y., Shang, L., Ye, J., Qian, Q., Li, Y.-F., Sun, B., Li, H., and Jin, R. Dash: Semi-supervised learning with dynamic thresholding. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 11525–11536, Virtual, 2021b.

Yang, F., Wu, K., Zhang, S., Jiang, G., Liu, Y., Zheng, F., Zhang, W., Wang, C., and Zeng, L. Class-aware contrastive semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14421–14430, New Orleans, LA, 2022.

Yao, Y., Liu, T., Gong, M., Han, B., Niu, G., and Zhang, K. Instance-dependent label-noise learning under a structural causal model. *Advances in Neural Information Processing Systems*, 34:4409–4420, 2021.

Yazici, V. O., Gonzalez-Garcia, A., Ramisa, A., Twardowski, B., and Weijer, J. v. d. Orderless recurrent models for multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13440–13449, Seattle, WA, 2020.

Yu, Y., Chen, J., Gao, T., and Yu, M. Dag-gnn: Dag structure learning with graph neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 7154–7163, Long Beach, CA, 2019.

Yu, Z. and Zhang, M. Multi-label classification with label-specific feature generation: A wrapped approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5199–5210, 2022.

Zhan, W. and Zhang, M.-L. Inductive semi-supervised multi-label learning with co-training. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1305–1314, Halifax, NS, Canada, 2017.

Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., and Shinozaki, T. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34: 18408–18419, 2021.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. Self-attention generative adversarial networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 7354–7363, Los Angeles, CA, 2019.

Zhang, M.-L. and Zhang, K. Multi-label learning by exploiting label dependency. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 999–1008, Washington, DC, 2010.

Zhang, M.-L. and Zhou, Z.-H. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2013.

Zhao, F. and Guo, Y. Semi-supervised multi-label learning with incomplete labels. In *Proceedings of International Joint Conference on Artificial Intelligence*, pp. 4062–4068, Buenos Aires, Argentina, 2015.

Zheng, M., You, S., Huang, L., Wang, F., Qian, C., and Xu, C. Simmatch: Semi-supervised learning with similarity matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14471–14481, New Orleans, LA, 2022.

A. Proofs

A.1. Proof of Theorem 5.2

Proof. Due to the correlation between labels in the label space, there are at least two labels in the label space that are not independent. This implies that the probability density of $\boldsymbol{y}$ cannot be factorized. Since the pseudo-label generator $S^\star$ aligns with the true labels as defined in Definition 5.1, for all $j = 1, \ldots, c$, there exists a σ such that $S^\star(\boldsymbol{x})_j = \sigma(y^j)$. This means that the probability density of $S^\star(\boldsymbol{x})$ is not factorized.

Note that the family of ground-turth label prior distribution is contained in the factorized distribution family, i.e., $\{p(\boldsymbol{y}) : p(\boldsymbol{y}) = \prod_{j=1}^{c} p(y_j)\}$. Therefore, the intersection of the marginal distribution families of $\boldsymbol{y}$ and $S^\star(\boldsymbol{x})$ is empty. Then, the joint distribution families of $(\boldsymbol{x}, S^\star(\boldsymbol{x}), E(\boldsymbol{x}))$ and $(G(\boldsymbol{d}, \boldsymbol{z}), \boldsymbol{d}, \boldsymbol{z})$ also have an empty intersection. $\mathcal{L}_{\text{gen}}(S^\star, E, G) = 0$ implies that $p_{S,E}(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z}) = p_G(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z})$, which contradicts the above. Hence, $a = \min_{E,G} \mathcal{L}_{\text{gen}}(S^\star, E, G) > 0$.

