FREDIS: A Fusion Framework of Refinement and Disambiguation for Unreliable Partial Label Learning

Congyu Qiao [1] **Ning Xu** [1] **Jiaqi Lv** [2] **Yi Ren** [3] **Xin Geng** [1]

Abstract

To reduce annotation difficulty, Partial label learning (PLL) uses ambiguous annotations with candidate labels instead of the exact correct label. It assumes the candidate label set contains the correct label, inducing disambiguation, which is commonly adopted in PLL methods. However, this assumption is impractical as no one could guarantee the existence of the correct label in the candidate label set under real-world scenarios. Therefore, Unreliable Partial Label Learning (UPLL) is investigated where the correct label of each example may not exist in the candidate label set. In this paper, we propose a fusion framework of refinement and disambiguation named FREDIS to handle the UPLL problem. Specifically, with theoretical guarantees, not only does disambiguation move incorrect labels from candidate labels to non-candidate labels but also refinement, an opposite procedure, moves correct labels from non-candidate labels to candidate labels. Besides, we prove that the classifier trained by our framework could eventually approximate the Bayes optimal classifier. Extensive experiments on widely used benchmark datasets validate the effectiveness of our proposed framework. Source code is available at https://github.com/palm-ml/fredis.

1. Introduction

Partial Label Learning (PLL), known as a typical weakly supervised learning (Patrini et al., 2017; Zhou, 2018; Kamnitsas et al., 2018; Lu et al., 2018; Gong et al., 2019), learns a multi-class classifier from instances annotated with candidate label sets, where the exact correct labels lie fixed but unknown. PLL has been widely encountered in a variety of real-world domains including web mining (Luo & Orabona, 2010), ecoinformatics (Liu & Dietterich, 2012), and multimedia content analysis (Zeng et al., 2013).

To tackle the PLL problem, a large number of approaches have been proposed. Identification-based approaches (Jin & Ghahramani, 2002; Nguyen & Caruana, 2008; Liu & Dietterich, 2012; Yu & Zhang, 2016; Chen et al., 2017; Feng & An, 2019; Wang et al., 2021; Ni et al., 2021) consider the correct label as a latent variable and aim to identify it through various techniques. Average-based approaches (Hüllermeier & Beringer, 2006; Cour et al., 2011; Zhang & Yu, 2015) treat all the candidate labels equally and predict by averaging the modeling outputs. Additionally, recent advancements in deep PLL algorithms have emerged oriented at regularization items (Yao et al., 2020a;b; Wu et al., 2022), classifier or risk consistency (Lv et al., 2020; Feng et al., 2020; Wen et al., 2021), and intrinsic representations (Zhang et al., 2021; Wang et al., 2022), providing new avenues for research in this field.

The preceding methods are all founded on the PLL assumption that each candidate label set inevitably contains the correct label. Nevertheless, this assumption is impractical as no one could guarantee the existence of the correct label in the candidate label set. Let us reconsider how candidate labels are typically generated: One case is that an annotator (or more abstractly, an annotation system) cannot decide which one is the correct label for a given instance, and thus uses multiple labels for annotation (Cour et al., 2011). Another is that multiple annotators disagree on the same instance (Jin & Ghahramani, 2002; Zhang et al., 2017). The two cases do not completely avoid inherent ambiguity, and the correct label cannot be guaranteed to be included by the corresponding candidate label set. Hence, we need to investigate Unreliable Partial Labeling Learning (UPLL) (Lv et al., 2023), where the candidate label set is not guaranteed to contain the correct label. In this way, PLL can be generalized to more real-world scenarios, such as crowdsourcing (CROWDSOURCING, 2008). Typically, several non-expert crowdworkers can be organized to annotate a large-scale

[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China [2]RIKEN Center for Advanced Intelligence Project [3]Research Center for Healthcare Data Science, Zhejianglab, China. Correspondence to: Ning Xu <xning@seu.edu.cn>, Xin Geng <xgeng@seu.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

dataset, saving a significant amount of annotation costs.

