Towards Robust Multi-Label Learning against Dirty Label Noise

Yuhai Zhao[1,2] , **Yejiang Wang**[1,2*] , **Zhengkui Wang**[3] , **Wen Shan**[4]
Miaomiao Huang[1,2] , **Meixia Wang**[1,2] , **Min Huang**[5] , **Xingwei Wang**[1]
[1]School of Computer Science and Engineering, Northeastern University, China
[2]Key Laboratory of Intelligent Computing in Medical Image
of Ministry of Education, Northeastern University, China
[3]InfoComm Technology Cluster, Singapore Institute of Technology, Singapore
[4]Singapore University of Social Sciences, Singapore
[5]College of Information Science and Engineering, Northeastern University, China
zhaoyuhai@mail.neu.edu.cn, wyejiang@gmail.com, zhengkui.wang@singaporetech.edu.sg

Abstract

In multi-label learning, one of the major challenges is that the data are associated with label noise including the random noisy labels (e.g., data encoding errors) and noisy labels created by annotators (e.g., missing, extra, or error label), where noise is promoted by different structures (e.g., gaussian, sparse, or subjective). Existing methods are tailored to handle noise with one specific structure. However, they lack of consideration of the fact that the data are always with *dirty* noisy labels, *simutaneously* gaussian, sparse, and subjective, in real applications. In this paper, we formalize the multi-label learning with dirty noise as a new learning problem, namely *Noisy Multi-label Learning* (NML). To solve the NML problem, we decompose a corrupted label matrix as the noise matrix plus a true label matrix (maybe high-rank). For the noise matrix, a mixed norm penalty is developed as regularizer for dirty noise distribution. Under this norm, the conditions required for *exact* noise recovery are provided theoretically. For the true label matrix that is not necessarily low-rank, we apply a non-linear mapping to ensure its low-rankness such that the high-order label correlation can be utilized. Experimental results show that the proposed method outperforms the state-of-the-art methods significantly.

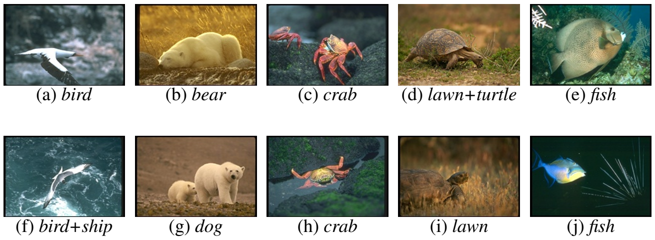
(a) *bird* (b) *bear* (c) *crab* (d) *lawn+turtle* (e) *fish*
(f) *bird+ship* (g) *dog* (h) *crab* (i) *lawn* (j) *fish*

Figure 1: Examples of labels with/without noise label in corel5k.

1 Introduction

Multi-label learning (MLL), which involves training on instances with multiple labels, is increasingly researched across various fields [Zhang *et al.*, 2022; Xu *et al.*, 2023; Lin, 2023]. Traditional MLL relies on accurately labeled data; however, real-world datasets often contain noisy labels that significantly impair learning. There are two main types of noise in these datasets. Category 1 is the partial random noise incurred by human experts or automated classification tools. For example, this type of noise may be generated by the annotators due to the ambiguous contents (missing label, e.g., turtle in Fig.1(i) is obscured by the lawn), complicated structure of objects (extra labels, e.g., ship in Fig.1(f) shouldn't exist) or a combination of the above two (missing and extra labels, e.g., bear in Fig.1(g) is mislabeled as dog, where bear label is missing and dog label is extra, also known as error labels). The missing label noise can be regarded as negative noise, and the extra label noise as positive noise. This kind of artificial noise may be *sparse* (e.g., only three out of ten pictures are mislabeled in Fig.1) or specific to some labels, named *subjective* (e.g., crab is easier to distinguish than bird in Fig.1). Category 2 refers to the completely random noise by encoding errors or communication problems introduced by sensors in fields such as electrical engineering, signal processing, or hardware sensor calibration [Brodley and Friedl, 1999; Zhu and Wu, 2004], which is often *gaussian*. The instance and the label matrices are located in the same data packet, it implies that the label may be affected by real-valued noise (e.g., gaussian) during propagation just like feature instance. Examples of these noises (gaussian, subjective and sparse) reflected in noise matrices for 8 labels and 4 instances are shown in Figs.2(a), (b) and (c), respectively.

Recent research has made effort in handling the data noise in the learning model. For example, multi-label learning with missing labels (MLML) [Wu *et al.*, 2014] trains the model for the case of missing labels (negative noise). Partial multi-label learning (PML) [Xie and Huang, 2018] considers the training samples with several extra distractor labels (positive noise). [Sun *et al.*, 2021] handles both missing and extra labels. These works assume that the true noise only obeys a

[*] Corresponding author.

gaussian or laplace distribution (sparse) and is recovered by the ℓ_2- or ℓ_1-norm (to denoise), respectively. However, the noise distributions are intricate in real applications. For example, the noise may depend on some specific classes or obey multiple distributions simultaneously, and we define this kind of noise as the **dirty noise**. As shown in Fig.1, it is very common that the real datasets contain a dirty label noise, satisfying two numerical conditions (i.e., negative and positive) and multiple structural requirements (i.e., sparse, subjective and gaussian), simultaneously. A representation of the dirty noise on the noise matrix is shown in Fig.2(d). Note that the dirty noise is a different type of noise from a simple summation of the previous noises. Thus, it cannot be solved by simply combining existing methods. This calls for new technologies to handle the dirty noise that exists in real-world datasets.

We formalize this problem as a new learning framework called Noisy Multi-label Learning (NML). NML is a novel learning framework with significant difference from existing settings. More specifically, NML tries to learn a multi-label model from the training datasets with dirty label noise, where each instance in the dataset is assigned with multiple labels and the labeled datasets contain the label noise obeyed 3 noise distributions (i.e. gaussian, sparsity and subjective), simultaneously. And the label noise can be a real number, not just positive or negative alone.

To tackle the NML problem, we propose a *robust Noisy Multi-label Learning against Dirty label noise* (NMLD) approach, which learns a multi-label model from the dataset with dirty label noise using the idea of "*simple but contracted* ": while any one structure penalty might not capture the noise, a superposition of structural classes might. Specifically, we decompose the corrupted label matrix as a noise matrix plus a true label matrix (maybe high-rank), and use a mixed norm penalty of noise matrix based on 3 noise distributions. We provide theoretical guarantee for *exact noise recovery* from such a structured corruption. To more accurately identify the true label matrix, we apply a nonlinear mapping to it to ensure that the mapped matrix is of low-rank such that the high-order label correlation can be utilized. A novel accelerated proximal alternating algorithm is developed, and the problem is jointly optimized for the noise and true label matrix. To summarize, our contributions are:

- A new framework NML is proposed to learn multi-label models from noisy labeled data where the noise is promoted by various different types of noise patterns.
- An algorithm NMLD is proposed for solving the NML problem. NMLD unifies a mixed penalty on noise matrix and true matrix exploration in a unified objective. Theoretical conditions for exact noise recovery are provided.
- Experiments demonstrate that NMLD is effective on noisy data and with missing or extra labels.