Let (S', E', G') be the solution of the optimization problem $\min_{\{(S,E,G):\mathcal{L}_{\text{gen}}=0\}} \mathcal{L}_{\text{sup}}(S)$. From the above analysis, S' cannot be a ground-truth aligned pseudo-label generator with respect to $\boldsymbol{y}$. Therefore, $\mathcal{L}' = \mathcal{L}(S', E', G') = \lambda b$, and $\mathcal{L}^\star = \mathcal{L}(S^\star, E, G) \geq a + \lambda b^\star > \lambda b^\star$ for any E and G. When $b^\star \geq b$ we directly have $\mathcal{L}' < \mathcal{L}^\star$. When $b^\star < b$ and λ is not large enough, i.e., $\lambda < \frac{a}{b - b^\star}$, we have $\mathcal{L}' < \mathcal{L}^\star$. □

A.2. Proof of Theorem 5.3

Proof. For each $j = 1, \ldots, c$, we consider the binary cross-entropy loss for each label y_j:

$$\begin{aligned}\mathcal{L}_{\text{sup},j} &= \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p(\boldsymbol{x},\boldsymbol{y})}\left[-y^j \log S(\boldsymbol{x})_j - (1 - y^j)\log(1 - S(\boldsymbol{x})_j)\right] \\ &= -\int p(\boldsymbol{x})p(y^j|\boldsymbol{x})\left[y^j \log S(\boldsymbol{x})_j + (1 - y^j)\log(1 - S(\boldsymbol{x})_j)\right] d\boldsymbol{x}dy^j.\end{aligned} \tag{20}$$

Let:

$$\frac{\partial \mathcal{L}_{\text{sup},j}}{\partial S(\boldsymbol{x})_j} = -\int p(\boldsymbol{x})p(y^j|\boldsymbol{x})\left(\frac{y^j}{S(\boldsymbol{x})_j} - \frac{1 - y^j}{1 - S(\boldsymbol{x})_j}\right) d\boldsymbol{x}dy^j = 0. \tag{21}$$

Then we have that $S(\boldsymbol{x})_j = p(y^j|\boldsymbol{x})$ minimizes $\mathcal{L}_{\text{sup},j}$.

By the assumption that there exists $\beta_0 = (g_0, h_0, A_0)$ such that $p(\boldsymbol{y}) = p_{\beta_0}$ and the infinite capacity of G, we have the distribution family of $p_{G,F}(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z})$ contains $p_{S^\star,E^\star}(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z})$. Then by minimizing Eq. (11) over G, we can find $G^\star$ such that $p_{G^\star,F^\star}(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z}) = p_{S^\star,E^\star}(\boldsymbol{x}, \boldsymbol{d}, \boldsymbol{z})$, where $F^\star$ corresponds to $\beta^\star = (g^\star, h^\star, A^\star)$.

Hence, the optimal solution $S^\star$ of Eq. (11) is a Bayesian-informed pseudo-label generator. □

A.3. Proof of Theorem 5.4

Proof. Firstly, we define the function space:

$$\mathcal{G}_l = \left\{g : (\boldsymbol{x}, \boldsymbol{y}) \mapsto \mathcal{L}(f(\boldsymbol{x}), \boldsymbol{y}) | f \in \mathcal{F}\right\}, \mathcal{G}_u = \left\{g : (\boldsymbol{x}, \boldsymbol{d}) \mapsto \mathcal{L}(f(\boldsymbol{x}), \boldsymbol{d}) | f \in \mathcal{F}\right\},$$

and define the denote the expected Rademacher complexity (Mohri et al., 2012) of the function space :

$$\widetilde{\mathfrak{R}}_n(\mathcal{G}_l) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}_l} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(\boldsymbol{x}_i, \boldsymbol{y}_i)\right], \widetilde{\mathfrak{R}}_m(\mathcal{G}_u) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}_u} \frac{1}{m}\sum_{i=1}^{m} \sigma_i g(\boldsymbol{x}_i, \boldsymbol{d}_i)\right],$$

where $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \cdots, \sigma_n\}$ is n Rademacher variables with σ_i independently uniform variable taking value in $\{+1, -1\}$. Then we have:

Lemma A.1. *We suppose that the loss function $\mathcal{L}$ and $\mathcal{L}_u$ could be bounded by M, and for any $\delta > 0$, with probability at least $1 - \delta$, we have:*

$$\sup_{f\in\mathcal{F}} |\mathcal{R}_l(f) - \hat{\mathcal{R}}_l(f)| \leq 2\widetilde{\mathfrak{R}}_n(\mathcal{G}_l) + \frac{M}{2}\sqrt{\frac{\log\frac{2}{\delta}}{2n}},$$

$$\sup_{f\in\mathcal{F}} |\mathcal{R}_u(f) - \hat{\mathcal{R}}_u(f)| \leq 2\widetilde{\mathfrak{R}}_m(\mathcal{G}_u) + \frac{M}{2}\sqrt{\frac{\log\frac{2}{\delta}}{2m}},$$

where $\mathcal{R}_l(f) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p(\boldsymbol{x},\boldsymbol{y})}\left[\mathcal{L}(f(\boldsymbol{x}),\boldsymbol{y})\right]$, $\mathcal{R}_u(f) = \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\mathcal{L}(f(\boldsymbol{x}),\boldsymbol{d})\right]$ and $\boldsymbol{d} = S(\boldsymbol{x})$.

Proof. Suppose an example $(\boldsymbol{x},\boldsymbol{y})$ is replaced by another arbitrary example $(\boldsymbol{x}',\boldsymbol{y}')$, then the change of $\sup_{f\in\mathcal{F}}\mathcal{R}_l(f) - \hat{\mathcal{R}}_l(f)$ is no greater than $\frac{M}{2n}$. By applying McDiarmid's inequality, for any $\delta > 0$, with probility at least $1-\frac{\delta}{2}$,

$$\sup_{f\in\mathcal{F}}\mathcal{R}_l(f) - \hat{\mathcal{R}}_l(f) \leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}\mathcal{R}_l(f) - \hat{\mathcal{R}}_l(f)\right] + \frac{M}{2}\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

By symmetry, we can obtain

$$\sup_{f\in\mathcal{F}}|\mathcal{R}_l(f) - \hat{\mathcal{R}}_l(f)| \leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}\mathcal{R}_l(f) - \hat{\mathcal{R}}_l(f)\right] + \frac{M}{2}\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

Next is to bound the term $\mathbb{E}\left[\sup_{f\in\mathcal{F}}\mathcal{R}_l(f) - \hat{\mathcal{R}}_l(f)\right]$:

$$\begin{aligned}
&\mathbb{E}\left[\sup_{f\in\mathcal{F}}\mathcal{R}_l(f) - \hat{\mathcal{R}}_l(f)\right] = \mathbb{E}_{\mathcal{D}_L}\left[\sup_{f\in\mathcal{F}}\mathcal{R}_l(f) - \hat{\mathcal{R}}_l(f)\right] \\
&= \mathbb{E}_{\mathcal{D}_L}\left[\sup_{f\in\mathcal{F}}\mathbb{E}_{\mathcal{D}'_L}\left[\hat{\mathcal{R}}'_l(f) - \hat{\mathcal{R}}_l(f)\right]\right] \\
&\leq \mathbb{E}_{\mathcal{D}_L,\mathcal{D}'_L}\left[\sup_{f\in\mathcal{F}}\left[\hat{\mathcal{R}}'_l(f) - \hat{\mathcal{R}}_l(f)\right]\right] \\
&= \mathbb{E}_{\mathcal{D}_L,\mathcal{D}'_L,\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sigma_i\left(\hat{\mathcal{R}}'_l(f) - \hat{\mathcal{R}}_l(f)\right)\right] \\
&\leq \mathbb{E}_{\mathcal{D}'_L,\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sigma_i\left(\hat{\mathcal{R}}'_l(f)\right)\right] + \mathbb{E}_{\mathcal{D}_L,\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sigma_i\left(\hat{\mathcal{R}}_l(f)\right)\right] \\
&= 2\mathbb{E}_{\mathcal{D}_L,\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sigma_i\left(\hat{\mathcal{R}}_l(f)\right)\right] \\
&= 2\widetilde{\mathfrak{R}}_n\left(\mathcal{G}_l\right).
\end{aligned}$$