In this paper, we propose a theoretically grounded framework for UPLL named FREDIS, i.e., *Fusion of REfinement and DISambiguation*. Specifically, since correct labels may hide in non-candidate labels, we perform not only disambiguation but also refinement of moving the correct label from non-candidate labels into candidate labels to avoid missing supervision information and misleading. Under mild assumptions, we prove that by selecting appropriate thresholds based on the classifier output, correct labels can be possibly refined from non-candidate labels and incorrect labels can be disambiguated from candidate labels. During each fusion round of refinement and disambiguation, we control disambiguated labels to be far more than refined labels in order to mitigate the risk posed by mistaken incorrect labels. As a result, the entire candidate label sets would be gradually purified, and the classifier is proved capable of approximating the Bayes optimal classifier eventually. Our main contributions are summarized below:

- We propose a novel framework named FREDIS from the perspective of data calibration. The framework progressively purifies the whole candidate label sets via performing refinement and disambiguation simultaneously.
- We demonstrate that FREDIS has the competence to refine correct labels out of non-candidate labels and is guaranteed to eliminate incorrect labels out of candidate labels. Disambiguation and refinement will be accommodated harmoniously in FREDIS, which is proved to improve the performance of the classifier iteratively.
- We prove that the classifier trained by our proposed framework approximates the Bayes optimal classifier under mild assumptions. To the best of our knowledge, this is the first theoretically guaranteed framework for the UPLL problem along the research line of identification.

2. Related Work

In this section, we will provide a brief overview of two aspects of PLL research: traditional PLL and deep PLL, both of which have achieved tremendous theoretical and empirical improvements.

Traditional PLL involves the use of linear models to perform disambiguation, which can be classified into two types:: identification and average. Those based on identification focus on distinguishing the correct label from the candidate label set by selecting one label as the ground truth based on certain criteria and directly maximizing the model's output for that label. For instance, (Jin & Ghahramani, 2002) apply the EM algorithm to determine which label among the candidate label set is more appropriate for training than the others. (Nguyen & Caruana, 2008) generalize the margin-based multi-class approach to margin-based partial label classification to disambiguate the candidate label set. (Liu & Dietterich, 2012) maximize the likelihood of data based on the assumption of a noise distribution, consider the correct label as a hidden variable, and discriminate it using variational EM. On the other hand, those based on average thought pursue disambiguation between candidate labels and non-candidate labels, thereby often putting all labels in the candidate label set in the equal position and averaging the outputs of the model on all candidate labels to make predictions. Typically, (Hüllermeier & Beringer, 2006) apply the k-nearest neighbour technique to vote for each acceptable label. In (Cour et al., 2011; Zhang et al., 2016), the parameters of the classifier are optimized to maximize the difference between the average score of candidate labels and non-candidate labels.

Deep PLL lifts the restriction on model structures, data dimensions and optimization strategies. Equipped with deep neural networks, the PLL problem has been studied on many benchmark datasets, in which parameters of the classifier is efficiently updated by various stochastic optimization, such as SGD (Robbins & Monro, 1951), and ADAM (Kingma & Ba, 2014). (Yao et al., 2020a) is the pioneering work to integrates deep neural networks into PLL to improve the representation ability of the models. Building upon the estimation error bound, (Lv et al., 2020) theoretically prove that its classifier learned from candidate label sets can converge to the optimal one trained with correct labels. From the perspective of the generation model of candidate label sets, (Feng et al., 2020) propose one risk-consistent estimator and one classifier-consistent estimator, which makes the assumption that the candidate label set is uniformly sampled with the correct label contained. (Wen et al., 2021) generalize the uniform generation process and assume the candidate label set to be label-specific. (Xu et al., 2021b) first notice that the instance-dependent case is more realistic and adopt the variational inference technique to estimate the latent label distribution (Xu et al., 2021a; 2023) of each instance. (Zhang et al., 2021) discover that class activation maps can be utilized for disambiguation purposes, and further propose using class activation values to capture learned representation information in a more general manner. (Wang et al., 2022) aim to improve the extracted representation and introduce constrastive learning. (Wu et al., 2022) consider the manifold consistency, which attempt to maintain the manifold in both feature space and label space. (Lv et al., 2023) first consider UPLL and propose average partial-label losses for UPLL along the research line of average. (Lian et al., 2023) focuses on detecting and correcting unreliable training examples to reduce the unreliability in UPLL.

At the core of most PLL approaches, whether traditional or deep, is disambiguation, which excludes incorrect labels from candidate labels or, more flexibly, exerts more weight on correct labels and less weight on incorrect labels within

candidate labels when training the classifier. However, in UPLL, disambiguation is ineffective because the correct label may not be present in the candidate label set. To avoid missing supervision information, refinement, which attempts to move the correct label from non-candidate labels to candidate labels, is supposed to be implemented. Nevertheless, disambiguation and refinement naturally interact on each other in a state of competition. Therefore, we consider fusing disambiguation and refinement with theoretical guarantees for UPLL in this paper.

