Unbiased Risk Estimator to Multi-Labeled Complementary Label Learning

Yi Gao[1,4] , **Miao Xu**[3] , **Min-Ling Zhang**[2,4*]
[1]School of Cyber Science and Engineering, Southeast University, China
[2]School of Computer Science and Engineering,Southeast University, China
[3]University of Queensland, Australia
[4]Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China
{gao_yi, zhangml}@seu.edu.cn, miao.xu@uq.edu.au

Abstract

Multi-label learning (MLL) usually requires assigning multiple relevant labels to each instance. While a fully supervised MLL dataset needs a large amount of labeling effort, using complementary labels can help alleviate this burden. However, current approaches to learning from complementary labels are mainly designed for multi-class learning and assume that each instance has a single relevant label. This means that these approaches cannot be easily applied to MLL when only complementary labels are provided, where the number of relevant labels is unknown and can vary across instances. In this paper, we first propose the unbiased risk estimator for the *multi-labeled complementary label learning* (MLCLL) problem. We also provide an estimation error bound to ensure the convergence of the empirical risk estimator. In some cases, the unbiased estimator may give unbounded gradients for certain loss functions and result in overfitting. To mitigate this problem, we improve the risk estimator by minimizing a proper loss function, which has been shown to improve gradient updates. Our experimental results demonstrate the effectiveness of the proposed approach on various datasets.

1 Introduction

Multi-label learning (MLL) is a method of training a classifier that can predict multiple labels for an unseen instance simultaneously [Zhang and Zhou, 2014a]. It has been widely used in open-environment [Zhou, 2022] and various real-world applications such as text categorization [Maltoudoglou *et al.*, 2022; Zhang *et al.*, 2022] and image retrieval [Xu *et al.*, 2022a; Ma *et al.*, 2022]. However, collecting accurate multi-labeled data can be challenging due to the difficulty of identifying small objects or complex semantic labels in complex images, as well as the unknown number of relevant labels. This requires a significant amount of labeling effort as annotators need to exercise caution and possess specialized knowledge. Precisely annotating images, such as the one shown in Fig. 1, requires a high level of attention and expertise, particularly when it comes to identifying the easily overlooked label "people" or specific geographic location label "Europe". Additionally, accurate identification of other relevant labels involves a thorough examination of each label within the entire label space.

Figure 1: An example. The image is labeled with a variety of relevant labels including "cloud", "plant", "house", "balloon", "people", "Europe", and others. The complementary label of this image is "river". Labeling the image with the label "Europe" requires specialized knowledge as it is a complex semantic label that is challenging to annotate. On the other hand, the label "people" refers to a small object that is difficult to spot, as it is highlighted by a blue box in the image.

To release the laborious of collecting multi-labeled data, we study the setting of *multi-labeled complementary label learning* (MLCLL) that could significantly reduce labeling effort, whose goal is to learn a multi-labeled classifier that can assign relevant labels for unseen instances. In MLCLL, each instance is associated with a single complementary label that specifies an irrelevant label of the instance. Obviously, providing weakly supervised information – complementary labels – releases annotation costs since selecting a complementary label does not need one-by-one checking of the entire label space and prior knowledge. For example, selecting the label "river" as the complementary label for the image in Fig. 1 is much easier than selecting all relevant labels.

Complementary label learning (CLL) is a method previously proposed for multi-class scenarios by Ishida *et al.* [2017]. Ishida *et al.* [2019] proposed an unbiased risk estimator by deriving a transition matrix using the fact that the complementary label is uniformly sampled besides the only one relevant label, such that any loss function can be

*Corresponding author

used. With such an unbiased risk estimator, the classification risk of fully-supervised learning can be evaluated on CLL training data, enabling empirical risk minimization [Vapnik, 1991]. CLL problems are further studied for biased complementary labels [Yu *et al.*, 2018b], generative adversarial network [Xu *et al.*, 2020], multiple complementary labels [Feng *et al.*, 2020], improved gradient estimation [Chou *et al.*, 2020], or easing the dependence on estimating the transition matrix [Gao and Zhang, 2021]. These existing works are based on the fact that only one relevant label exists for multi-class cases, and most solutions are strongly dependent on the known number of irrelevant labels. They cannot be used to solve the MLCLL problem, due to the unknown number of relevant labels in MLL given only a complementary label.

To address the MLCLL problem, we first propose an unbiased risk estimator that utilizes the distribution of complementary data to approximate the distribution of fully-supervised data under certain assumptions. We also establish an estimation error bound for this estimator, guaranteeing that the classifier learned from complementary labeled data will converge to the optimal one from fully-supervised MLL. Furthermore, we analyze the impact of cross-entropy loss on learning when a crucial assumption is not met and propose a gradient-friendly loss function to prevent overfitting. Our experimental results demonstrate the effectiveness of the proposed risk minimization methods. Our contributions are summarized as follows:

- The proposal of an unbiased risk estimator for MLCLL and the derivation of an estimation error bound to ensure that the classifier learned from complementary labeled data is similar to the optimal one learned from fully-supervised MLL
- The improvement of the risk estimator by using a gradient-friendly loss function to prevent overfitting, which can occur when a crucial assumption does not hold.

The rest of this paper is organized as follows. Section 2 briefly reviews related work. We introduce our proposed approach and the gradient-friendly loss function Section 3 and 4 respectively. Section 5 describes experimental results, and Section 6 gives the conclusion.

2 Related Work

In MLL, the main challenge is that the number of the output space grows exponentially as the number of labels increases [Zhang and Zhou, 2014b]. There are three routes to cope with MLL problems: *first-order approach* [Zhang *et al.*, 2018; Zhang and Zhou, 2007], *second-order approach* [Elisseeff and Weston, 2001; Fürnkranz *et al.*, 2008] and *high-order approach* [Read *et al.*, 2011; Tsoumakas *et al.*, 2011]. The first-order approaches decompose MLL problems into a series of binary classification problems to solve [Zhang *et al.*, 2018; Boutell *et al.*, 2004]. Subsequently, Zhang et al. [Zhang and Zhou, 2014b; Zhang and Wu, 2015] revealed that label correlations exist in multi-labeled data, more and more studies consider label correlations to address MLL problems [Gerych *et al.*, 2021; Li *et al.*, 2017]. Among them, the second-order approaches consider label correlations between label pairs [Fürnkranz *et al.*, 2008; Zhang and Zhou, 2006; Li *et al.*, 2017]. Beyond the second-order relationship, the high-order approaches pay attention to exploring label correlations among label sets [Read *et al.*, 2011; Gerych *et al.*, 2021; Zhao *et al.*, 2021].

In practice, collecting precisely multi-labeled data is difficult because labeling information is often incomplete [Xu *et al.*, 2022b]. A weakly supervised framework – *partial multi-label learning* (PML) – is designed to alleviate the pain of collecting precisely labeled data [Zhou, 2018]. PML was firstly proposed by Xie and Huang [2018], where each instance is associated with a set of *candidate labels* that consists of relevant labels and irrelevant (noisy) labels. Existing PML approaches handle PML problems according to the assumption that noisy labels only compose a small portion of candidate labels [Sun *et al.*, 2022; Xie and Huang, 2020; Sun *et al.*, 2019; Yu *et al.*, 2018a]. Then, these approaches use the matrix factorization technique to decompose the candidate label matrix into the low-rank multi-label matrix and the sparse noisy label matrix to solve PML problems.

CLL was first applied in multi-class learning [Ishida *et al.*, 2017], whose emergence significantly reduces annotation costs in multi-class learning. CLL aims to recover relevant labels from complementary labels. Ishida *et al.* [2019] used uniformly sampled complementary labels to derive an unbiased risk estimator that is available for arbitrary loss functions to solve the CLL problem. In addition, studies of CLL problems further involve biased complementary labels, whose implementation depends on estimating a transition matrix [Yu *et al.*, 2018b]. To ease the dependence on estimating transition matrix, Gao and Zhang [2021] designed a discriminative way to directly model the probabilities of complementary labels. Moreover, multiple complementary labels were proposed to make the learning process have more labeling information compared with one complementary label [Feng *et al.*, 2020].

Compared with PML, MLCLL considers the hardest version of this problem – a high-noise PML problem, where the candidate label set of an instance is all labels other than its complementary label. In this case, existing PML approaches may be inapplicable for MLCLL since their implementations are based on the assumption of few noisy included in candidate labels. For CLL, existing approaches are accomplished by the fact that each instance has one relevant label in multi-class learning and irrelevant labels are known. However, the number of relevant labels is unknown and can vary across instances in MLL, which leads to CLL approaches that could not handle MLCLL problems.

3 The Proposed Approach

In this section, we first introduce notations and problem setting (Sec. 3.1) and the recovery of data distribution given complementary labels (Sec. 3.2). Then, we propose an unbiased risk estimator with its estimation error bound (Sec. 3.3).

3.1 Preliminaries

Let $\mathcal{X} \subset \mathbb{R}^d$ be the feature space with d dimensions and $\mathcal{Y} = \{1, 2, 3, \ldots, K\}$ be the label space with K ($K > 2$)

possible labels. In MLL, an instance $\boldsymbol{x} \in \mathcal{X}$ with its relevant label set $Y \subseteq \mathcal{Y}$ is drawn from the unknown joint probability distribution $p(\boldsymbol{x}, Y)$. Given fully supervised data, MLL aims to learn a classifier $\boldsymbol{f} : \mathcal{X} \mapsto [0,1]^K$ by minimizing the following classification risk:

$$R(\boldsymbol{f}) = \mathbb{E}_{p(\boldsymbol{x},Y)}[\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}), Y)], \tag{1}$$

where

$$\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}), Y) = \sum_{j=1, j\in Y}^{K} \ell_j(\boldsymbol{x}) + \sum_{j=1, j\notin Y}^{K} \bar{\ell}_j(\boldsymbol{x}). \tag{2}$$

Denoting the j-th prediction of $\boldsymbol{f}$ as f_j which estimates $p(j = 1|\boldsymbol{x})$, $\ell_j(\boldsymbol{x})$ and $\bar{\ell}_j(\boldsymbol{x})$ calculate the loss of $f_j(\boldsymbol{x})$ respectively on relevant and irrelevant labels. Specially, when $\ell_j(\boldsymbol{x}) = -\log(f_j(\boldsymbol{x}))$ and $\bar{\ell}_j(\boldsymbol{x}) = -\log(1 - f_j(\boldsymbol{x}))$, $\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}), Y)$ is the popular BCE loss. $\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}), Y)$ is not limited to BCE loss only. The *mean absolute error* (MAE) loss can also be used, resulting in $\ell_j(\boldsymbol{x}) = 1 - f_j(\boldsymbol{x})$ and $\bar{\ell}_j(\boldsymbol{x}) = f_j(\boldsymbol{x})$.