Then we have:

$$\sup_{f\in\mathcal{F}}|\mathcal{R}_l(f) - \hat{\mathcal{R}}_l(f)| \leq 2\widetilde{\mathfrak{R}}_n\left(\mathcal{G}_l\right) + \frac{M}{2}\sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

Similarly, we can obtain:

$$\sup_{f\in\mathcal{F}}|\mathcal{R}_u(f) - \hat{\mathcal{R}}_u(f)| \leq 2\widetilde{\mathfrak{R}}_m\left(\mathcal{G}_u\right) + \frac{M}{2}\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

□

Lemma A.2. *Define* $\mathcal{H}_y = \{h : \boldsymbol{x} \mapsto f_y(\boldsymbol{x}) | f \in \mathcal{F}\}$ *and* $\mathfrak{R}_n\left(\mathcal{H}_y\right) = \mathbb{E}_{p(\boldsymbol{x})}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}_y}\frac{1}{n}\sum_{i=1}^{n}h\left(\boldsymbol{x}_i\right)\right]$. *And suppose that the loss function* $\mathcal{L}$ *and* $\mathcal{L}_u$ *is* ρ*-Lipschitz with respect to* $f(\boldsymbol{x})$ *Then, we have with Rademacher vector contraction inequality (Mohri et al., 2012):*

$$\widetilde{\mathfrak{R}}_n\left(\mathcal{G}_l\right) \leq \sqrt{2}\rho\sum_{y=1}^{c}\mathfrak{R}_n(\mathcal{H}_y), \qquad \widetilde{\mathfrak{R}}_m\left(\mathcal{G}_u\right) \leq \sqrt{2}\rho\sum_{y=1}^{c}\mathfrak{R}_m(\mathcal{H}_y),$$

Based on Lemma A.1 and Lemma A.2, we could obtain:

$$\begin{aligned}
\mathcal{R}(\hat{f}) - \mathcal{R}(f^\star) &= \mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f}) + \hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(f^\star) + \hat{\mathcal{R}}(f^\star) - \mathcal{R}(f^\star) \\
&\leq \mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f}) + \hat{\mathcal{R}}(f^\star) - \mathcal{R}(f^\star) \\
&= \mathcal{R}_l(\hat{f}) - \hat{\mathcal{R}}_l(\hat{f}) + \hat{\mathcal{R}}_l(f^\star) - \mathcal{R}_l(f^\star) \\
&+ \mathcal{R}_u(\hat{f}) - \hat{\mathcal{R}}_u(\hat{f}) + \hat{\mathcal{R}}_u(f^\star) - \mathcal{R}_u(f^\star) \\
&\leq 2 \sup_{f \in \mathcal{F}} |\mathcal{R}_l(f) - \hat{\mathcal{R}}_l(f)| + 2 \sup_{f \in \mathcal{F}} |\mathcal{R}_u(f) - \hat{\mathcal{R}}_u(f)| \\
&\leq 4\widetilde{\mathfrak{R}}_n\left(\mathcal{G}_l\right) + M\sqrt{\frac{\log \frac{4}{\delta}}{2n}} + 4\widetilde{\mathfrak{R}}_m\left(\mathcal{G}_u\right) + M\sqrt{\frac{\log \frac{4}{\delta}}{2m}} \\
&\leq 4\sqrt{2}\sum_{y=1}^{c} \mathfrak{R}_n(\mathcal{H}_y) + M\sqrt{\frac{\log \frac{4}{\delta}}{2n}} + 4\sqrt{2}\sum_{y=1}^{c} \mathfrak{R}_m(\mathcal{H}_y) + M\sqrt{\frac{\log \frac{4}{\delta}}{2m}} \\
&\leq 4\sqrt{2}\sum_{y=1}^{c} \mathfrak{R}_{n+m}(\mathcal{H}_y) + M\sqrt{\frac{\log \frac{4}{\delta}}{2n}} + M\sqrt{\frac{\log \frac{4}{\delta}}{2m}}
\end{aligned}$$

□