3. Problem Setup

3.1. Partial Label Learning

For partial label learning, each example is annotated with a candidate label set, which conceals the correct label. The objective is to induce a predictive model capable of assigning the correct label to an unseen instance. Let $\mathcal{X} = \mathbb{R}^q$ be the q-dimensional instance space and $\mathcal{Y} = \{1, 2, ..., c\}$ be the label space with c class labels. Given a PLL training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, S_i)|1 \leq i \leq n\}$, where $\boldsymbol{x}_i \in \mathcal{X}$ denotes the i-th q-dimensional instance and $S_i \in \mathcal{C}$ denotes the candidate label set associated with $\boldsymbol{x}_i$ where $\mathcal{C} = 2^{\mathcal{Y}} \setminus \{\emptyset, \mathcal{Y}\}$. The task of PLL is to train a multi-class classifier $f : \mathcal{X} \mapsto \mathcal{Y}$ using the PLL training dataset $\mathcal{D}$.

3.2. Unreliable Partial Label Learning

Given a UPLL training set $\tilde{\mathcal{D}} = \{(\boldsymbol{x}_i, \tilde{S}_i)|1 \leq i \leq n\}$ where $\tilde{S}_i \in \mathcal{C}$ denotes the candidate label set of $\boldsymbol{x}_i$ in UPLL. Note that in PLL the correct label y_i is guaranteed to exist in the corresponding candidate label set S_i, i.e., $y_i \in S_i$ always holds, while in UPLL the correct label y_i is not necessarily in the candidate label set $\tilde{S}_i$, i.e., $y_i \notin \tilde{S}_i$ sometimes exists, which is more challenging but more practical. The task of UPLL is also to induce a multi-class classifier $f : \mathcal{X} \mapsto \mathcal{Y}$ from $\tilde{\mathcal{D}}$, which can assign the correct label for the unseen instance. Compared to PLL, UPLL suffers from a loss of supervision information in non-candidate labels. Merely conducting disambiguation solely on the candidate label set, which lacks the correct label, fails to effectively utilize the corresponding instance during training. Hence, we need to perform refinement, which attempts to sieve correct labels in non-candidate labels to recover supervision information.

To formulate our fusion framework FREDIS, we let the output of the classifier f satisfy $f(\boldsymbol{x}) \in \Delta^{c-1}$, which denotes the c-dimensional probability simplex, and the predict label of the classifier f given the instance $\boldsymbol{x}_i$ is denoted by $y_{f(\boldsymbol{x}_i)} = \arg\max_j f_j(\boldsymbol{x}_i)$. To simplify the following theoretical induction, $\eta_j(\boldsymbol{x}) = \mathbb{P}[y_i = j|\boldsymbol{x}_i]$ is employed to denote the posterior probability of the label $y_i = j$ given the instance $\boldsymbol{x}_i$, and the multi-class Bayes optimal classifier prediction is denoted by $\eta^\star(\boldsymbol{x}) = \arg\max_j \eta_j(\boldsymbol{x})$, which is also the correct label $y_{\boldsymbol{x}}$ according to (Zheng et al., 2020). A classifier is considered consistent with or approximating the Bayes optimal classifier when its predictions align precisely with those of the Bayes optimal classifier.

4. The Proposed Method

4.1. Overview

In this section, a fusion framework of refinement and disambiguation for UPLL named FREDIS is introduced following theoretical guarantees. Theoretically, we first prove that for a given UPLL distribution, there exist appropriate thresholds, which are set on the output of the classifier, to sieve correct and incorrect labels, inducing refinement and disambiguation. Then we prove that if the disambiguated labels overwhelms the refined labels, the boundary of the gap between the output of the classifier and the posterior probability will decrease after one fusion round of refinement and disambiguation. Additionally, disambiguation is proved capable of eliminating incorrect labels previously misidentified by refinement. Finally, we prove that the classifier trained in our framework FREDIS could have a good chance to approximate the Bayes optimal classifier.