For the MLCLL problem studied in this paper, fully supervised data is unavailable. Instead, $\bar{D} = \{(\boldsymbol{x}_i, \bar{y}_i)\}_{i=1}^{n}$ containing n training instances is given, where $\bar{y}_i \in \{\mathcal{Y} - Y_i\}$ is the complementary label of $\boldsymbol{x}_i \in \mathcal{X}$. Even given the complementary training set, the goal of MLCLL is still to learn a multi-labeled classifier $\boldsymbol{f} : \mathcal{X} \mapsto [0,1]^K$, i.e., the same goal as fully supervised MLL. As $(\boldsymbol{x}, Y)$ follows a distribution $p(\boldsymbol{x}, Y)$, $(\boldsymbol{x}, \bar{y})$ also follows a distribution, which is denoted as $\bar{p}(\boldsymbol{x}, \bar{y})$ on which $\bar{D}$ is drawn. In the next subsection, we will construct the relationship between $p(\boldsymbol{x}, Y)$ and $\bar{p}(\boldsymbol{x}, \bar{y})$ under some assumptions, and propose one scenario which guarantees assumptions to hold.

3.2 Recovering the Distribution

Without any additional knowledge, it is naturally difficult to construct the relationship between $p(\boldsymbol{x}, Y)$ and $\bar{p}(\boldsymbol{x}, \bar{y})$, especially when the number of unlabeled data is unknown. To enable the construction, below we provide some assumptions.

Assumption 1. *Instance-Independent Assumption: Given the complementary label $\bar{y}$, the relevant label set Y is independent of $\boldsymbol{x}$, i.e. $p(Y|\bar{y}) = p(Y|\boldsymbol{x}, \bar{y})$.*

Assumption 1 is motivated by existing works on complementary labels [Yu *et al.*, 2018b; Feng *et al.*, 2020; Ishida *et al.*, 2019], which assumes that the complementary label is conditionally independent of $\boldsymbol{x}$ given relevant labels. This assumption does not necessarily lead to the prediction of $p(Y|\boldsymbol{x})$, because the information on complementary labels in MLL is not complete with only one complementary label provided. Information on other complementary labels and learning on $\boldsymbol{x}$ are still essential. The following result provided a conclusion based on Assumption 1.

Lemma 1. *Under Assumption 1, $\sum_{\bar{y}=1,\bar{y}\notin Y}^{K} \frac{\bar{p}(\boldsymbol{x},\bar{y})}{2^{K-1}-1}$ is a valid probability mass function with respect to $\boldsymbol{x}$ and Y, i.e., it is non-negative and*

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} \sum_{\bar{y}=1,\bar{y}\notin Y}^{K} \frac{\bar{p}(\boldsymbol{x}, \bar{y})}{2^{K-1}-1} \mathrm{d}\boldsymbol{x}\mathrm{d}Y = 1. \tag{3}$$

The proof is stated in Appendix A. We then give another assumption, which describes the relationship between Y and $\bar{y}$.

Assumption 2. *Uniform Generation Assumption:*

$$p(Y|\bar{y}) = \begin{cases} \frac{1}{2^{K-1}-1}, & \bar{y} \notin Y \\ 0, & \bar{y} \in Y \end{cases}. \tag{4}$$

Assumption 2 provides the information on how complementary labels can contribute to the prediction of Y, i.e., Y can be any subset of $\mathcal{Y} - \bar{y} - \emptyset$ with uniform probability. This is a strong assumption for the difficult MLCLL problem, which implies that the complementary label $\bar{y}$ will contribute to learning no more than an irrelevant label. Such an assumption will be a necessary ground to give the unbiased risk estimation for MLCLL in Section 3.3, and will discuss an alternative solution in Section 4 if such an assumption does not hold.

With these assumptions, the following result gives the relationship between $p(\boldsymbol{x}, Y)$ and $\bar{p}(\boldsymbol{x}, \bar{y})$.

Theorem 1. *Under Assumption 1 and Assumption 2,*

$$p(\boldsymbol{x}, Y) = \sum_{\bar{y}=1,\bar{y}\notin Y}^{K} \frac{\bar{p}(\boldsymbol{x}, \bar{y})}{2^{K-1}-1}. \tag{5}$$

It has been verified in Lemma 1 that the right side of Eq. (5) forms a valid probability mass function under Assumption 1.

3.3 Unbiased Risk Estimator

An unbiased risk estimator enables the classification risk of fully-supervised MLL to be evaluated on the given MLCLL training data. The following result shows an unbiased risk estimator of the fully supervised MLL classification risk $R(\boldsymbol{f})$ defined in Eq. (1).

Theorem 2. *With $p(\boldsymbol{x}, Y)$ defined in Eq. (5) and $R(\boldsymbol{f})$ defined in Eq. (1), $R(\boldsymbol{f}) = \bar{R}(\boldsymbol{f})$, where $\bar{R}(\boldsymbol{f}) = \mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}[\bar{\mathcal{L}}(\boldsymbol{f}(\boldsymbol{x}), \bar{y})]$ is the expected risk on complementary data, and*

$$\bar{\mathcal{L}}(\boldsymbol{f}(\boldsymbol{x}), \bar{y}) = \frac{2^{K-2}}{2^{K-1}-1} \sum_{j=1, j\neq\bar{y}}^{K} \ell_j(\boldsymbol{x}) + \frac{2^{K-2}-1}{2^{K-1}-1} \sum_{j=1, j\neq\bar{y}}^{K} \bar{\ell}_j(\boldsymbol{x}) + \bar{\ell}_{\bar{y}}(\boldsymbol{x}). \tag{6}$$

The proof is proved in Appendix B. Theorem 2 shows that the fully-supervised classification risk can be estimated using the complementary labels by the unbiased estimator $\bar{R}(\boldsymbol{f})$, with the corresponding complementary loss function defined in Eq. (6). Eq. (6) shows that the complementary loss gives low confidence in treating other labels as either relevant or irrelevant with smaller-than-one weights for them.

Estimation Error Bound. With Theorem 2, the expected risk $\bar{R}(\boldsymbol{f})$ can be approximated by its empirical estimation $\bar{R}_n(\boldsymbol{f})$, which is

$$\bar{R}_n(\boldsymbol{f}) = \frac{1}{n} \sum_{i=1}^{n} \bar{\mathcal{L}}(\boldsymbol{f}(\boldsymbol{x}_i), \bar{y}_i). \tag{7}$$

Denote $\mathcal{F}$ as the hypothesis class and $\mathcal{G}_j = \{g : \boldsymbol{x} \mapsto f_j(\boldsymbol{x}) | \boldsymbol{f} \in \mathcal{F}\}$ as the functional space for the label $j \in \mathcal{Y}$. $\mathfrak{R}_n(\mathcal{G}_j)$ gives the *Rademacher Complexity* [Bartlett and Mendelson, 2002] of $\mathcal{G}_j$, defined as $\mathfrak{R}_n(\mathcal{G}_j) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}_j} \frac{1}{n} \sum_{i=1}^{n} g(\boldsymbol{x}_i)\right]$. We further denote $\boldsymbol{f}_n \in \mathcal{F}$

Algorithm 1 MLCLL with the GDF loss

Input:
$\bar{D} = \{(\boldsymbol{x}_i, \bar{y}_i)\}_{i=1}^n$: the training data;
E : the number of epochs;
$\mathcal{A}$: an external stochastic optimization algorithm
Output:
$\boldsymbol{\theta}$: model parameter for $\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\theta})$

1: **for** $t = 1$ to E **do**
2: Let $\mathcal{L}$ be the risk, $\mathcal{L} = \frac{1}{n}\sum_{i=1}^n \bar{\mathcal{L}}_{\mathrm{GDF}}(\boldsymbol{f}(\boldsymbol{x}_i), \bar{y}_i) = \frac{1}{n}\sum_{i=1}^n \left\{ -\sum_{j=1, j\neq\bar{y}_i}^K \log f_j(\boldsymbol{x}_i) - \log(1 - f_{\bar{y}_i}(\boldsymbol{x}_i)) \right\}$;
3: Set gradient $-\bigtriangledown_{\boldsymbol{\theta}} \mathcal{L}$;
4: Update $\boldsymbol{\theta}$ by $\mathcal{A}$;
5: **end for**

and $\boldsymbol{f}^* \in \mathcal{F}$ as the minimizers of $\bar{R}_n(\boldsymbol{f})$ and $R(\boldsymbol{f})$ respectively. Based on these notations, we give the estimation error bound.

Theorem 3. *Suppose* $M = \sup_{\boldsymbol{x}\in\mathcal{X}, \boldsymbol{f}\in\mathcal{F}} \bar{\mathcal{L}}(\boldsymbol{f}(\boldsymbol{x}), \bar{y})$. *For any* $j \in \mathcal{Y}$, *assuming* $\ell_j(\boldsymbol{x})$ *and* $\bar{\ell}_j(\boldsymbol{x})$ *are* β^+*-Lipschitz and* β^-*-Lipschitz with respect to* $\boldsymbol{f}(\boldsymbol{x})$ *respectively. For any* δ, *with the probability at least* $1-\delta$,

$$R(\boldsymbol{f}_n) - R(\boldsymbol{f}^*) \leq M\sqrt{\frac{\log 2/\delta}{2n}} + 4\sqrt{2}\left[\frac{(K-1)2^{K-2}}{2^{K-1}-1}\beta^+ + \frac{(K-1)2^{K-2}-K}{2^{K-1}-1}\beta^-\right]\sum_{j=1}^K \mathfrak{R}_n(\mathcal{G}_j). \tag{8}$$

The proof is shown in Appendix C. Theorem 3 demonstrates that the empirical risk minimizer converges to the true risk minimizer as $n \to \infty$. It also demonstrates the impact of K on learning complementary labeled problems.

4 Gradient Descent Friendly Loss

In the above section, necessary assumptions are made for giving an unbiased risk estimator for the MLCLL problem, especially Assumption 2, which assumes that there exists no additional bias in learning besides that $\bar{y}$ is an irrelevant label. In this section, we will investigate the situation that Assumption 2 does not hold from the perspective of the loss function and give a moderated solution.