2 Related Work

To mitigate the label noise issues in MLL, weakly supervision learning have been proven to be practical [Natarajan *et al.*, 2013; Wang *et al.*, 2023; Van Rooyen and Williamson, 2017; Xie *et al.*, 2022; Xie *et al.*, 2023; Wang *et al.*, 2024a; Wang *et al.*, 2024b]. Multi-label learning with missing labels (MLML) is a learning task that conducts the learning over the data with partial set of positive labels and missing labels. Most of the MLML methods attempt to complete the missing labels first, and then train the classifiers with the complete labels [Wu *et al.*, 2014].

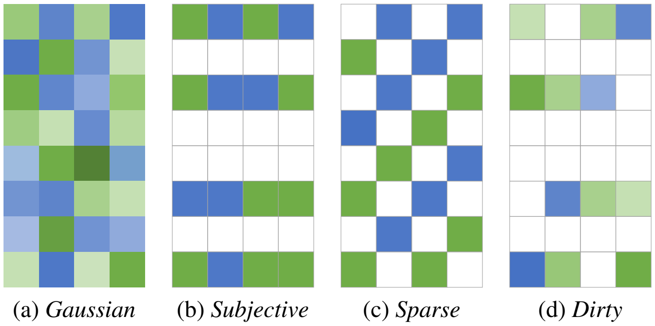
(a) *Gaussian* (b) *Subjective* (c) *Sparse* (d) *Dirty*

Figure 2: Different types of label noise matrices $\mathbf{E}$, where blue and green are the negative and positive noises, respectively. The darker the color, the larger the absolute value.

Partial multi-label learning deals with training instances annotated with a candidate label set that contains true labels and extra labels. To identify the extra labels, [Xie and Huang, 2018] first utilizes the label confidence to measure the probability of being the true labels for each candidate label, and obtains the true labels according to label ranking. CORALS [Sun *et al.*, 2021] propose to deal with the missing and extra labels simultaneously, where a label confidence matrix is constructed and the ordering of labels are optimized.

However, these methods consider one specific label noise only, which lacks of the capability of handling common possible type of dirty label noise. Therefore, they can not solve the NML problem that widely exist in real-world datasets.

3 Methodology

3.1 NML Definition

Let $\mathbf{X} \in \mathbb{R}^{d\times n}$ be the feature matrix for n instances in the d-dimensional feature space. The corresponding ground-truth label of $\mathbf{X}$ is denoted as matrix $\mathbf{Y} \in \{1,-1\}^{q\times n}$, where each column corresponds to an instance and each row corresponds to a label. If $\mathbf{x}_i$ is associated with the c-th label, $[\mathbf{Y}]_{ci} = 1$; otherwise, $[\mathbf{Y}]_{ci} = -1$. And let $\tilde{\mathbf{Y}} \in \mathbf{\Gamma}^{q\times n}$ denote the corrupted label matrix with dirty label noise, which satisfy two numerical conditions (i.e. negative and positive) and multiple structural requirements (i.e. sparse, subjective and gaussian), simultaneously. Generally, $\mathbf{\Gamma} := \mathbb{R}$ due to the existence of noise. The *goal* of NML is to identify noisy labels of corrupted label matrix $\tilde{\mathbf{Y}}$ and train a multi-label classifier $f : \mathbb{R}^{d\times n} \to \{1,-1\}^{q\times n}$ from $\mathbf{X}$.

3.2 NMLD Algorithm

To deal with the NML problem, in this paper, we propose a robust noisy multi-label learning method for dirty noise NMLD. Specifically, we decompose the corrupted label matrix as a noise matrix plus a true label matrix (maybe high-rank), and use a mixed norm penalty of noise matrix for different noise distribution to fit different types of noise. To improve the generalization ability, we apply a non-linear mapping on true label matrix to ensure the low-rankness such that the high-order label correlation can be easily utilized. To solve the resulting objective, we develop an optimization method to minimize the loss by alternating restricted minimization over the subsets of variables.

Our goal is to use the instance matrix $\mathbf{X}$ and the corrupted label matrix $\tilde{\mathbf{Y}}$ for training a new NML model and to predict the labels for unlabeled data. To solve this problem, we assume the linear regression prediction model. Therefore, we can optimize the matrix $\mathbf{W} \in \mathbb{R}^{q \times d}$ by minimizing the square loss with the ℓ_1-norm regularization

$$\min_{\mathbf{W}} \frac{1}{2}\|\tilde{\mathbf{Y}} - \mathbf{W}\mathbf{X}\|_F^2 + \alpha\|\mathbf{W}\|_1 \tag{1}$$

where $\|\cdot\|_F$ and $\|\cdot\|_1$ are Frobenius- and ℓ_1-norm, respectively, and α is the trade-off parameter.

However, the corrupted label matrix $\tilde{\mathbf{Y}}$ often includes label noise in noisy multi-label learning. Many weakly supervised MLL methods assume the label matrix $\tilde{\mathbf{Y}}$ is low-rank and use nuclear norm regularization to recover the true label matrix. Unfortunately, the low-rank assumption does not necessarily hold. For example, if each sample has a distinct label, the rank of the label matrix is $\mathrm{rank}(\tilde{\mathbf{Y}}) = \min\{q, n\}$. Obviously, the low-rank assumption fails when both n and c are large. In addition, the presence of noise may increase the rank of $\tilde{\mathbf{Y}}$. To deal with this problem, we decompose the corrupted label matrix $\tilde{\mathbf{Y}}$ into a true label matrix $\mathbf{Y}$ and a noise label matrix $\mathbf{E}$, i.e. $\tilde{\mathbf{Y}} = \mathbf{Y} + \mathbf{E}$. Motivated by the kernel principal component analysis (KPCA) method, we use a smooth nonlinear function ϕ to map the the columns of label matrix $\mathbf{Y}$ into a high-dimensional (possibly infinite) space such that $\phi(\mathbf{Y}) \in \mathbb{R}^{\kappa \times n}$ is exactly or approximately of low-rank [Schölkopf *et al.*, 1997; Fan and Chow, 2019; Ma *et al.*, 2019]. Thus, a low-rank constraint on $\phi(\mathbf{Y})$ is utilized to further exploit high-order label correlations [Xu *et al.*, 2014]

$$\min_{\mathbf{Q},\mathbf{P},\mathbf{E}} \frac{1}{2}\|\Upsilon - \mathbf{\Theta}\mathbf{P}\|_F^2 + \beta\|\mathbf{\Theta}\|_2 + \gamma\|\mathbf{P}\|_* + \Omega(\mathbf{E}) \tag{2}$$

to mitigate the impact of noise and strengthen the generalization performance, where β and γ are balancing parameters, $\|\cdot\|_2$ denotes the ℓ_2-norm, $\mathbf{\Theta} = \phi(\mathbf{Q}) \in \mathbb{R}^{\kappa \times r}$, and $\Upsilon = \phi(\mathbf{Y}) = \phi(\tilde{\mathbf{Y}} - \mathbf{E})$. We find that the RBF kernel mapping is more effective than polynomial kernel. Hence, we mainly focus on RBF mapping with parameter σ. To capture the rank property of the mapped label matrix, the trace norm $\|\cdot\|_*$ is employed to introduce the low-rank assumption.