Practically, FREDIS repeatedly trains a randomly initialized classifier on the UPLL dataset for enough epochs, and then updates the candidate labels in the dataset with the refinement and disambiguation procedure. During each fusion round of refinement and disambiguation, we calculate the difference between the output of the classifier on the predicted label and that on the rest labels, where we set a low threshold for refinement and a high threshold for disambiguation, respectively. We randomly sample a subset of the refined and disambiguated labels to control the sieving number. In this way of fusion, the predictive classifier can perform prediction for unseen instances.

4.2. The FREDIS framework

Recently weakly-supervised approaches (Feng et al., 2020; Gao & Zhang, 2021) assume that the output of deep neural networks with the softmax layer could be employed to directly approximate the posterior probability, which is empirically effective but theoretically unreasonable to some extent. It has only been proved by (Yu et al., 2018; Lv et al., 2020) that for a ordinary multi-class classifier g, if the hypothesis class is enough complex, given infinitely many data and the strict proper loss function ℓ such as the cross-entropy loss or the mean square loss, the optimal classifier $g^* = \arg\min_{g \in \mathcal{H}} \mathbb{E}_{(\boldsymbol{x}, y_{\boldsymbol{x}})}[\ell(g(\boldsymbol{x}), y_{\boldsymbol{x}})]$ can output the posterior probability , i.e., $\forall j \in \mathcal{Y}, g_j^*(\boldsymbol{x}) = \eta_j(\boldsymbol{x})$. In this paper, we also use a deep model with the softmax layer as our classifier but relax the approximation in UPLL by only assuming that the probabilistic output of the classifier

trained on a dataset with lower label ambiguity gets closer to the Bayesian posterior probability.

To formulate the assumption, we define a scoring function to decide the label ambiguity of an UPLL dataset, inspired by (Gong et al., 2021), and an (α, ϵ, ρ)-ambiguity bounded distribution of UPLL, inspired by (Cour et al., 2011).

Definition 1 *(Label Ambiguity Scoring Function) Let* $U(\tilde{\mathcal{D}}) = \sum_{(\boldsymbol{x},\tilde{S})\in\tilde{\mathcal{D}}} \sum_{j\in\mathcal{Y}} \mathbb{I}[j = y_{\boldsymbol{x}}, j \notin \tilde{S}] + \mathbb{I}[j \neq y_{\boldsymbol{x}}, j \in \tilde{S}]$ *denote the label ambiguity of the dataset* $\tilde{\mathcal{D}}$ *where* $\mathbb{I}$ *is the indicator function. A scoring function* $O(U(\tilde{\mathcal{D}})) : \mathbb{R}^+ \mapsto [0, 1]$ *is said to depict the concrete score of the label ambiguity if there exists* $\varepsilon > 0, \sigma > 1$ *making the following conditions holds:*

- *Suppose an data point* $(\boldsymbol{x}, \tilde{S})$ *in the dataset* $\tilde{\mathcal{D}}$*, which satisfies* $y_{\boldsymbol{x}} \notin \tilde{S}$*, is replaced by another point* $(\boldsymbol{x}, \tilde{S}')$*, which satisfies* $\tilde{S}' = \tilde{S} \cup \{y_{\boldsymbol{x}}\}$*, to form a new dataset* $\tilde{\mathcal{D}}'$*. Then* $O(U(\tilde{\mathcal{D}})) - O(U(\tilde{\mathcal{D}}')) > \sigma\varepsilon$ *will hold.*
- *Suppose an data point* $(\boldsymbol{x}, \tilde{S})$ *in the dataset* $\tilde{\mathcal{D}}$*, which satisfies* $j \in \tilde{S}$ *and* $j \neq y_{\boldsymbol{x}}$*, is replaced by another point* $(\boldsymbol{x}, \tilde{S}')$*, which satisfies* $\tilde{S}' = \tilde{S} \backslash \{j\}$*, to form a new dataset* $\tilde{\mathcal{D}}'$*. Then* $\varepsilon \leq O(U(\tilde{\mathcal{D}})) - O(U(\tilde{\mathcal{D}}')) \leq \sigma\varepsilon$ *will holds.*

Intuitively, the less label ambiguity the UPLL dataset $\tilde{\mathcal{D}}$ has, the smaller value the score function $O(U(\tilde{\mathcal{D}}))$ outputs. On the one hand, when one correct label is refined from non-candidate labels to candidate labels, the value of the score function $O(U(\tilde{\mathcal{D}}))$ will be reduced by more than $\sigma\varepsilon$. On the other hand, when one incorrect label is disambiguated from candidate labels to non-candidate labels, the value of the score function $O(U(\tilde{\mathcal{D}}))$ is reduced by at least ε. Here, we also assume that the gain of recovering one correct label is larger that that of eliminating one incorrect label in the scoring function.