Our proposed complementary loss function Eq. (6) is eligible to accommodate any loss functions. If we use the popular BCE loss, we have

$$\bar{\mathcal{L}}_{\mathrm{BCE}}(\boldsymbol{f}(\boldsymbol{x}), \bar{y}) = -\frac{2^{K-2}}{2^{K-1}-1}\sum_{j=1, j\neq\bar{y}}^K \log(f_j(\boldsymbol{x}_i)) \tag{9}$$
$$-\frac{2^{K-2}-1}{2^{K-1}-1}\sum_{j=1, j\neq\bar{y}}^K \log(1 - f_j(\boldsymbol{x})) - \log(1 - f_{\bar{y}}(\boldsymbol{x})).$$

Motivated by [Chou *et al.*, 2020] which concluded that gradient estimation is important for weakly supervised learning, we further give the gradient of $\bar{\mathcal{L}}_{\mathrm{BCE}}$ with respect to $\boldsymbol{\theta}$, which is learnable parameters for $f_j(\boldsymbol{x})$:

$$\frac{\partial\bar{\mathcal{L}}_{\mathrm{BCE}}}{\partial\boldsymbol{\theta}} = \begin{cases} -(w^+ + w^-)\bigtriangledown_{\boldsymbol{\theta}} f_j(\boldsymbol{x};\boldsymbol{\theta}), & \text{if } j \neq \bar{y}, \\ \frac{1}{1-f_j(\boldsymbol{x};\boldsymbol{\theta})}\bigtriangledown_{\boldsymbol{\theta}} f_j(\boldsymbol{x};\boldsymbol{\theta}), & \text{if } j = \bar{y}, \end{cases} \tag{10}$$

where $w^+ = \frac{2^{K-2}}{2^{K-1}-1}\frac{1}{f_j(\boldsymbol{x};\boldsymbol{\theta})}$ and $w^- = \frac{2^{K-2}-1}{2^{K-1}-1}\frac{1}{1-f_j(\boldsymbol{x};\boldsymbol{\theta})}$.

The calculation in Eq. (10) is divided into two parts: on the complementary label $\bar{y}$ and non-complementary labels $\mathcal{Y} - \bar{y}$. On the complementary label, the gradient descent favors a prediction of zero for $f_{\bar{y}}(\boldsymbol{x};\boldsymbol{\theta})$. However, serious overfitting can happen regarding the punishment on non-complementary labels $\mathcal{Y} - \bar{y}$ if Assumption 2 does not hold, i.e., $\bar{y}$ provides some bias in learning other labels.

To give a simple example, imagine there is another label y_s which is similar to $\bar{y}$, such that its prediction $f_{y_s}(\boldsymbol{x};\boldsymbol{\theta})$ will be close to the prediction on $f_{\bar{y}}(\boldsymbol{x};\boldsymbol{\theta})$, i.e., zero. When $f_{y_s}(\boldsymbol{x};\boldsymbol{\theta}) = 0$ or close to zero, w^+ in Eq. (10) will become ∞, resulting in ∞ gradient although the prediction f_{y_s} is already close to the groundtruth. In another example, when a label y_d is absolutely exclusive of $\bar{y}$, its close-to-groundtruth prediction should be close to one, resulting in an infinity w^-, still an ∞ gradient despite that the prediction is close to groundtruth. These two naive examples could demonstrate that if Assumption 2 is violated, serious overfitting could happen when using the proposed unbiased risk estimator.

To mitigate the overfitting caused due to violating Assumption 2, we design a new gradient by using the $\bar{y}$ part in Eq. (10), but a modification on the gradient of the $Y - \bar{y}$ part. Such a gradient is potentially calculated from a Gradient-Descent-Friendly (GDF) loss function and shown as

$$\frac{\partial\bar{\mathcal{L}}_{\mathrm{GDF}}}{\partial\boldsymbol{\theta}} = \begin{cases} -\frac{1}{f_j(\boldsymbol{x};\boldsymbol{\theta})}\bigtriangledown_{\boldsymbol{\theta}} f_j(\boldsymbol{x};\boldsymbol{\theta}), & \text{if } j \neq \bar{y}, \\ \frac{1}{1-f_j(\boldsymbol{x};\boldsymbol{\theta})}\bigtriangledown_{\boldsymbol{\theta}} f_j(\boldsymbol{x};\boldsymbol{\theta}), & \text{if } j = \bar{y}. \end{cases} \tag{11}$$

Eq. (11) achieves a kind of gradient that will enable any label belonging to $Y - \bar{y}$ to have a prediction higher than the prediction on the complementary label, i.e., achieving a good ranking effect. When the label absolutely exclusive from $\bar{y}$ exists, such a gradient will favor the prediction to be close to one. It seems that the problem still remains if there is any label similar to $\bar{y}$ exists, however, the learning on $\bar{y}$ will provide a way of balancing the learning of these two labels, and a good ranking performance can still be achieved in MLL metrics such as *ranking loss* or *average precision*. In Eq. (11), we remove all weights related to K to make the gradients caused by individual labels have uniform weights in case one label dominates the learning.

The corresponding loss function of Eq. (11) is

$$\bar{\mathcal{L}}_{\mathrm{GDF}}(\boldsymbol{f}(\boldsymbol{x}), \bar{y}) = -\sum_{j=1, j\neq\bar{y}}^K \log(f_j(\boldsymbol{x})) - \log(1 - f_{\bar{y}}(\boldsymbol{x})). \tag{12}$$

The procedure for optimizing the GDF loss is shown in Algorithm 1. We will show in the next section how optimizing Eq. (12) can mitigate overfitting and improve performance.

MAE loss is another widely used loss function, which is shown to be less likely to overfit but slower at convergence [Ghosh *et al.*, 2017]. Formally, the MLCLL corresponding MAE loss and its gradients are

$$\bar{\mathcal{L}}_{\mathrm{MAE}}(\boldsymbol{f}(\boldsymbol{x}), \bar{y}) = \frac{2^{K-2}}{2^{K-1}-1}\sum_{j=1, j\neq\bar{y}}^K (1 - f_j(\boldsymbol{x}))$$

Table 1: Experimental results (mean±std) on training data with first pre-processing way (each instance is given one complementary label and labels are kept under 15). The best performance of each dataset is shown in **boldface**, where •/○ denotes whether GDF is superior/inferior to baselines with pairwise t-test (at 0.05 significance level).

Methods	MLL		PML		CLL	Unbounded	Bounded	GDF
	ML-KNN	LIFT	fpml	PML-LRS	L-UW	BCE	MAE	
One Error↓								
bookmark	.801±.006•	.649±.015•	.885±.019•	.584±.004•	.504±.008•	.568±.006•	.613±.011•	**.483±.007**
Corel16k	.736±.054•	.789±.044•	.816±.023•	.730±.000•	.703±.058	.798±.019•	**.702±.053**	.706±.053
delicious	.592±.017•	.533±.014•	.617±.017•	.452±.007•	.482±.018•	.490±.011•	.467±.019•	**.426±.012**
mediamill	.198±.020•	.187±.019	.188±.019	.200±.002•	.188±.019	.190±.014•	.188±.019	**.184±.018**
eurlex_dc	.790±.033•	.759±.024•	.920±.024•	.517±.013•	.609±.024•	.575±.019•	.906±.010•	**.459±.015**
eurlex_sm	.667±.017•	.674±.013•	.868±.032•	.457±.005•	.552±.011•	.534±.012•	.834±.008•	**.400±.010**
scene	.692±.029•	.605±.022•	.815±.026•	.540±.022•	.764±.023•	.449±.031•	.805±.021•	**.383±.020**
yeast	.297±.028•	.284±.027•	.251±.023	.738±.097•	.253±.022	.277±.016•	.251±.023	**.249±.024**
Ranking Loss↓								
bookmark	.348±.006•	.310±.007•	.468±.019•	.260±.004•	.196±.007	.260±.005•	.301±.008•	**.196±.005**
Corel16k	.328±.045•	.392±.026•	.420±.031•	.303±.005	**.301±.039**	.400±.016•	.316±.032	.307±.040
delicious	.398±.004•	.383±.003•	.438±.008•	.305±.002•	.319±.005•	.336±.003•	.305±.004•	**.292±.005**
mediamill	.200±.014•	.202±.015•	.206±.019•	.160±.001○	.193±.020•	**.157±.009**○	.192±.019•	.174±.013
eurlex_dc	.283±.008•	.294±.012•	.470±.037•	.179±.006•	.267±.015•	.216±.017•	.437±.008•	**.161±.005**
eurlex_sm	.322±.007•	.333±.013•	.472±.034•	.238±.004•	.329±.005•	.273±.014•	.461±.006•	**.209±.006**
scene	.340±.030•	.289±.019•	.504±.024•	.258±.006•	.457±.018•	.184±.019•	.492±.023•	**.153±.009**
yeast	.247±.011•	.298±.012•	.233±.012•	.464±.018•	.252±.012•	.250±.015•	.238±.014•	**.219±.013**
Average Precision↑								
bookmark	.383±.006•	.480±.010•	.267±.018•	.534±.004•	.604±.005•	.544±.005•	.494±.008•	**.619±.006**
Corel16k	.405±.047	.350±.033•	.325±.021•	.423±.006	**.434±.047**	.343±.017•	.426±.040	.429±.044
delicious	.487±.006•	.511±.004•	.457±.006•	.580±.002•	.554±.006•	.544±.005•	.567±.006•	**.586±.004**
mediamill	.711±.010•	.710±.009•	.709±.012•	.748±.001○	.717±.012•	**.765±.005**○	.715±.012•	.732±.007
eurlex_dc	.417±.018•	.429±.016•	.240±.028•	.616±.009•	.525±.017•	.563±.014•	.267±.009•	**.658±.010**
eurlex_sm	.426±.009•	.425±.012•	.285±.026•	.584±.004•	.475±.008•	.524±.012•	.288±.003•	**.623±.009**
scene	.543±.023•	.600±.016•	.417±.020•	.637±.010•	.463±.018•	.717±.022•	.429±.018•	**.759±.012**
yeast	.677±.018•	.636±.016•	.688±.016•	.459±.031•	.679±.016•	.679±.015•	.693±.017•	**.712±.018**

$$+\frac{2^{K-2}-1}{2^{K-1}-1}\sum_{j=1,j\neq\bar{y}}^{K} f_j(\boldsymbol{x})+f_{\bar{y}}(\boldsymbol{x}), \tag{13}$$

and

$$\frac{\partial\bar{\mathcal{L}}_{\mathrm{MAE}}}{\partial\boldsymbol{\theta}}=\begin{cases}-\frac{1}{2^{K-1}-1}\bigtriangledown_{\boldsymbol{\theta}} f_j(\boldsymbol{x};\boldsymbol{\theta}), & \text{if } j\neq\bar{y},\\ \bigtriangledown_{\boldsymbol{\theta}} f_j(\boldsymbol{x};\boldsymbol{\theta}), & \text{if } j=\bar{y}.\end{cases} \tag{14}$$

Eq. (14) shows the reason why the $\bar{\mathcal{L}}_{\mathrm{MAE}}$ is less likely to overfit but slower in convergence compared to the $\bar{\mathcal{L}}_{\mathrm{BCE}}$. By comparing Eqs. (10), (14) and (11), we can see that the GDF loss is a trade-off between the MAE one and the BCE one from the perspective of the gradient. The relationship between $\bar{\mathcal{L}}_{\mathrm{GDF}}$ and $\bar{\mathcal{L}}_{\mathrm{MAE}}$ can be further shown by the fact that $\bar{\mathcal{L}}_{\mathrm{MAE}} \leq \bar{\mathcal{L}}_{\mathrm{GDF}}$. The empirical comparison of optimizing these three loss functions is shown in Section 5.