In this work, we aims to handle the dirty label noise (Fig.2(d)), which is promoted by the different noise priors: *Gaussian* (Fig.2(a)), *Sparsity* (Fig.2(c)) and *Subjective* (Fig.2(b)), simultaneously. As shown in Fig.2, the distribution of the dirty noise is completely different from any of the individual noise patterns, rather than a simple combination of them. To identify the dirty label noise, we use a *simple but interesting* idea: while any one structure constraint might not capture the dirty noise, a superposition of structural constraint classes might

$$\Omega(\mathbf{E}) = \eta_1\|\mathbf{E}\|_2 + \eta_2\|\mathbf{E}\|_1 + \eta_3\|\mathbf{E}\|_{2,1} \tag{3}$$

where η_1, η_2 and η_3 are balancing parameters. 1) $\|\mathbf{E}\|_2$ is to capture the gaussian noise which is incurred by data encoding or communication errors; 2) Both $\|\mathbf{E}\|_1$ and $\|\mathbf{E}\|_{2,1}$ are to capture the label noise incurred by annotators (e.g., missing labels, extra labels etc.), where $\|\mathbf{E}\|_1$ is for the case when $\tilde{\mathbf{Y}}$ is partially and randomly corrupted (i.e., $\mathbf{E}$ is a sparse matrix) and $\|\mathbf{E}\|_{2,1}$ is for the case when a few columns of $\tilde{\mathbf{Y}}$ are corrupted by noise. We will prove later that this superposition of norms can identify with dirty label noise.

The unified objective function for NMLD is defined as

$$\begin{aligned}\min_{\mathbf{W},\mathbf{Q},\mathbf{P},\mathbf{E}} \frac{1}{2}\|(\tilde{\mathbf{Y}} - \mathbf{E}) - \mathbf{W}\mathbf{X}\|_F^2 + \frac{1}{2}\|\Upsilon - \mathbf{\Theta}\mathbf{P}\|_F^2 \\ + \alpha\|\mathbf{W}\|_1 + \beta\|\mathbf{\Theta}\|_2 + \gamma\|\mathbf{P}\|_* + \Omega(\mathbf{E})\end{aligned} \tag{4}$$

where α, β, γ are parameters to keep the balance of the model. We introduce the specific noise matrix $\mathbf{E}$ to the first term which helps our method better identify the true labels and eliminate the impact of noisy labels.

Theoretically, we are also interested in providing sufficient condition on the recovering of the true noise label matrix. The following result provides sufficient condition for exactly reconstructing true noise $\mathbf{E}^\dagger$ using Eq.(4).

Theorem 1. *If the true noise matrix* $\mathbf{E}^\dagger$ *(*$>\mathbf{0}$ *for* PML, $<\mathbf{0}$ *for* MLML, *both for* NML*) obeys 3 distribution as described in Figs.2(a), (b) and (c), simultaneously (i.e. Fig.2(d)). In Eq.(4),* $\tilde{\mathbf{Y}} - \mathbf{W}\mathbf{X}$ *in the first term is an observation of the real noise* $\mathbf{E}^\dagger$*, assume that the observation error* $\Delta{=}\tilde{\mathbf{Y}}{-}\mathbf{W}\mathbf{X}{-}\mathbf{E}^\dagger$ *obeys the gaussian distribution. Let* $\eta_1 = \nu_1\sqrt{3nq}/2$, $\eta_2 = \nu_2\sqrt{2\log(q/s_1)}$ *and* $\eta_3 = \nu_3(\sqrt{2\log(m/s_2)} + \sqrt{k})$*, where* $\nu_i \in [0,1]$ *and* $\sum_i \nu_i = 1$, $s_\circ$ *controls the sparsity proportion of* ℓ_1 *and* $\ell_{2,1}$*, respectively.* m *denotes the number of groups for* $\ell_{2,1}$*.* t *denotes arbitrary constant. Then, whenever the number of observations*

$$n_o \geq \left(\frac{3nq - \sqrt{nq}}{4\nu_1^{-1}} + 2\nu_2 s_1\sqrt{\log\frac{q}{s_1}} + 4\nu_3 s_2\sqrt{\log\frac{m}{s_2}} + t\right)^2 + 1,$$

$\mathbf{E}^\dagger$ *can be recovered exactly with probability* $1 - 2e^{-t^2/2}$.

This result shows how the noise recovery is governed by the number of observations for the real noise matrices.

3.3 Optimization

To solve the problem in Eq.(4), we iteratively update $\mathbf{Q}$, $\mathbf{P}$, $\mathbf{E}$ and $\mathbf{W}$. We summarize the key steps in Algorithm 1.

Update Q. When $\mathbf{W}$, $\mathbf{P}$ and $\mathbf{E}$ are fixed, the problem in Eq.(4) w.r.t $\mathbf{Q}$ is

$$\min_{\mathbf{Q}} \|\Upsilon - \mathbf{\Theta}\mathbf{P}\|_F^2/2 + \beta\|\mathbf{\Theta}\|_2 \tag{5}$$

We have $\|\mathbf{\Theta}\|_2 = tr(\phi(\mathbf{Q})^\top\phi(\mathbf{Q})) = r$, which means the norm has no effect on the problem and can be discarded, where $tr(\cdot)$ denotes the trace operator. The minimization of Eq.(5) has no closed-form solution, we optimize $\mathbf{Q}$ via the gradient descent algorithm. The gradient of Eq.(5) is

$$\nabla_{\mathbf{Q}} = \sigma^{-1}(\mathbf{Y}\mathbf{R}_1 - \mathbf{Q}(\check{\mathbf{R}}_1 - 2\mathbf{R}_2 + 2\check{\mathbf{R}}_2))/2 \tag{6}$$

Algorithm 1 NMLD

```
Input: train data $\mathbf{X}$, label matrix $\tilde{\mathbf{Y}}$
Output: $\mathbf{W}$
$\dot{\omega} \leftarrow 1$ and initialize $\dot{\mathbf{W}}, \dot{\mathbf{Q}}, \dot{\mathbf{P}}, \dot{\mathbf{E}}$;
while not converged
    Compute $\nabla_{\mathbf{Q}}$ using Eq.(6);                 % update $\mathbf{Q}$
    $\mathbf{Q} \leftarrow \dot{\mathbf{Q}} - \delta\nabla_{\mathbf{Q}}(\dot{\mathbf{Q}})$;
    Update $\mathbf{P}$ using Eq.(10);                          % update $\mathbf{P}$
    $\dot{\mathbf{P}} \leftarrow \mathbf{P}$;
    Update $\mathbf{E}$ using Eq.(19);                          % update $\mathbf{E}$
    $\dot{\mathbf{E}} \leftarrow \mathbf{E}$;
    Update $\mathbf{W}$ using Eq.(22) and Eq.(23);              % update $\mathbf{W}$
    $\omega \leftarrow 1/2 + \sqrt{4\dot{\omega}^2 + 1}/2$;
    $\ddot{\mathbf{W}} \leftarrow \dot{\mathbf{W}}$, $\dot{\mathbf{W}} \leftarrow \mathbf{W}$ and $\dot{\omega} \leftarrow \omega$;
return $\mathbf{W}$
```

where $\mathbf{R}_1 = -\mathbf{P}^\top \odot \Upsilon^\top\mathbf{\Theta}$, $\mathbf{R}_2 = (1/2\mathbf{P}\mathbf{P}^\top) \odot \mathbf{\Theta}^\top\mathbf{\Theta}$, $\check{\mathbf{R}}_i = \mathrm{diag}(\mathbf{1}^\top\mathbf{R}_i)$ and $\odot$ denotes the Hadamard product. Then we update $\mathbf{Q} \coloneqq \dot{\mathbf{Q}} - \delta\nabla_{\mathbf{Q}}(\dot{\mathbf{Q}})$, where δ is the step size and $\dot{\mathbf{Q}}$ denotes the solution of previous iteration.