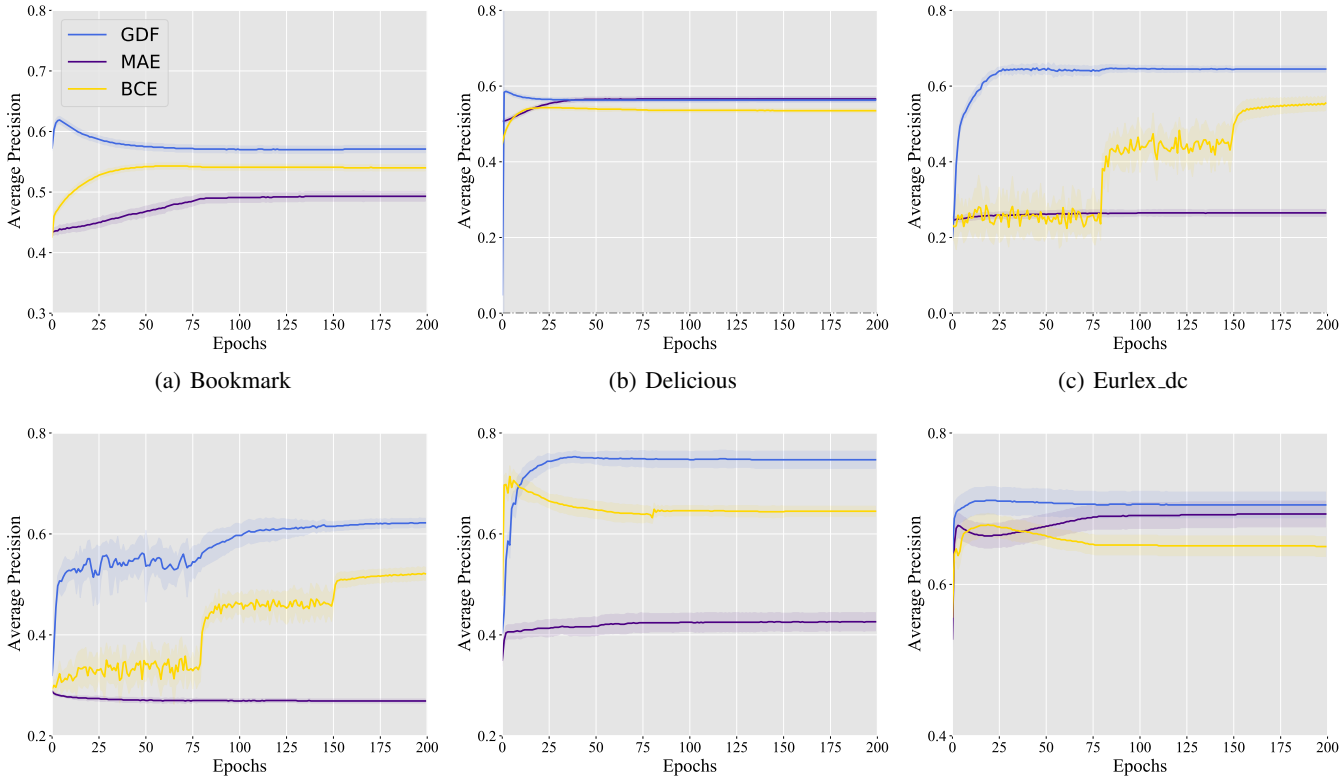
5 Experiments

In this section, we conduct experiments to evaluate the performance of the unbiased risk estimator with various loss functions and our proposed GDF loss. Here, we adopt five MLL criteria, including *ranking loss*, *hamming loss*, *one error*, *coverage* and *average precision*, to measure the performance of approaches. For *average precision*, the greater the value, the better the performance, while for the remaining four criteria, the smaller the values the better the performance. We use PyTorch [Paszke *et al.*, 2019] and NVIDIA TITAN RTX to implement our experiments, where the code is available at https://github.com/GaoYi439/GDF.

5.1 Experimental Settings

Datasets & pre-processing. We use eight widely-used MLL datasets to experiments[1], where we adopt two pre-processing ways to process datasets to verify the performance of the proposed approach. The first way follows Xie et al. [2018; 2020], rare labels and instances being relevant to these rare labels are removed for datasets with more than 15 labels, whose

[1]Publicly available at https://mulan.sourceforge.net/datasets-mlc.html.



(a) Bookmark (b) Delicious (c) Eurlex_dc

(d) Eurlex_sm (e) Scene (f) Yeast

Figure 2: *Average precision* on various datasets adopted the first pre-processing way, where each instance is associated with one complementary label and labels are kept under 15. Dark colors show the mean of testing average precision and light colors are corresponding to the std.

labels are kept under 15. Datasets use the first way to process, where each instance is associated with one complementary label. The second way keeps the original number of labels for datasets, while the number of complementary labels per instance is half the original number of labels in datasets. Detailed descriptions of these datasets are shown in Appendix D.

Baselines. We adopt two PML approaches (fpml [Yu *et al.*, 2018a] and PML-LRS [Sun *et al.*, 2019]) as baselines, which use candidate labels ($\mathcal{Y} \setminus \bar{y}$) to learn. Similarly, a CLL approach, L-UW [Gao and Zhang, 2021], is used in our experiments. Due to that L-UW is applied in multi-class learning, we use the Sigmoid layer and BCE loss to replace the Softmax layer and cross-entropy loss respectively, and help it suit the MLCLL problem. In addition, ML-KNN [Zhang and Zhou, 2007] and LIFT [Zhang and Wu, 2015] are two MLL approaches, which are absorbed to compare with our approach. They deal with MLCLL by taking all candidate labels ($\mathcal{Y} \setminus \bar{y}$) of instances as possible labels. To verify the feasibility of theoretical analysis in Section 4, we use the BCE loss $\bar{\mathcal{L}}_{\mathrm{BCE}}$ and MAE loss $\bar{\mathcal{L}}_{\mathrm{MAE}}$ to our empirical risk estimator Eq. (7) as baselines. The proposed GDF loss is inserted into Eq. (7).

Setup. We adopt the linear model as the predictive model for fair comparisons and apply SGD with momentum 0.9 for optimization. We set batch-size and training epoch as 256 and 200 respectively. Weight decay is set as 10^{-3} and learning rate is selected from $\{10^{-1}, 10^{-2}, 10^{-3}\}$, where the learning rate is multiplied by 0.1 at 100 and 150 epochs [Wu *et al.*, 2018]. We apply the same model and hyper-parameters of ours for L-UW, $\bar{\mathcal{L}}_{\mathrm{BCE}}$ and $\bar{\mathcal{L}}_{\mathrm{MAE}}$. Ten-fold cross-validation is used to evaluate the performance of all approaches. Here, training data is only equipped with complementary labels and test data associated with the sets of relevant labels is used to verify the performance of approaches. We adopt the mean and *standard deviation* (std) of five criteria. $\downarrow$ / $\uparrow$ indicates that criteria are smaller/larger, the performance is better.

5.2 Main Empirical Results

Results. Table 1 reports *one error*, *ranking loss*, and *average precision* of various approaches over eight datasets (the results of *hamming loss* and *coverage* are shown in Appendix D). Here, all datasets are processed by the first pre-processing way, where each instance is associated with one complementary label. As shown in Table 1, GDF outperforms the most approaches on eight datasets. Specifically, we improve upon the best baseline on *one error*, *ranking loss*, and *average precision* by 0.066, 0.031, and 0.099 respectively on scene and eurlex_sm datasets. This indicates the effectiveness of our proposed approach GDF to solve the MLCLL problem.

Moreover, GDF achieves comparable results on three criteria against two MLL approaches over all datasets. For example, GDF on *average precision* are significantly superior to ML-KNN and LIFT by 0.241 and 0.229 respectively on eurlex_dc dataset. This is because our approach is more suitable for dealing with complementary labeled data than MLL approaches. GDF achieves similar or better performance to

Table 2: Experimental results (mean±std) on the training data adopted the second pre-processing way. The label space is original and the number of complementary labels per instance is half the original number of labels. The best performance of each dataset is shown in **boldface**, where •/◦ indicates whether GDF is superior/inferior to baselines with pairwise t-test (at 0.05 significance level).

Methods	MLL	PML		CLL	Bounded	GDF
	ML-KNN	fpml	PML-LRS	L-UW	MAE	
Ranking Loss↓						
Corel16k	.276±.043	.286±.054	.281±.013	.293±.052•	.307±.051•	.286±.036
delicious	.235±.002•	.204±.002•	.261±.009•	.195±.002•	.201±.002•	.186±.002
eurlex_dc	.267±.004•	.263±.005•	.243±.003•	.252±.006•	.248±.008•	.239±.005
eurlex_sm	.173±.004•	.208±.007•	.170±.002•	.167±.004•	.167±.003•	.159±.004
Average Precision↑						
Corel16k	.224±.042•	.238±.045	.223±.009•	.244±.046	.245±.046	.240±.042
delicious	.205±.003•	.201±.002•	.232±.001•	.235±.003•	.234±.003•	.299±.004
eurlex_dc	.300±.011	.179±.007•	.299±.003•	.278±.006•	.233±.008•	.300±.008
eurlex_sm	.407±.009•	.204±.004•	.413±.002	.370±.006•	.321±.009•	.413±.008

PML approaches. Especially, the *average precision* of GDF is 0.418 higher than fpml on eurlex_dc dataset, which shows that our approach can work well on MLCLL with the high-noisy problem compared with PML approaches.

Results of GDF shown in Table 1 outperform L-UW on all datasets, which indicates that the CLL approach can not enough handle the MLCLL problem. This is because their implementations depend on the assumption of one relevant label per instance, while the number of relevant labels for each instance is unknown in MLL. That leads to CLL approaches could not hold the relationship of multiple relevant labels and complementary labels in MLCLL. In addition, we observe that the unbounded loss BCE used to the empirical risk estimator Eq. (7) is inferior to our proposed GDF loss in most cases, which proves that GDF is indeed better than the unbounded one in MLCLL. Accordingly, experimental results of the empirical risk estimator Eq. (7) with MAE loss on three criteria are lower than ours in all datasets, which further proves the effectiveness of our proposed strategy – minimizing the empirical risk estimator by the GDF loss.