Update P. Fixing $\mathbf{W}$, $\mathbf{Q}$ and $\mathbf{E}$, we have

$$\min_{\mathbf{P}} \|\Upsilon - \mathbf{\Theta}\mathbf{P}\|_F^2/2 + \gamma\|\mathbf{P}\|_* \tag{7}$$

Eq.(7) has no closed-form solution, since the presence of kernel mapping. We use the proximal gradient method to update $\mathbf{P}$ [Bauschke and Combettes, 2017]. Let $g(\mathbf{P}) = \gamma\|\mathbf{P}\|_*$ and $f(\mathbf{P}) = \frac{1}{2}\|\Upsilon - \mathbf{\Theta}\mathbf{P}\|_F^2$. $f(\mathbf{P})$ is approximated by its first order expansion at the solution $\dot{\mathbf{P}}$ of previous iteration

$$f(\mathbf{P}) \leq f(\dot{\mathbf{P}}) + \langle(\mathbf{P} - \dot{\mathbf{P}}), \nabla_P f(\dot{\mathbf{P}})\rangle + L_f^{\mathbf{P}}\|\mathbf{P} - \dot{\mathbf{P}}\|_2^2/2 \tag{8}$$

where $L_f^{\mathbf{P}}$ is the Lipschitz constant of $\nabla_P f = \mathbf{\Theta}^\top(\mathbf{\Theta}\mathbf{P} - \Upsilon)$. We therefore turn to optimize the following problem

$$\min_{\mathbf{P}} L_f^{\mathbf{P}}\|\mathbf{P} - (\dot{\mathbf{P}} - \nabla_P f(\dot{\mathbf{P}})/L_f^{\mathbf{P}})\|_F^2/2 + \gamma\|\mathbf{P}\|_* \tag{9}$$

This can be solved by singular value thresholding (SVT) [Cai *et al.*, 2010], and the optimizing rules are

$$\mathbf{P}^* = \mathcal{T}_{\gamma/L_f^{\mathbf{P}}}(\dot{\mathbf{P}} - \nabla_P f(\dot{\mathbf{P}})/L_f^{\mathbf{P}}) \tag{10}$$

where $\mathcal{T}_\tau(\mathbf{Z}) = \mathbf{U}\mathcal{S}_\tau(\mathbf{\Sigma})\mathbf{V}^\top$ is the SVT operator, which $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}$ are given by the singular value decomposition of $\mathbf{Z} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. $\mathcal{S}$ denotes the elementwise soft-thresholding operator defined by $\mathcal{S}_\tau(v) = (v - \tau)_+ - (-v - \tau)_+$, which $(x)_+$ replaces x with zero if $x < 0$, otherwise unchanged.

Update E. Fixing $\mathbf{W}$, $\mathbf{Q}$ and $\mathbf{P}$, we have

$$\min_{\mathbf{E}} \|(\tilde{\mathbf{Y}} - \mathbf{E}) - \mathbf{W}\mathbf{X}\|_F^2/2 + \|\Upsilon - \mathbf{\Theta}\mathbf{P}\|_F^2/2 + \Omega(\mathbf{E}) \tag{11}$$

There is no closed-form solution for Eq.(11). We update $\mathbf{E}$ of three norms uniformly by the proximal gradient method. Let $g(\mathbf{E}) = \Omega(\mathbf{E})$, $f(\mathbf{E}) = \|(\tilde{\mathbf{Y}} - \mathbf{E}) - \mathbf{W}\mathbf{X}\|_F^2/2 + \|\Upsilon - \mathbf{\Theta}\mathbf{P}\|_F^2/2$, and $L_f^{\mathbf{E}}$ is the Lipschitz constant of the gradient $\nabla_{\mathbf{E}} f = (\mathbf{W}\mathbf{X} - \tilde{\mathbf{Y}} + \mathbf{E}) + \sigma^{-1}(\mathbf{Y}\check{\mathbf{R}}_3 - \mathbf{Q}\mathbf{R}_3)/2$, where $\mathbf{R}_3 = -\mathbf{P} \odot \mathbf{\Theta}^\top\Upsilon$ and $\check{\mathbf{R}}_3$ is defined as in Eq.(6). The inequality

$$f(\mathbf{E}) \leq f(\dot{\mathbf{E}}) + \langle(\mathbf{E} - \dot{\mathbf{E}}), \nabla_{\mathbf{E}} f(\dot{\mathbf{E}})\rangle + L_f^{\mathbf{E}}\|\mathbf{E} - \dot{\mathbf{E}}\|_F^2/2 \tag{12}$$

holds, where $\dot{\mathbf{E}}$ is the solution of previous iteration. We solve

$$\min_{\mathbf{E}} L_f^{\mathbf{E}}\|\mathbf{E} - (\dot{\mathbf{E}} - \nabla_{\mathbf{E}} f(\dot{\mathbf{E}})/L_f^{\mathbf{E}})\|_F^2/2 + g(\mathbf{E}) \tag{13}$$

Next, we provide two useful lemmas related to proximity in the following, which are then used in Theorem 2.

Lemma 1. *[Jenatton* et al.*, 2011] Let groups $g_1 \preccurlyeq \ldots \preccurlyeq g_m$ such that $G = \{g_1, \ldots, g_m\}$, where $\preccurlyeq$ is a total order relation defined by $g \preceq h \Rightarrow \{g \subseteq h \text{ or } g \cap h = \emptyset\}$. The proximal operator $\mathbf{Prox}_{\lambda\Omega}$ associated with the norm Ω can be written as the composition of elementary operators ($\circ$)*

$$\mathbf{Prox}_{\lambda\Omega} = \mathbf{Prox}_{g_m} \circ \ldots \circ \mathbf{Prox}_{g_1} \tag{14}$$

Lemma 2. *[Yu, 2013] We denote $\partial f(x)$ as the subdifferential of the group f at point x, a sufficient condition for $\mathbf{Prox}_{f+g} = \mathbf{Prox}_f \circ \mathbf{Prox}_g$ is*

$$\partial g\left(\mathbf{Prox}_f(x)\right) \supseteq \partial g(x) \tag{15}$$

Theorem 2. *Let $g_1 = \|\cdot\|_2$, $g_2 = \|\cdot\|_1$ and $g_3 = \|\cdot\|_{2,1}$, then $\mathbf{Prox}_{\eta_1 g_1 + \eta_2 g_2 + \eta_3 g_3} = \mathbf{Prox}_{\eta_3 g_3} \circ \mathbf{Prox}_{\eta_1 g_1} \circ \mathbf{Prox}_{\eta_2 g_2}$, where $\mathbf{Prox}_g(\mathbf{J}) = \arg\min_{\mathbf{E}} L_f^{\mathbf{E}}\|\mathbf{E} - \mathbf{J}\|_F^2/2 + g(\mathbf{E})$.*

For $\eta_1 g_1$, the solution of related proximal map is

$$\mathbf{Prox}_{\eta_1\|\cdot\|_2}(\mathbf{J}) = L_f^{\mathbf{E}}\mathbf{J}/(L_f^{\mathbf{E}} + \eta_1) \tag{16}$$

for $\eta_2 g_2$, the related proximal map are given

$$\mathbf{Prox}_{\eta_2\|\cdot\|_1}(\mathbf{J}) = \mathcal{S}_{\eta_2/L_f^{\mathbf{E}}}(\mathbf{J}) \tag{17}$$

for $\eta_3 g_3$, the solution is

$$\mathbf{Prox}_{\eta_3\|\cdot\|_{2,1}}(\mathbf{J}) = \mathcal{B}_{\eta_3/L_f^{\mathbf{E}}}(\mathbf{J}) \tag{18}$$

where $\mathcal{B}_\tau(\mathbf{v}) = (1 - \tau/\|\mathbf{v}\|_2)_+\mathbf{v}$ is a block soft-thresholding operator which process matrix by column [Jenatton *et al.*, 2011]. Let $\mathbf{J} = \dot{\mathbf{E}} - \nabla_{\mathbf{E}} f(\dot{\mathbf{E}})/L_f^{\mathbf{E}}$, the solution of Eq.(13) follows directly from Theorem 2

$$\mathbf{E}^* = \mathbf{Prox}_{\eta_3\|\cdot\|_{2,1}} \circ \mathbf{Prox}_{\eta_1\|\cdot\|_2} \circ \mathbf{Prox}_{\eta_2\|\cdot\|_1}(\mathbf{J}) \tag{19}$$

Update W. When $\mathbf{Q}$, $\mathbf{P}$ and $\mathbf{E}$ are fixed, the optimization problem in Eq.(4) w.r.t $\mathbf{W}$ can be reformulated as follows

$$\min_{\mathbf{W}} \frac{1}{2}\|(\tilde{\mathbf{Y}} - \mathbf{E}) - \mathbf{W}\mathbf{X}\|_F^2 + \alpha\|\mathbf{W}\|_1 \tag{20}$$

The minimization of Eq.(20) is convex, but non-smooth due to the ℓ_1-norm terms. We use the *accelerated proximal gradient (APG)* method to solve it [Beck and Teboulle, 2009]. Let $g(\mathbf{W}) = \alpha\|\mathbf{W}\|_1$ and $f(\mathbf{W}) = \frac{1}{2}\|(\tilde{\mathbf{Y}} - \mathbf{E}) - \mathbf{W}\mathbf{X}\|_F^2$. The derivation of f is denoted as $\nabla_{\mathbf{W}} f = (\mathbf{W}\mathbf{X} - \tilde{\mathbf{Y}} + \mathbf{E})\mathbf{X}^\top$, and let $L_f^{\mathbf{W}}$ denotes the Lipschitz constant, we consider the following *QP* problem

$$\min_{\mathbf{W}} L_f^{\mathbf{W}}\|\mathbf{W} - (\dot{\mathbf{W}} - \nabla_{\mathbf{W}} f(\dot{\mathbf{W}})/L_f^{\mathbf{W}})\|_F^2/2 + g(\mathbf{W}) \tag{21}$$

The *APG* method includes an *extrapolation* step in the algorithm, the optimizing rules are given

$$\dot{\mathbf{Z}} = \dot{\mathbf{W}} + \frac{\ddot{\omega} - 1}{\dot{\omega}}(\dot{\mathbf{W}} - \ddot{\mathbf{W}}) \tag{22}$$

and

$$\mathbf{W}^* = \mathcal{S}_{\alpha/L_f^{\mathbf{W}}}(\dot{\mathbf{Z}} - \nabla_{\mathbf{W}} f(\dot{\mathbf{Z}})/L_f^{\mathbf{W}}) \tag{23}$$

where $\ddot{\mathbf{W}}$ denotes the optimal solution at the iteration before last. $\dot{\omega}$, $\ddot{\omega} \in [0, 1)$ are extrapolation parameter at previous iteration and before last iteration, respectively. In practice, we update $\omega = 1/2 + \sqrt{4\dot{\omega}^2 + 1}/2$.

Datasets	NMLD	fPML	PML-NI	PART-VLS	PART-MAP	PML-LC	PML-FP	MUSER
Coverages (↓)								
slashdot	**.317±.010**	.423±.065	.437±.026	.389±.013	.379±.011	.433±.019	.427±.016	.429±.020
medical	**.230±.009**	.389±.005	.466±.022	.412±.004	.391±.003	.463±.005	.472±.014	.484±.012
enron	**.146±.012**	.212±.043	.232±.013	.224±.001	.214±.013	.231±.015	.232±.013	.237±.016
scene	**.614±.032**	.633±.016	.700±.008	.723±.017	.728±.019	.734±.009	.725±.013	.753±.008
yeast	**.652±.024**	.637±.044	.666±.022	.705±.019	.708±.013	.714±.013	.707±.011	.709±.018
20ng	**.136±.003**	.173±.015	.188±.013	.145±.012	**.136±.011**	.219±.002	.235±.003	.193±.011
corel5k	.202±.011	**.190±.041**	.217±.006	.213±.001	.218±.020	.224±.013	.228±.016	.223±.007
mirflickr	**.114±.016**	.115±.010	.119±.011	.167±.012	.169±.013	.237±.012	.244±.008	.243±.009
eurlex_dc	**.286±.010**	.326±.009	.291±.021	.327 ±.026	.325 ±.045	.378±.021	.379±.011	.306±.032
m_emotion	**.411±.008**	.418±.015	.423±.017	.423±.011	.426±.018	.443±.003	.460±.012	.417±.017
AveragePrecision (↑)								
slashdot	**.666±.005**	.537±.010	.573±.024	.633±.016	.655±.022	.552±.004	.583±.021	.620±.006
medical	**.813±.011**	.794±.006	.714±.041	.754±.031	.775±.021	.703±.011	.715±.008	.732±.019
enron	**.689±.021**	.668±.012	.459±.012	.594±.026	.668±.012	.554±.012	.568±.011	.569±.007
scene	**.827±.006**	.820±.022	.794±.012	.754±.021	.758±.001	.707±.015	.714±.008	.641±.005
yeast	.756±.014	**.765±.021**	.754±.006	.713±.012	.714±.011	.706±.028	.709±.015	.709±.012
20ng	**.785±.004**	.744±.009	.746±.022	.777±.022	.783±.010	.672±.013	.677±.018	.725±.010
corel5k	**.200±.003**	.193±.001	.133±.006	.173±.007	.174±.003	.124±.011	.125±.007	.019±.003
mirflickr	**.781±.012**	.776±.015	.779±.023	.720±.017	.727±.019	.574±.010	.581±.008	.586±.025
eurlex_dc	**.725±.004**	.629±.017	.724±.013	.633 ±.051	.633 ±.030	.596±.015	.599±.009	.721±.005
m_emotion	**.609±.006**	.604±.011	.600±.011	.589±.010	.595±.022	.563±.027	.566±.003	**.609±.012**

Table 1: Experimental results on *redundant multi-label data*. ↑ indicates the larger, the better; ↓ indicates the smaller, the better.