Effect of the gradient descent friendly loss GDF. Fig. 2 shows *average precision* of BCE loss, MAE loss and GDF loss on various datasets over 200 epochs, where six datasets both adopt the first pre-processing way to process and each instance is associated with one complementary label. As can be seen from Fig. 2, the curve of MAE loss is lower than that of BCE loss, but MAE loss is more stable than BCE loss in Fig. 2. It clearly illustrates the advantages and disadvantages of MAE loss: the bounded loss MAE is more robust for noisy data than unbounded loss BCE, while its convergence rate is inferior to the unbounded loss [Ghosh *et al.*, 2017]. In addition, we observe that the curve of our proposed GDF loss is higher than these of the unbounded loss BCE and the bounded loss MAE, which demonstrates the effectiveness of our approach. Specifically, the curve of GDF loss shows that our approach has both the robustness of the bounded loss MAE and the superior convergence of the unbounded loss BCE. This demonstrates that removing the easily caused overfitting part of BCE loss to design GDF loss is a reasonable strategy in MLCLL to improve performance of our approach, which also confirms that GDF loss to optimize in MLCLL brings benefits to gradient update.

5.3 Additional Experiments

Table 2 shows *ranking loss* and *average precision* of different approaches on four datasets that are adopted the second pre-processing way to process, the remaining experimentations of three criteria are shown in Appendix D. Here, each dataset keeps the original number of labels and complementary labels given per instance is half of $L(\mathcal{S})$ (the number of labels). In Table 2, the proposed approach achieves comparable performance to excellent baselines on four datasets, which indicates that GDF loss can also work on datasets with the original label space. Specifically, results of fpml and PML-LRS are inferior to GDF loss, which indicates that PML approaches depending on the assumption of sparse noisy labels can not deal with the high-noisy label problem of MLCLL. Table 2 illustrates that GDF loss outperforms L-UW, which proves that the proposal is better than the CLL approach in tackling the relationship between multiple relevant labels and complementary labels. GDF loss obtains better performance compared with MAE loss, which indicates that the proposed loss function designed by improving gradient updates can also deal with the multiple complementary-label problem.

6 Conclusion

In this paper, we study the setting of MLCLL, which aims to learn a multi-labeled classifier from complementary labeled data. To solve this problem, we propose an unbiased risk estimator with an estimation error bound, which supports that the learned risk minimizer from complementary labeled data converges to the optimal one of fully supervised MLL. Although our unbiased risk estimator has no restrictions on loss functions, it will produce unbounded gradients if certain unbounded loss functions are used and result in overfitting. To alleviate this issue, we design GDF loss to prevent the overfitting problem and find that it brings benefits to gradient updates. We verify the effectiveness of the proposed upper bound loss on various datasets.

References

[Bartlett and Mendelson, 2002] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 2111:224–240, 2002.

[Boutell *et al.*, 2004] Matthew R. Boutell, Jie-Bo Luo, Xi-Peng Shen, and Christopher M. Brown. Learning multi-label scene classification. *Pattern Recognit.*, 37(9):1757–1771, 2004.

[Chou *et al.*, 2020] Yu-Ting Chou, Gang Niu, Hsuan-Tien Lin, and Masashi Sugiyama. Unbiased risk estimators can mislead: A case study of learning with complementary labels. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1929–1938, Virtual Event, 2020.

[Elisseeff and Weston, 2001] André Elisseeff and Jason Weston. A kernel method for multi-labelled classification. In *Advances in Neural Information Processing Systems 14*, pages 681–687, Vancouver, Canada, 2001.

[Feng *et al.*, 2020] Lei Feng, Takuo Kaneko, Bo Han, Gang Niu, Bo An, and Masashi Sugiyama. Learning with multiple complementary labels. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3072–3081, Virtual Event, 2020.

[Fürnkranz *et al.*, 2008] Johannes Fürnkranz, Eyke Hüllermeier, Eneldo Loza Mencía, and Klaus Brinker. Multilabel classification via calibrated label ranking. *Machine learning*, 73(2):133–153, 2008.

[Gao and Zhang, 2021] Yi Gao and Min-Ling Zhang. Discriminative complementary-label learning with weighted loss. In *Proceedings of the 38th International Conference on Machine Learning*, pages 3587–3597, Virtual Event, 2021.

[Gerych *et al.*, 2021] Walter Gerych, Thomas Hartvigsen, Luke Buquicchio, Emmanuel Agu, and Elke A. Rundensteiner. Recurrent bayesian classifier chains for exact multi-label classification. In *Advances in Neural Information Processing Systems 34*, pages 15981–15992, Virtual Event, 2021.

[Ghosh *et al.*, 2017] Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 1919–1925, San Francisco, CA, 2017.

[Ishida *et al.*, 2017] Takashi Ishida, Gang Niu, Wei-Hua Hu, and Masashi Sugiyama. Learning from complementary labels. In *Advances in Neural Information Processing Systems 30*, pages 5639–5649, Long Beach, CA, 2017.

[Ishida *et al.*, 2019] Takashi Ishida, Gang Niu, Aditya Krishna Menon, and Masashi Sugiyama. Complementary-label learning for arbitrary losses and models. In *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 2971–2980, Long Beach, CA, 2019.

[Li *et al.*, 2017] Yun-Cheng Li, Yale Song, and Jie-Bo Luo. Improving pairwise ranking for multi-label image classification. In *Proceedings of 2017 IEEE conference on computer vision and pattern recognition*, pages 3617–3625, Honolulu, HI, 2017.

[Ma *et al.*, 2022] Ze-Yu Ma, Xiao Luo, Ying-Jie Chen, Mi-Xiao Hou, Jin-Xing Li, Ming-Hua Deng, and Guang-Ming Lu. Improved deep unsupervised hashing with fine-grained semantic similarity mining for multi-label image retrieval. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, pages 1254–1260, Vienna, Austria, 2022.

[Maltoudoglou *et al.*, 2022] Lysimachos Maltoudoglou, Andreas Paisios, Ladislav Lenc, Jirí Martínek, Pavel Král, and Harris Papadopoulos. Well-calibrated confidence measures for multi-label text classification with a large number of labels. *Pattern Recognit.*, 122:108271, 2022.

[Maurer, 2016] A. Maurer. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory*, 2016.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Ze-Ming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Jun-Jie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035, Vancouver, Canada, 2019.

[Read *et al.*, 2011] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Mach. Learn.*, 85(3):333–359, 2011.

[Sun *et al.*, 2019] Li-Juan Sun, Song-He Feng, Tao Wang, Cong-Yan Lang, and Yi Jin. Partial multi-label learning by low-rank and sparse decomposition. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pages 5016–5023, Honolulu, HI, 2019.

[Sun *et al.*, 2022] Li-Juan Sun, Song-He Feng, Jun Liu, Geng-Yu Lyu, and Cong-Yan Lang. Global-local label correlation for partial multi-label learning. *IEEE Trans. Multim.*, 24:581–593, 2022.

[Tsoumakas *et al.*, 2011] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis P. Vlahavas. Random k-labelsets for multilabel classification. *IEEE Trans. Knowl. Data Eng.*, 23(7):1079–1089, 2011.

[Vapnik, 1991] Vladimir Vapnik. Principles of risk minimization for learning theory. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, pages 831–838, 1991.

[Wu *et al.*, 2018] Zhi-Rong Wu, Yuan-Jun Xiong, Stella X. Yu, and Da-Hua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 2018.

[Xie and Huang, 2018] Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 4302–4309, New Orleans, LA, 2018.

[Xie and Huang, 2020] Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning with noisy label identification. In *Proceedings of 34th AAAI Conference on Artificial Intelligence*, pages 6454–6461, New York, NY, 2020.

[Xu *et al.*, 2020] Yan-Wu Xu, Ming-Ming Gong, Jun-Xiang Chen, Tong-Liang Liu, Kun Zhang, and Kayhan Batmanghelich. Generative-discriminative complementary learning. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 6526–6533, New York, NY, 2020.

[Xu *et al.*, 2022a] Cheng-Yin Xu, Zeng-Hao Chai, Zheng-Zhuo Xu, Chun Yuan, Yan-Bo Fan, and Jue Wang. Hyp^2 loss: Beyond hypersphere metric space for multi-label image retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3173–3184, Lisboa, Portugal, 2022.

[Xu *et al.*, 2022b] Ning Xu, Cong-Yu Qiao, Jia-Qi Lv, Xin Geng, and Min-Ling Zhang. One positive label is sufficient: Single-positive multi-label learning with label enhancement. In *Advances in Neural Information Processing Systems 35*, in press, 2022.

[Yu *et al.*, 2018a] Guo-Xian Yu, Xia Chen, Carlotta Domeniconi, Jun Wang, Zhao Li, Zi-Li Zhang, and Xin-Dong Wu. Feature-induced partial multi-label learning. In *Proceedings of 2018 IEEE International Conference on Data Mining*, pages 1398–1403, Singapore, 2018.

[Yu *et al.*, 2018b] Xi-Yu Yu, Tong-Liang Liu, Ming-Ming Gong, and Da-Cheng Tao. Learning with biased complementary labels. In *Proceedings of the 15th European Conference on Computer Vision*, pages 69–85, Munich, Germany, 2018.

[Zhang and Wu, 2015] Min-Ling Zhang and Lei Wu. Lift: Multi-label learning with label-specific features. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(1):107–120, 2015.

[Zhang and Zhou, 2006] Min-Ling Zhang and Zhi-Hua Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE transactions on Knowledge and Data Engineering*, 18(10):1338–1351, 2006.

[Zhang and Zhou, 2007] Min-Ling Zhang and Zhi-Hua Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognit.*, 40(7):2038–2048, 2007.

[Zhang and Zhou, 2014a] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.*, 26(8):1819–1837, 2014.

[Zhang and Zhou, 2014b] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.*, 26(8):1819–1837, 2014.

[Zhang *et al.*, 2018] Min-Ling Zhang, Yu-Kun Li, Xu-Ying Liu, and Xin Geng. Binary relevance for multi-label learning: an overview. *Frontiers Comput. Sci.*, 12(2):191–202, 2018.

[Zhang *et al.*, 2022] Yu Zhang, Zhi-Hong Shen, Chieh-Han Wu, Bo-Ya Xie, Jun-Heng Hao, Ye-Yi Wang, Kuan-San Wang, and Jia-Wei Han. Metadata-induced contrastive learning for zero-shot multi-label text classification. In *Proceedings of the ACM Web Conference 2022*, pages 3162–3173, Virtual Event, 2022.