4 Experiments

4.1 Experimental Setting

Datasets. To evaluate NMLD's performance, we conducted experiments on ten multi-label datasets: slashdot, medical, enron, scene, yeast, 20ng, corel5k, mirflickr, eurlex_dc, and m_emotion [Pestian *et al.*, 2007; Tidake and Sane, 2018]. Table 2 shows their characteristics, including number of instances, features, label classes, and maximum and average true labels per dataset. We also generated 3 synthetic datasets with different label noises for further testing. 1) For comparison on dirty noise, given a noise rate ρ, to make the noise matrix $\mathbf{E}$ with the dirty distribution, we first set $\mathbf{E} = 0$ and randomly sample the row $\mathbf{E}_{i:}$ from $\mathbf{E}$ with probability (subjective). Then, we draw elements from each row $\mathbf{E}_{i:}$ randomly (sparsity) and assign a value to each element such that the value obeys the gaussian distribution. 2) For partial multi-label learning comparison purpose, we generate synthetic datasets by adding random redundant label noises. We use the same approach as the existing PML works [Xie and Huang, 2018]. 3) For multi-label learning with missing labels comparison, we randomly remove few classes or positive labels as MLML [Wu *et al.*, 2014]. In our work, the dirty, excess and missing rate are set to 50%, 20% and 50%, respectively.

Baselines. Existing approaches fall into two categories: partial multi-label learning (PML) and multi-label learning with missing labels (MLML). We evaluated our method in both contexts by comparing it with 7 state-of-the-art PML algorithms (i.e., fPML [Yu *et al.*, 2018], PML-NI [Xie and Huang, 2021], PART-VLS, PART-MAP [Fang and Zhang, 2019], PML-LC, PML-FP [Xie and Huang, 2018] and MUSER [Li *et al.*, 2020]) and 7 state-of-the-art MLML algorithms (D2ML-L, D2ML-NL [Ma and Chen, 2021], LEML [Yu *et al.*, 2014], MLMLFS [Zhu *et al.*, 2018], GLOCAL [Zhu *et al.*, 2017], MLMLV1 [Wu *et al.*, 2014] and MAXIDE [Xu *et al.*, 2013]). Consider there is no existing methods designed for the NML problem. We compare with the following methods CORALS [Sun *et al.*, 2021], ECC [Ferng and Lin, 2011], CBMLC [Zhao and Gomes, 2021], where CORALS and CBMLC handle missing and noisy labels and ECC is full-supervised.

Datasets	*#instance*	*#dim*	*#label*	*#max*	*#avg*
slashdot	3782	1079	22	3	1.18
medical	978	1449	45	3	1.25
enron	1702	1001	53	12	3.37
scene	2407	294	6	3	1.07
yeast	2417	103	14	11	4.23
20ng	19300	1006	20	14	1.02
corel5k	5000	499	374	5	3.52
mirflickr	25000	15	24	5	1.77
eurlex_dc	19384	5000	412	7	1.29
m_emotion	6833	98	11	7	2.42

Table 2: Statistics of the Datasets.

4.2 Comparison Results

We use 4 MLL evaluation metrics: Coverages, Average Precision, One Error and Ranking Loss. We compare the proposed NMLD with the 17 comparison methods on the 10 datasets.

Datasets	NMLD	D2ML-L	D2ML-NL	LEML	MLMLFS	GLOCAL	MLML	MAXIDE
Coverages (↓)								
slashdot	**.257±.019**	.514±.023	.471±.009	.397±.001	.496±.026	.434±.030	.412±.013	.512±.010
medical	**.157±.002**	.900±.038	.724±.024	.418±.014	.908±.015	.642±.031	.374±.009	.933±.052
enron	**.122±.005**	.263±.020	.241±.013	.305±.012	.210±.009	.163±.003	.125±.007	.299±.014
scene	.648±.027	.922±.049	.742±.012	.911±.034	.908±.081	.795±.033	**.539±.014**	1.84±.045
yeast	.663±.061	.744±.042	.684±.020	.676±.023	.743±.024	.741±.041	**.632±.017**	.913±.039
20ng	**.145±.014**	.698±.056	.313±.012	.373±.015	.674±.010	.551±.029	.386±.026	.900±.013
corel5k	.163±.014	.174±.008	.193±.031	.221±.013	.166±.012	.127±.003	**.111±.014**	.191±.021
mirflickr	**.123±.010**	.128±.013	.128±.016	.145±.009	.165±.022	.148±.002	.134±.010	.244±.011
eurlex_dc	**.269±.003**	.277±.022	.272±.021	.304 ±.030	.347 ±.025	.326±.026	.291±.015	.420±.031
m_emotion	**.402±.026**	.506±.014	.490±.023	.407±.007	.524±.011	.435±.015	.428±.017	.580±.027
AveragePrecision (↑)								
slashdot	**.675±.014**	.359±.015	.468±.019	.610±.003	.424±.019	.463±.022	.479±.001	.465±.012
medical	**.836±.001**	.443±.021	.520±.021	.773±.022	.367±.012	.507±.005	.713±.012	.342±.013
enron	**.694±.012**	.323±.013	.352±.011	.496±.025	.501±.024	.624±.014	.623±.011	.448±.003
scene	.809±.020	.736±.011	.744±.010	.663±.009	.710±.018	.774±.012	**.819±.004**	.645±.028
yeast	**.753±.019**	.673±.025	.681±.009	.750±.004	.694±.018	.686±.013	.714±.024	.716±.021
20ng	**.777±.002**	.313±.005	.315±.011	.538±.013	.469±.028	.424±.015	.634±.001	.455±.027
corel5k	**.274±.014**	.150±.003	.196±.009	.247±.019	.195±.006	.222±.013	.231±.027	.212±.010
mirflickr	**.911±.012**	.879±.013	.888±.021	.872±.016	.817±.025	.873±.033	.871±.001	.725±.018
eurlex_dc	**.743±.004**	.693±.036	.733±.005	.700±.042	.655±.008	.630±.005	.684±.031	.639±.015
m_emotion	.667±.004	.598±.023	.618±.013	**.676±.009**	.536±.013	.646±.020	.664±.013	.599±.022

Table 3: Experimental results on *missing multi-label data*. ↑ indicates the larger, the better; ↓ indicates the smaller, the better.

Comparison on Redundant Multi-Label Data. We compared the performance of NMLD with baselines in PML settings for label prediction across 10 datasets, as shown in Table 1. NMLD consistently outperformed other PML methods. Notably, NMLD significantly surpassed PART-VLS, PML-NI, PML-LC, and PML-FP, and often exceeded PML, PART-MAP, and MUSER. These results highlight NMLD's capability in managing redundant noisy labels and validate its effectiveness in nonlinear mapping for partial multi-label matrices.

Comparison on Missing Mutli-Label Data. In the MLML setting, NMLD demonstrates competitive performance against established multi-label algorithms, as seen in Table 3. It notably surpasses LEML and MLML—the top-performing baselines—across most datasets for all evaluation metrics. LEML considers only a low-rank decomposition of the model coefficient, while the low-rank label matrix assumption is violated in many real-world application due to the presence of tail labels [Xu *et al.*, 2016]. The super performance of NMLD against LEML and others indicates the effectiveness of NMLD on solving MLML problem.

Comparison on Dirty Multi-Label Data. After that, we study the prediction of the proposed algorithm on dirty noisy multi-label learning. Table 4 provides the experimental result of each compared method on 10 different datasets. As shown in Table 4, out of 40 statistical tests (10 data sets × 4 evaluation metrics), NMLD ranks in 1st place among the 4 comparing methods at 90% cases and ranks in 2st place at 7% cases. NMLD achieves the best performance in most cases. Regarding the CORALS and CBMLC, NMLD significantly outperforms them on most datasets. Regarding the multi-label learning approaches, the performance of NMLD is statistically superior to ECC. This indicates that label noisy significantly influences the performance of multi-label classifiers and NMLD can identify noisy labels effectively.