[Zhao *et al.*, 2021] Wen-Ting Zhao, Shu-Feng Kong, Jun-Wen Bai, Daniel Fink, and Carla P. Gomes. HOT-VAE: learning high-order label correlation for multi-label classification via attention-based variational autoencoders. In *Proceedings of 35th AAAI Conference on Artificial Intelligence*, pages 15016–15024, Virtual Event, 2021.

[Zhou, 2018] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2018.

[Zhou, 2022] Zhi-Hua Zhou. Open-environment machine learning. *National Science Review*, 9(8):nwac123, 2022.

A The Proof of Lemma 1

Lemma 1. *Under Assumption 1, $\sum_{\bar{y}=1,\bar{y}\notin Y}^{K} \frac{\bar{p}(\boldsymbol{x},\bar{y})}{2^{K-1}-1}$ is a valid probability mass function with respect to $\boldsymbol{x}$ and Y, i.e., it is non-negative and*

$$\int_{\mathcal{X}}\int_{\mathcal{Y}} \sum_{\bar{y}=1,\bar{y}\notin Y}^{K} \frac{\bar{p}(\boldsymbol{x},\bar{y})}{2^{K-1}-1}\mathrm{d}\boldsymbol{x}\mathrm{d}Y = 1.$$

Proof. To make MLCLL hold, we remove two special subsets $Y=\emptyset$ and $Y=\mathcal{Y}$ from $\mathcal{Y}$, so let $\mathcal{Y}'=\{2^{\mathcal{Y}}-\emptyset-\mathcal{Y}\}$. Then, we have

$$\begin{aligned}
\int_{\mathcal{X}}\int_{\mathcal{Y}} \sum_{\bar{y}=1,\bar{y}\notin Y}^{K} \frac{\bar{p}(\boldsymbol{x},\bar{y})}{2^{K-1}-1}\mathrm{d}\boldsymbol{x}\mathrm{d}Y &= \int_{\mathcal{X}} \sum_{Y\in\mathcal{Y}'} \sum_{\bar{y}=1,\bar{y}\notin Y}^{K} \frac{\bar{p}(\boldsymbol{x},\bar{y})}{2^{K-1}-1}\mathrm{d}\boldsymbol{x} \\
&= \int_{\mathcal{X}} \sum_{\bar{y}=1}^{K} \sum_{Y\in\mathcal{Y}',\bar{y}\notin Y} \frac{\bar{p}(\boldsymbol{x},\bar{y})}{2^{K-1}-1}\mathrm{d}\boldsymbol{x} \quad \because |\{Y|Y\in\mathcal{Y}',\bar{y}\notin Y\}| = (2^{K-1}-1) \\
&= \int_{\mathcal{X}} \sum_{\bar{y}=1}^{K} \bar{p}(\boldsymbol{x},\bar{y})\mathrm{d}\boldsymbol{x} \\
&= 1.
\end{aligned}$$

□

B The Proof of Theorem 2

Theorem 2. *With $p(\boldsymbol{x},Y)$ defined in Eq. (5) and $R(\boldsymbol{f})$ defined in Eq. (1), $R(\boldsymbol{f})=\bar{R}(\boldsymbol{f})$, where $\bar{R}(\boldsymbol{f})=\mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}[\bar{\mathcal{L}}(\boldsymbol{f}(\boldsymbol{x}),\bar{y})]$ is the expected risk on complementary data, and*

$$\bar{\mathcal{L}}(\boldsymbol{f}(\boldsymbol{x}),\bar{y}) = \frac{2^{K-2}}{2^{K-1}-1}\sum_{j=1,j\neq\bar{y}}^{K}\ell_j(\boldsymbol{x}) + \frac{2^{K-2}-1}{2^{K-1}-1}\sum_{j=1,j\neq\bar{y}}^{K}\bar{\ell}_j(\boldsymbol{x}) + \bar{\ell}_{\bar{y}}(\boldsymbol{x}).$$

Proof. According to the definition of $R(\boldsymbol{f})$, we have

$$\begin{aligned}
R(\boldsymbol{f}) &= \mathbb{E}_{p(\boldsymbol{x},Y)}\left[\mathcal{L}(\boldsymbol{f}(\boldsymbol{x}),Y)\right] \\
&= \int_{\mathcal{X}} \sum_{Y\in\mathcal{Y}'} \mathcal{L}(\boldsymbol{f}(\boldsymbol{x}),Y)p(\boldsymbol{x},Y)\mathrm{d}\boldsymbol{x} \\
&= \int_{\mathcal{X}} \sum_{Y\in\mathcal{Y}'} \sum_{\bar{y}=1,\bar{y}\notin Y}^{K} \mathcal{L}(\boldsymbol{f}(\boldsymbol{x}),Y)\bar{p}(\boldsymbol{x},\bar{y})\mathrm{d}\boldsymbol{x} \\
&= \frac{1}{2^{K-1}-1}\int_{\mathcal{X}} \sum_{\bar{y}=1}^{K} \sum_{Y\in\mathcal{Y}',\bar{y}\notin Y} \mathcal{L}(\boldsymbol{f}(\boldsymbol{x}),Y)\bar{p}(\boldsymbol{x},\bar{y})\mathrm{d}\boldsymbol{x} \\
&= \frac{1}{2^{K-1}-1}\mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}\left[\sum_{Y\in\mathcal{Y}',\bar{y}\notin Y} \mathcal{L}(\boldsymbol{f}(\boldsymbol{x}),Y)\right] \\
&= \frac{1}{2^{K-1}-1}\mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}\left[\sum_{Y\in\mathcal{Y}',\bar{y}\notin Y}\left\{\sum_{j=1,j\in Y}^{K}\ell_j(\boldsymbol{x}) + \sum_{j=1,j\notin Y}^{K}\bar{\ell}_j(\boldsymbol{x})\right\}\right] \\
&= \frac{1}{2^{K-1}-1}\mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}\left[\sum_{Y\in\mathcal{Y}',\bar{y}\notin Y}\left\{\sum_{j=1,j\neq\bar{y},j\in Y}^{K}\ell_j(\boldsymbol{x}) + \sum_{j=1,j\neq\bar{y},j\notin Y}^{K}\bar{\ell}_j(\boldsymbol{x}) + \sum_{j=1,j=\bar{y},j\notin Y}^{K}\bar{\ell}_j(\boldsymbol{x})\right\}\right] \\
&= \frac{1}{2^{K-1}-1}\mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}\left[\sum_{j=1,j\neq\bar{y}}^{K}\sum_{Y\in\mathcal{Y}',j\in Y,\bar{y}\notin Y}\ell_j(\boldsymbol{x}) + \sum_{j=1,j\neq\bar{y}}^{K}\sum_{Y\in\mathcal{Y}',j\notin Y,\bar{y}\notin Y}\bar{\ell}_j(\boldsymbol{x}) + \sum_{Y\in\mathcal{Y}',\bar{y}\notin Y}\bar{\ell}_{\bar{y}}(\boldsymbol{x})\right]
\end{aligned}$$

$$= \frac{1}{2^{K-1}-1}\mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}\left[\sum_{j=1,j\neq\bar{y}}^{K} 2^{K-2}\ell_j(\boldsymbol{x}) + \sum_{j=1,j\neq\bar{y}}^{K}(2^{K-2}-1)\bar{\ell}_j(\boldsymbol{x}) + (2^{K-1}-1)\bar{\ell}_{\bar{y}}(\boldsymbol{x})\right]$$
$$= \mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}\left[\frac{2^{K-2}}{2^{K-1}-1}\sum_{j=1,j\neq\bar{y}}^{K}\ell_j(\boldsymbol{x}) + \frac{2^{K-2}-1}{2^{K-1}-1}\sum_{j=1,j\neq\bar{y}}^{K}\bar{\ell}_j(\boldsymbol{x}) + \bar{\ell}_{\bar{y}}(\boldsymbol{x})\right]$$
$$= \mathbb{E}_{\bar{p}(\boldsymbol{x},\bar{y})}[\bar{\mathcal{L}}(\boldsymbol{f}(\boldsymbol{x}),\bar{y})] = \bar{R}(\boldsymbol{f}).$$

□

C The Proof of Theorem 3

We start investigating an estimation error bound from the following two lemmas.

Lemma 2. *Suppose* $M = \sup_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{f}\in\mathcal{F}}\bar{\mathcal{L}}(\boldsymbol{f}(\boldsymbol{x}),\bar{y})$, $\mathcal{H} = \{h : (\boldsymbol{x},\bar{y}) \mapsto \bar{\mathcal{L}}(\boldsymbol{f}(\boldsymbol{x}),\bar{y}) | \boldsymbol{f} \in \mathcal{F}\}$ *is a class of measurable functions. For any* $\delta > 0$*, we are with probability at least* $1-\delta$*,*

$$\sup_{\boldsymbol{f}\in\mathcal{F}}|\bar{R}_n(\boldsymbol{f}) - \bar{R}(\boldsymbol{f})| \leq 2\mathfrak{R}_n(\mathcal{H}) + \frac{M}{2}\sqrt{\frac{\log 2/\delta}{2n}},$$

where $\mathfrak{R}_n(\mathcal{H}) = \mathbb{E}_{\boldsymbol{x},\bar{y},\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h(\boldsymbol{x}_i,\bar{y}_i)\right]$ *is the expected Rademacher complexity of* $\mathcal{H}$*, in which* $\boldsymbol{\sigma} = \{\sigma_1,\ldots,\sigma_n\}$ *are* n *Rademacher variables.*

Proof. To proof this lemma, we show that the single direction $\sup_{\boldsymbol{f}\in\mathcal{F}}(\bar{R}_n(\boldsymbol{f}) - \bar{R}(\boldsymbol{f}))$ is bounded with the probability $1-\sigma/2$. Based on $\bar{\mathcal{L}}(\boldsymbol{f}(\boldsymbol{x}),\bar{y})$, suppose an instance $(\boldsymbol{x}_i,\bar{y}_i)$ is replaced by an arbitrary instance $(\boldsymbol{x}_i',\bar{y}_i')$, whose change is no greater than $M/2n$. By applying McDiarmid's inequality, for any $\delta > 0$, with the probability at least $1-\delta/2$, we have

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

By sysmmetrization, we have

$$\mathbb{E}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\left(\bar{R}_n(\boldsymbol{f}) - \bar{R}(\boldsymbol{f})\right)\right] \leq 2\mathfrak{R}_n(\mathcal{H}).$$

□

Then, we give an upper bound of $\mathfrak{R}_n(\mathcal{H})$.