Ablation Analysis. We set the corresponding parameters $\eta_i = 0$ for the ablation experiments respectively. For example, set $\eta_1 = 0$ to get $\ell_1 + \ell_{2,1}$ or $\eta_2 = 0, \eta_3 = 0$ to get ℓ_2. The evaluation result Average Precision (AP) is given in Table 5. As shown in Table 5, we found that the performance of combined norm is generally better than that of single norm, such as the model with $\ell_1 + \ell_{2,1}$ is better than ℓ_1 and $\ell_{2,1}$. Moreover, the model with $\ell_2 + \ell_1 + \ell_{2,1}$ has the best performance, which is in accordance with the Theorem 1.

Time Comparison. In this section we perform time comparison on two datasets medical and 20ng. The running time of each method on data with above dirty noise are recorded in Table 6. As shown in Table 6, our approach is the most efficient one. For example, in the MLML setting, NMLD significantly outperforms all MLML methods on these two datasets. NMLD is more than 3 times faster than LEML, which is the most efficient existing MLML approach.

4.3 Experiments on Large Scale Datasets

To validate the performance of NMLD on large-scale data, we conduct experiments on the extreme multi-label benchmark data *EURLex-4K*, which contains 15,539 instances and *Mediamill* containing 43,970 instances [Bhatia *et al.*, 2016]. We add PML and MLML noise as described previously. Ta-

Datasets	**NMLD**	**CORALS**	**CBMLC**	**ECC**	**NMLD**	**CORALS**	**CBMLC**	**ECC**
	RankingLoss (↓)				OneError (↓)			
slashdot	**.141±.004**	.181±.016	.153±.014	.271±.005	**.470±.012**	.627±.008	.531±.012	.707±.008
medical	**.055±.005**	.066±.011	.060±.012	.164±.010	**.294±.007**	.414±.007	.331±.004	.609±.012
enron	**.134±.012**	.215±.015	.137±.004	.333±.018	**.234±.008**	.401±.005	.266±.009	.675±.007
scene	**.089±.006**	.267±.003	.103±.006	.199±.012	**.264±.011**	.520±.006	.267±.021	.419±.002
yeast	**.177±.001**	.180±.003	**.177±.003**	.193±.006	.234±.006	**.229±.007**	.260±.010	.246±.009
20ng	**.089±.013**	.147±.007	.113±.001	.115±.005	**.359±.021**	.407±.005	.361±.012	.401±.005
corel5k	**.309±.014**	.350±.007	.317±.013	.325±.017	**.799±.002**	.880±.013	.805±.003	.892±.008
mirflickr	**.052±.008**	.064±.010	.060±.008	.061±.005	**.113±.007**	.116±.011	.124±.007	.129±.010
eurlex_dc	**.046±.015**	.120±.004	.048±.001	.056±.007	**.069±.008**	.212±.001	.073±.016	.070±.011
m_emotion	**.202±.005**	.322±.004	.258±.008	.204±.009	.369±.002	.522±.002	.445±.009	**.367±.003**
	AveragePrecision (↑)				Coverages (↓)			
slashdot	**.623±.019**	.567±.011	.602±.003	.418±.029	**.370±.014**	.442±.007	.427±.013	.626±.008
medical	**.783±.007**	.698±.009	.770±.005	.469±.012	**.298±.002**	.722±.008	.314±.012	.814±.016
enron	**.660±.006**	.440±.026	.644±.032	.299±.009	**.156±.020**	.240±.014	.170±.013	.298±.025
scene	**.834±.011**	.683±.018	.830±.014	.732±.015	**.050±.013**	.135±.028	.057±.002	.103±.007
yeast	.749±.008	**.759±.011**	.751±.021	.713±.007	**.629±.005**	.632±.003	.632±.012	.711±.005
20ng	**.749±.012**	.667±.007	.729±.016	.693±.009	**.173±.003**	.269±.012	.189±.004	.230±.003
corel5k	**.140±.010**	.119±.007	.137±.003	.099±.011	**.217±.007**	.241±.019	.222±.002	.246±.007
mirflickr	**.907±.006**	.894±.013	.900±.008	.890±.010	**.114±.013**	.124±.020	.123±.011	.129±.015
eurlex_dc	**.705±.019**	.523±.016	.703±.007	.695±.002	.311±.012	.381±.003	**.305±.009**	.315±.006
m_emotion	**.667±.010**	.595±.014	.610±.002	.663±.028	**.419±.016**	.570±.021	.467±.004	.422±.006

Table 4: Experimental results on *dirty multi-label data*. ↑ indicates the larger, the better; ↓ indicates the smaller, the better.

AP (↑)	ℓ_2	ℓ_1	$\ell_{2,1}$
medical	.777±.006	.769±.030	.775±.012
20ng	.735±.009	.732±.016	.740±.004
	$\ell_1 + \ell_2$	$\ell_1 + \ell_{2,1}$	$\ell_2 + \ell_1 + \ell_{2,1}$
medical	.775±.003	.781±.007	**.783±.007**
20ng	.737±.003	.741±.004	**.749±.012**

Table 5: Ablation experiment in data medical and 20ng.

Time	**NMLD**	**CORALS**	**CBMLC**	**PML-NI**	**fPML**
medical	**1.12s**	879.43s	15.33s	14.08s	68.17s
20ng	**10.93s**	>1day	96.88s	23.07s	233.83s
	PART-VLS	**PML-LC**	**D2ML-L**	**LEML**	**GLOCAL**
medical	24.97s	1480.21s	1.35s	7.49s	15.72s
20ng	2875.53s	>1day	1040.55s	33.41s	43.75s

Table 6: Time comparison in datasets medical and 20ng.

ble 7 reports the detailed experimental results of the proposed NMLD and comparing algorithms in term of Coverages. As show in Table 7, NMLD achieves the best performance against other SOTA contenders on large scale data.

5 Conclusion

In this paper, we proposed a novel noisy multi-label learning framework NML, which tackles a more practical learning problem over training data with dirty noise. We presented a robust approach NMLD, which unifies a mixed penalty on noise matrix and true label matrix exploration in a unified objective. A theoretical guarantee for exact noise recovery from dirty noise data has been provided. We formulated the method as a non-convex and non-smooth problem with accelerated proximal alternating techniques to jointly optimize the noise and true label matrix in a mutually beneficial manner. Empirical studies on 10 datasets against 17 baselines confirmed that our proposed approach outperform state-of-the-art baseline algorithms significantly in the new NML problem as well as existing partial multi-label learning and multi-label learning with missing labels settings.

PML Data	**NMLD**	**fPML**	**PML-NI**	**PART-VLS**
EURLex-4K	**.294±.004**	.396±.011	.331±.007	.348±.004
Mediamill	**.192±.005**	.265±.007	.249±.003	.224±.002
MLML Data	**NMLD**	**GLOCAL**	**MLML**	**LEML**
EURLex-4K	**.286±.010**	.377±.008	.385±.011	.322±.007
Mediamill	**.188±.003**	.240±.002	.233±.005	.201±.012

Table 7: Experimental results on *large scale* multi-label data with PML and MLML baselines in terms of Coverages (↓).

Acknowledgments

This research is supported in part by the National Natural Science Foundation of China under Grant No.92267206 and No.62032013, Singapore Institute of Technology Ignition Grant (R-IE2-A405-0001), AcRF Tier 1 Grant (R-R12-A405-0009).

References

[Bauschke and Combettes, 2017] Heinz H Bauschke and Patrick L Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer New York, 2017.

[Beck and Teboulle, 2009] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[Bhatia *et al.*, 2016] Kush Bhatia, Kunal Dahiya, Himanshu Jain, Purushottam Kar, Anshul Mittal, Yashoteja Prabhu, and Manik Varma. The extreme classification repository: Multi-label datasets and code, 2016.

[Brodley and Friedl, 1999] Carla E Brodley and Mark A Friedl. Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11:131–167, 1999.

[Cai *et al.*, 2010] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[Fan and Chow, 2019] Jicong Fan and Tommy WS Chow. Exactly robust kernel principal component analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 31(3):749–761, 2019.

[Fang and Zhang, 2019] Jun-Peng Fang and Min-Ling Zhang. Partial multi-label learning via credible label elicitation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3518–3525, 2019.

[Ferng and Lin, 2011] Chung-Sung Ferng and Hsuan-Tien Lin. Multi-label classification with error-correcting codes. In *Asian Conference on Machine Learning*, pages 281–295. PMLR, 2011.

[Jenatton *et al.*, 2011] Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, and Francis Bach. Proximal methods for hierarchical sparse coding. *The Journal of Machine Learning Research*, 12:2297–2334, 2011.

[Li *et al.*, 2020] Ziwei Li, Gengyu Lyu, and Songhe Feng. Partial multi-label learning via multi-subspace representation. In *International Joint Conference on Artificial Intelligence*, pages 2612–2618, 2020.

[Lin, 2023] Dekun Lin. Probability guided loss for long-tailed multi-label image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1577–1585, 2023.

[Ma and Chen, 2021] Zhongchen Ma and Songcan Chen. Expand globally, shrink locally: Discriminant multi-label learning with missing labels. *Pattern Recognition*, 111:107675, 2021.

[Ma *et al.*, 2019] Jingting Ma, Anqi Wang, Feng Lin, Stefan Wesarg, and Marius Erdt. A novel robust kernel principal component analysis for nonlinear statistical shape modeling from erroneous data. *Computerized Medical Imaging and Graphics*, 77:101638, 2019.

[Natarajan *et al.*, 2013] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in Neural Information Processing Systems*, 26, 2013.

[Pestian *et al.*, 2007] John Pestian, Chris Brew, Pawel Matykiewicz, Dj J Hovermale, Neil Johnson, K Bretonnel Cohen, and Wlodzislaw Duch. A shared task involving multi-label classification of clinical free text. In *Biological, Translational, and Clinical Language Processing*, pages 97–104, 2007.

[Schölkopf *et al.*, 1997] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International Conference on Artificial Neural Networks*, pages 583–588. Springer, 1997.

[Sun *et al.*, 2021] Lijuan Sun, Gengyu Lyu, Songhe Feng, and Xiankai Huang. Beyond missing: Weakly-supervised multi-label learning with incomplete and noisy labels. *Applied Intelligence*, 51(3):1552–1564, 2021.

[Tidake and Sane, 2018] Vaishali S Tidake and Shirish S Sane. Multi-label classification: A survey. *International Journal of Engineering and Technology*, 7(4.19):1045–1054, 2018.

[Van Rooyen and Williamson, 2017] Brendan Van Rooyen and Robert C Williamson. A theory of learning with corrupted labels. *The Journal of Machine Learning Research*, 18(1):8501–8550, 2017.

[Wang *et al.*, 2023] Yejiang Wang, Yuhai Zhao, Zhengkui Wang, and Ling Li. Galopa: Graph transport learning with optimal plan alignment. *Advances in Neural Information Processing Systems*, 36, 2023.

[Wang *et al.*, 2024a] Yejiang Wang, Yuhai Zhao, Zhengkui Wang, Wen Shan, and Xingwei Wang. Limited-supervised multi-label learning with dependency noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15662–15670, 2024.

[Wang *et al.*, 2024b] Yejiang Wang, Yuhai Zhao, Zhengkui Wang, Chengqi Zhang, and Xingwei Wang. Robust multi-graph multi-label learning with dual-granularity labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[Wu *et al.*, 2014] Baoyuan Wu, Zhilei Liu, Shangfei Wang, Bao-Gang Hu, and Qiang Ji. Multi-label learning with missing labels. In *22nd International Conference on Pattern Recognition*, pages 1964–1968. IEEE, 2014.

[Xie and Huang, 2018] Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[Xie and Huang, 2021] Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning with noisy label identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[Xie *et al.*, 2022] Ming-Kun Xie, Jiahao Xiao, and Sheng-Jun Huang. Label-aware global consistency for multi-label learning with single positive labels. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 18430–18441. Curran Associates, Inc., 2022.

[Xie *et al.*, 2023] Ming-Kun Xie, Jiahao Xiao, Hao-Zhe Liu, Gang Niu, Masashi Sugiyama, and Sheng-Jun Huang. Class-distribution-aware pseudo-labeling for semi-supervised multi-label learning. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 25731–25747. Curran Associates, Inc., 2023.

[Xu *et al.*, 2013] Miao Xu, Rong Jin, and Zhi-Hua Zhou. Speedup matrix completion with side information: Application to multi-label learning. In *Advances in Neural Information Processing Systems*, pages 2301–2309, 2013.

[Xu *et al.*, 2014] Linli Xu, Zhen Wang, Zefan Shen, Yubo Wang, and Enhong Chen. Learning low-rank label correlations for multi-label classification with missing labels. In *2014 IEEE International Conference on Data Mining*, pages 1067–1072, 2014.

[Xu *et al.*, 2016] Chang Xu, Dacheng Tao, and Chao Xu. Robust extreme multi-label learning. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1275–1284, 2016.

[Xu *et al.*, 2023] Pengyu Xu, Lin Xiao, Bing Liu, Sijin Lu, Liping Jing, and Jian Yu. Label-specific feature augmentation for long-tailed multi-label text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10602–10610, 2023.

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

[Yu *et al.*, 2018] Guoxian Yu, Xia Chen, Carlotta Domeniconi, Jun Wang, Zhao Li, Zili Zhang, and Xindong Wu. Feature-induced partial multi-label learning. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1398–1403. IEEE, 2018.

[Yu, 2013] Yaoliang Yu. On decomposing the proximal map. In *Advances in Neural Information Processing Systems*, pages 91–99, 2013.

[Zhang *et al.*, 2022] Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. Use all the labels: A hierarchical multi-label contrastive learning framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16660–16669, June 2022.

[Zhao and Gomes, 2021] Wenting Zhao and Carla Gomes. Evaluating multi-label classifiers with noisy labels. *arXiv:2102.08427*, 2021.

[Zhu and Wu, 2004] Xingquan Zhu and Xindong Wu. Class noise vs. attribute noise: A quantitative study. *Artificial Intelligence Review*, 22(3):177–210, 2004.

[Zhu *et al.*, 2017] Yue Zhu, James T Kwok, and Zhi-Hua Zhou. Multi-label learning with global and local label correlation. *IEEE Transactions on Knowledge and Data Engineering*, 30(6):1081–1094, 2017.

[Zhu *et al.*, 2018] Pengfei Zhu, Qian Xu, Qinghua Hu, Changqing Zhang, and Hong Zhao. Multi-label feature selection with missing labels. *Pattern Recognition*, 74:488–502, 2018.