Lemma 3. *For any* $j \in \mathcal{Y}$*, assuming* $\ell_j(\boldsymbol{x})$ *and* $\bar{\ell}_j(\boldsymbol{x})$ *are* β^+*-Lipschitz and* β^-*-Lipschitz with respect to* $\boldsymbol{f}(\boldsymbol{x})$ *respectively, then we have*

$$\mathfrak{R}_n(\mathcal{H}) \leq \sqrt{2}\left[\frac{(K-1)2^{K-2}}{2^{K-1}-1}\beta^+ + \frac{(K-1)2^{K-2}-K}{2^{K-1}-1}\beta^-\right]\sum_{j=1}^{K}\mathfrak{R}_n(\mathcal{G}_j),$$

where $\mathcal{G}_j = \{g : \boldsymbol{x} \mapsto f_j(\boldsymbol{x}) | \boldsymbol{f} \in \mathcal{F}\}$ *and* $\mathfrak{R}_n(\mathcal{G}_j) = \mathbb{E}_{\boldsymbol{x},\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}_j}\frac{1}{n}\sum_{i=1}^{n}g(\boldsymbol{x}_i)\right]$.

Proof. Firstly, suppose $\ell\circ\mathcal{F}$ and $\bar{\ell}\circ\mathcal{F}$ denote $\{\ell\circ\mathcal{F}|\boldsymbol{f}\in\mathcal{F}\}$ and $\{\bar{\ell}\circ\mathcal{F}|\boldsymbol{f}\in\mathcal{F}\}$ respectively. Then we apply the Rademacher vector contraction inequality [Maurer, 2016]:

$$\mathfrak{R}_n(\mathcal{H}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i h(\boldsymbol{x}_i,\bar{y}_i)\right]$$
$$= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\bar{\mathcal{L}}(\boldsymbol{f}(\boldsymbol{x}_i),\bar{y}_i)\right]$$
$$= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left\{\frac{2^{K-2}}{2^{K-1}-1}\sum_{j=1,j\neq\bar{y}_i}^{K}\ell_j(\boldsymbol{x}_i) + \frac{2^{K-2}-1}{2^{K-1}-1}\sum_{j=1,j\neq\bar{y}_i}^{K}\bar{\ell}_j(\boldsymbol{x}_i) + \bar{\ell}_{\bar{y}_i}(\boldsymbol{x}_i)\right\}\right]$$

$$
\begin{aligned}
&\leq \frac{2^{K-2}}{2^{K-1}-1}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\sum_{j=1,j\neq\bar{y}_i}^{K}\ell_j(\boldsymbol{x}_i)\right]+\frac{2^{K-2}-1}{2^{K-1}-1}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\sum_{j=1,j\neq\bar{y}_i}^{K}\bar{\ell}_j(\boldsymbol{x}_i)\right]+\\
&\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\bar{\ell}_{\bar{y}_i}(\boldsymbol{x}_i)\right]\\
&\leq \frac{2^{K-2}}{2^{K-1}-1}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\sum_{j=1}^{K}\ell_j(\boldsymbol{x}_i)\right]-\frac{2^{K-2}}{2^{K-1}-1}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell_{\bar{y}_i}(\boldsymbol{x}_i)\right]\\
&+\frac{2^{K-2}-1}{2^{K-1}-1}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\sum_{j=1}^{K}\bar{\ell}_j(\boldsymbol{x}_i)\right]-\frac{2^{K-2}-1}{2^{K-1}-1}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\bar{\ell}_{\bar{y}_i}(\boldsymbol{x}_i)\right]+\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\bar{\ell}_{\bar{y}_i}(\boldsymbol{x}_i)\right]\\
&\leq \frac{2^{K-2}}{2^{K-1}-1}\sum_{j=1}^{K}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell_j(\boldsymbol{x}_i)\right]-\frac{2^{K-2}}{2^{K-1}-1}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell_{\bar{y}_i}(\boldsymbol{x}_i)\right]\\
&+\frac{2^{K-2}-1}{2^{K-1}-1}\sum_{j=1}^{K}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\bar{\ell}_j(\boldsymbol{x}_i)\right]+\frac{2^{K-2}}{2^{K-1}-1}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{f}\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\bar{\ell}_{\bar{y}_i}(\boldsymbol{x}_i)\right]\\
&\leq \frac{2^{K-2}}{2^{K-1}-1}\sum_{j=1}^{K}\mathfrak{R}_n(\ell\circ\mathcal{F})-\frac{2^{K-2}}{2^{K-1}-1}\mathfrak{R}_n(\ell\circ\mathcal{F})+\frac{2^{K-2}-1}{2^{K-1}-1}\sum_{j=1}^{K}\mathfrak{R}_n(\bar{\ell}\circ\mathcal{F})+\frac{2^{K-2}}{2^{K-1}-1}\mathfrak{R}_n(\bar{\ell}\circ\mathcal{F})\\
&=\frac{2^{K-2}(K-1)}{2^{K-1}-1}\mathfrak{R}_n(\ell\circ\mathcal{F})+\frac{2^{K-2}(K-1)-K}{2^{K-1}-1}\mathfrak{R}_n(\bar{\ell}\circ\mathcal{F})\\
&\leq\frac{2^{K-2}(K-1)}{2^{K-1}-1}\sqrt{2}\beta^{+}\sum_{j=1}^{K}\mathfrak{R}_n(\mathcal{G}_j)+\frac{2^{K-2}(K-1)-K}{2^{K-1}-1}\sqrt{2}\beta^{-}\sum_{j=1}^{K}\mathfrak{R}_n(\mathcal{G}_j)\\
&=\sqrt{2}\left[\frac{2^{K-2}(K-1)}{2^{K-1}-1}\beta^{+}+\frac{2^{K-2}(K-1)-K}{2^{K-1}-1}\beta^{-}\right]\sum_{j=1}^{K}\mathfrak{R}_n(\mathcal{G}_j).
\end{aligned}
$$

□

Based on Lemma 2 and 3, we can derive the estimation error bound.

Theorem 3. *Suppose* $M=\sup_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{f}\in\mathcal{F}}\bar{\mathcal{L}}(\boldsymbol{f}(\boldsymbol{x}),\bar{y})$. *For any* $j\in\mathcal{Y}$, *assuming* $\ell_j(\boldsymbol{x})$ *and* $\bar{\ell}_j(\boldsymbol{x})$ *are* β^{+}*-Lipschitz and* β^{-}*-Lipschitz with respect to* $\boldsymbol{f}(\boldsymbol{x})$ *respectively. For any* δ, *with the probability at least* $1-\delta$,

$$
R(\boldsymbol{f}_n)-R(\boldsymbol{f}^*)\leq M\sqrt{\frac{\log 2/\delta}{2n}}+4\sqrt{2}\left[\frac{(K-1)2^{K-2}}{2^{K-1}-1}\beta^{+}+\frac{(K-1)2^{K-2}-K}{2^{K-1}-1}\beta^{-}\right]\sum_{j=1}^{K}\mathfrak{R}_n(\mathcal{G}_j).
$$

Proof. Combining Theorem 3, Lemma 2 and 3, the following inequality holds:

$$
\begin{aligned}
R(\boldsymbol{f}_n)-R(\boldsymbol{f}^*)&=\bar{R}(\boldsymbol{f}_n)-\bar{R}(\boldsymbol{f}^*)\\
&=[\bar{R}(\boldsymbol{f}_n)-\bar{R}_n(\boldsymbol{f}^*)]+[\bar{R}_n(\boldsymbol{f}_n)-\bar{R}_n(\boldsymbol{f}^*)]+[\bar{R}_n(\boldsymbol{f}^*)-\bar{R}(\boldsymbol{f}^*)]\\
&\leq\bar{R}(\boldsymbol{f}_n)-\bar{R}_n(\boldsymbol{f}_n)+\bar{R}_n(\boldsymbol{f}^*)-\bar{R}(\boldsymbol{f}^*)\\
&=2\sup_{\boldsymbol{f}\in\mathcal{F}}|\bar{R}_n(\boldsymbol{f})-\bar{R}(\boldsymbol{f})|\\
&\leq M\sqrt{\frac{\log 2/\delta}{2n}}+4\sqrt{2}\left[\frac{(K-1)2^{K-2}}{2^{K-1}-1}\beta^{+}+\frac{(K-1)2^{K-2}-K}{2^{K-1}-1}\beta^{-}\right]\sum_{j=1}^{K}\mathfrak{R}_n(\mathcal{G}_j).
\end{aligned}
$$

□

Table 3: Characteristics of datasets. $dim(\mathcal{S})$, $|\mathcal{S}|$ and $L(\mathcal{S})$ refer to the number of features, instances and labels respectively. $LCard(\mathcal{S})$ and avg. #CL denote the average number of labels and complemenatry labels per instance respectively.

Datasets	$dim(\mathcal{S})$	$\|\mathcal{S}\|$	$L(\mathcal{S})$	$LCard(\mathcal{S})$	avg. #CL
scene	294	2407	6	1.07	1
yeast	103	2417	14	4.23	1
bookmark	2150	38912	15	1.25	1
mediamill	120	41701	15	3.63	1
eurlex_dc	100	8636	15	1.02	1
		19340	410	1.29	205
eurlex_sm	100	13270	15	1.74	1
		19338	201	2.21	100
Corel16k	500	11103	15	1.77	1
		13766	153	2.87	77
delicious	500	14784	15	4.32	1
		16105	983	19.02	492

D Details of Experiments

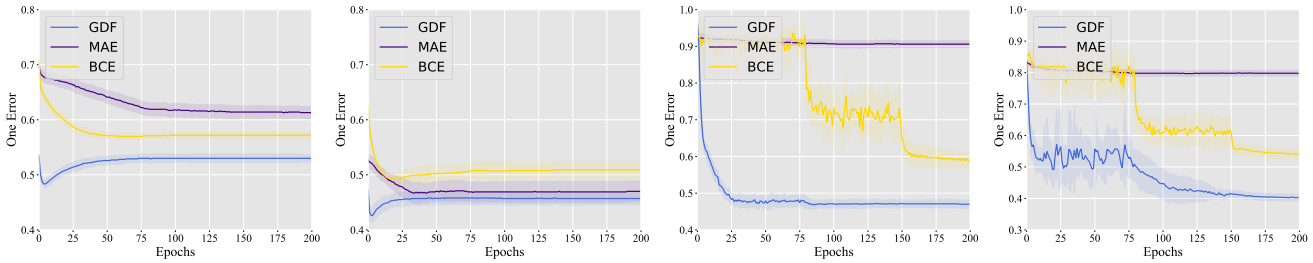
Details of datasets are presented in Table 3, including the number of features ($dim(\mathcal{S})$), the number of instances ($|\mathcal{S}|$), the number of labels ($L(\mathcal{S})$), the average number of relevant labels per instance ($LCard(\mathcal{S})$), and the average number of complementary labels per instance (avg. #CL). Besides, Table 4 and 5 are shown results of various approaches on different datasets that are applied the first pre-processing way and second pre-processing way to process respectively. As can be seen from Table 4 and 5, our approach can perform well in most cases. Accordingly, we draw the curves of our empirical risk estimator with MAE loss and BCE loss, and GDF loss on *one error*, *coverage*, and *ranking loss*. Due to the results of *hamming loss* relying on the threshold value, its curve does not be presented here. From Fig. 3, GDF achieves a good performance in most cases, which demonstrates the superiority of our proposed approach.

Table 4: *Hamming loss* and *coverage* (mean±std) on training data with first pre-processing way (each instance is given one complementary label and labels are kept under 15). The best performance of each dataset is presented in **boldface**, where •/○ denotes whether GDF is superior/inferior to baselines with pairwise t-test (at 0.05 significance level).

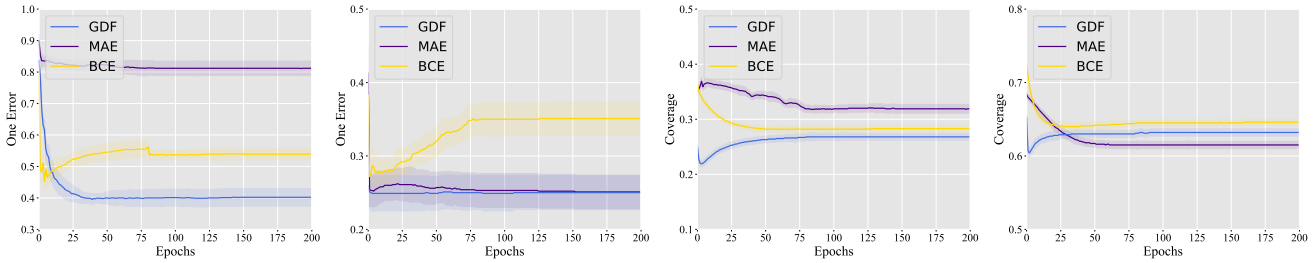
Methods	MLL		PML		CLL	Unbounded	Bounded	GDF
	ML-KNN	LIFT	fpml	PML-LRS	L-UW	BCE	MAE	
Hamming Loss↓								
bookmark	.917±.001•	.916±.001•	.420±.009•	.813±.003•	.295±.007•	**.147±.002**	.237±.006•	.149±.004
Corel16k	.882±.008•	.882±.008•	.882±.008•	.862±.001•	.365±.021•	.240±.006•	.200±.013	**.196±.019**
delicious	.711±.003•	.711±.003•	.712±.003•	.459±.002•	.331±.005•	.300±.002•	.285±.004	**.283±.003**
mediamill	.758±.011•	.758±.011•	.757±.011•	.760±.000•	.185±.016	**.173±.011**	.179±.016	.180±.011
eurlex_dc	.932±.000•	.932±.000•	.178±.012○	.777±.008•	.541±.014•	**.136±.004**○	.513±.013•	.225±.009
eurlex_sm	.883±.001•	.883±.001•	.178±.005○	.711±.002•	.533±.014•	**.176±.012**○	.500±.006•	.211±.004
scene	.821±.002•	.820±.003•	.819±.002•	.814±.000•	.526±.016•	.203±.014	.372±.009•	**.202±.010**
yeast	.697±.012•	.697±.012•	.697±.012•	.316±.000•	.287±.010•	.247±.008•	.240±.008•	**.231±.007**
Coverage↓								
bookmark	.359±.006•	.328±.008•	.474±.018•	.280±.004•	.219±.007	.281±.006•	.318±.008•	**.219±.006**
Corel16k	.430±.042	.487±.026•	.513±.034•	**.404±.008**	.405±.037	.493±.020•	.420±.031	.410±.039
delicious	.712±.006•	.703±.004•	.726±.009•	.609±.003	.634±.006•	.641±.005•	.615±.006•	**.605±.007**
mediamill	.503±.022•	.501±.023•	.512±.034•	**.436±.002**○	.483±.033	.449±.022	.494±.034•	.463±.023
eurlex_dc	.266±.007•	.277±.011•	.441±.035•	.170±.005•	.252±.014•	.205±.016•	.411±.008•	**.153±.005**
eurlex_sm	.411±.008•	.421±.012•	.552±.031•	.336±.005•	.423±.008•	.374±.017•	.546±.007•	**.308±.008**
scene	.299±.025•	.256±.016•	.434±.02•	.230±.006•	.397±.015•	.169±.017•	.425±.019•	**.144±.008**
yeast	.579±.017•	.649±.019•	.553±.031	.742±.025•	.585±.017•	.582±.024•	.567±.018•	**.540±.020**

Table 5: Experimental results (mean±std) on the training data adopted the second pre-processing way. The label space is original and the number of complementary labels per instance is half the original number of labels. The best performance of each dataset is shown in **boldface**, where •/○ indicates whether GDF is superior/inferior to baselines with pairwise t-test (at 0.05 significance level).

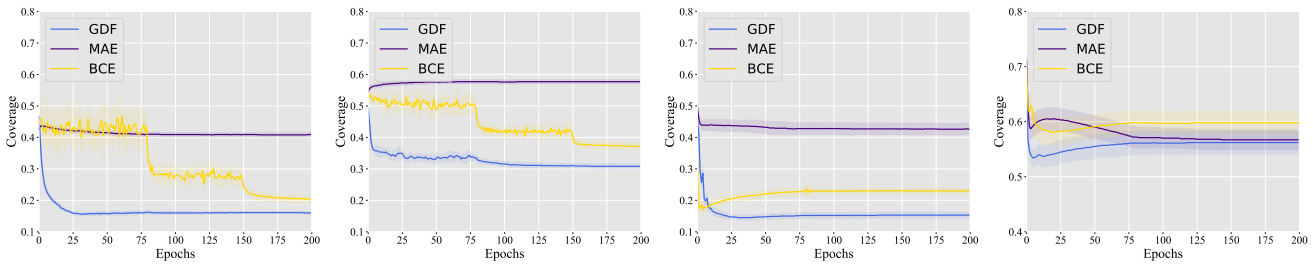
Methods	MLL	PML		CLL	Bounded	GDF
	ML-KNN	fpml	PML-LRS	L-UW	MAE	
Hamming Loss↓						
Corel16k	.299±.004•	.195±.011•	.452±.003•	.036±.002•	.060±.003•	.029±.003
delicious	.391±.003•	.262±.009•	.202±.001•	.025±.000•	.038±.001•	.019±.000
eurlex_dc	.474±.007•	.091±.005•	.136±.003•	.271±.015•	.196±.006•	.020±.001
eurlex_sm	.483±.008•	.066±.005•	.146±.002•	.312±.019•	.233±.006•	.019±.001
One Error↓						
Corel16k	.789±.050•	.770±.052	.721±.014	.769±.053	.742±.048	.736±.053
delicious	.541±.018•	.593±.015•	.455±.012•	.526±.006•	.510±.009•	.383±.014
eurlex_dc	.716±.011○	.898±.011•	.722±.002	.804±.007•	.815±.009•	.728±.009
eurlex_sm	.535±.011•	.778±.007•	.525±.003	.599±.009•	.660±.009•	.523±.007
Coverage↓						
Corel16k	.507±.070	.458±.091	.519±.019	.451±.090	.449±.088	.519±.061
delicious	.845±.004•	.803±.008•	.803±.011•	.753±.005○	.746±.005○	.767±.003
eurlex_dc	.310±.005•	.305±.006•	.296±.004•	.293±.006•	.301±.010•	.283±.006
eurlex_sm	.285±.007•	.325±.007•	.278±.004•	.287±.006•	.273±.006•	.268±.008



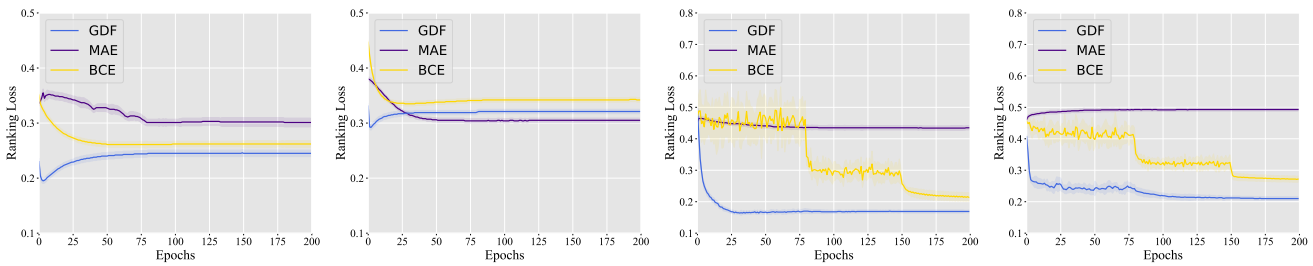
(a) *One error* on bookmark (b) *One error* on delicious (c) *One error* on eurlex_dc (d) *One error* on eurlex_sm



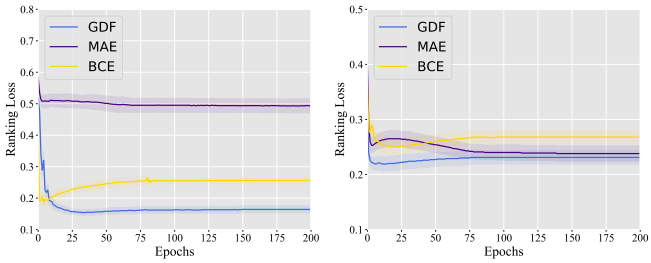
(e) *One error* on scene (f) *One error* on yeast (g) *Coverage* on bookmark (h) *Coverage* on delicious



(i) *Coverage* on eurlex_dc (j) *Coverage* on eurlex_sm (k) *Coverage* on scene (l) *Coverage* on yeast



(m) *Ranking Loss* on bookmark (n) *Ranking Loss* on delicious (o) *Ranking Loss* on eurlex_dc (p) *Ranking Loss* on eurlex_sm



(q) *Ranking Loss* on scene (r) *Ranking Loss* on yeast

Figure 3: *One error*, *coverage*, and *ranking loss* on various datasets adopted the first pre-processing way, where each instance is associated with one complementary label and labels are kept under 15. Dark colors show the mean of testing average precision and light colors are corresponding to the std.