Definition 2 (α, ϵ, ρ)***-ambiguity bounded distribution.*** *An UPLL distribution* $\mathbb{P}[\boldsymbol{x}, \tilde{S}]$ *is bounded by* (α, ϵ, ρ)*-ambiguity if there exists a subset* G *of the support of* $\mathbb{P}[\boldsymbol{x}, \tilde{S}]$*,* $G \subseteq \mathcal{X} \times \mathcal{C}$*, with probability mass at least* $1 - \rho$*, that is,* $\int_{(\boldsymbol{x},\tilde{S})\in G} \mathbb{P}[\boldsymbol{x}, \tilde{S}] \mathrm{d}\mu(\boldsymbol{x}, \tilde{S}) \geq 1 - \rho$*, integrated w.r.t the appropriate underlying measure* μ *on* $\mathcal{X} \times \mathcal{C}$*, for which when* $f(\boldsymbol{x}) = \arg\min_{f\in\mathcal{H}} \widehat{\mathcal{R}}(f)$ *and* $\forall \tilde{\mathcal{D}} \subseteq G$*,*

$$\sup_{(x,\tilde{S})\in\tilde{\mathcal{D}}, j\in\mathcal{Y}} |f_j(\boldsymbol{x}) - \eta_j(\boldsymbol{x})| \leq \alpha O(U(\tilde{\mathcal{D}})) + \epsilon, \quad (1)$$

where $\widehat{\mathcal{R}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), \tilde{S})$ *is the empirical risk estimator,* $\alpha \in (0, 1)$ *is used to resolve the scale problem, and* $\epsilon \in (0, 1)$ *is a minor value denoting the inherent difference between* f *and* η *influenced by the loss function, sample complexity, optimization and etc.*

Definition 2 indicates that for an UPLL dataset $\tilde{\mathcal{D}} \subseteq G$ in the (α, ϵ, ρ)-ambiguity bounded $\mathbb{P}[\boldsymbol{x}, \tilde{S}]$, the gap between f and η is bounded by the label ambiguity of the whole dataset D. If we can refine correct labels or disambiguate incorrect labels in the dataset $\tilde{\mathcal{D}}$, the boundary will be narrowed. From now on, we will assume:

Assumption 1 *The UPLL dataset* $\tilde{\mathcal{D}}$ *is always a subset of* G *in the* (α, ϵ, ρ)*-ambiguity bounded* $\mathbb{P}[\boldsymbol{x}, \tilde{S}]$*.*

Based on Assumption 1, we introduce the refinement theorem and the disambiguation theorem for our framework. The refinement theorem states that for each instance $\boldsymbol{x}$, there exists a threshold ζ to test the output difference of a classifier between the predictive label and the rest, and then decide which crowd correct labels hide themselves in. On the contrary, the disambiguation theorem states that there exists a threshold $\bar{\zeta}$ for the difference to determine which label is incorrect.

Theorem 1 *(Refinement) Under Assumption 1, suppose that for an instance* $\boldsymbol{x}$*, of which the correct label* $y_{\boldsymbol{x}} \notin \tilde{S}$*, and a constant* $\zeta = 2(\alpha O(U(\tilde{\mathcal{D}})) + \epsilon) - \eta_{y_{\boldsymbol{x}}}(\boldsymbol{x}) + \eta_{y_{f(\boldsymbol{x})}}(\boldsymbol{x})$*, there exists an instance-label level set* $I(f, \zeta) = \{(\boldsymbol{x}, j) \big| f_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) - f_j(\boldsymbol{x}) \leq \zeta, j \notin \tilde{S}\}$*. Then we have* $(\boldsymbol{x}, y_{\boldsymbol{x}}) \in I(f, \zeta)$*.*

The detailed proof can be found in Appendix A.1. Theorem 1 theoretically guarantees the refinement procedure in our framework FREDIS. It demonstrates that for an instance $\boldsymbol{x}_i$, there exists a special threshold ζ such that its correct label $y_{\boldsymbol{x}_i}$ satisfies the condition $f_{y_{f(\boldsymbol{x}_i)}}(\boldsymbol{x}_i) - f_{y_{\boldsymbol{x}_i}}(\boldsymbol{x}_i) \leq \zeta$ and $(\boldsymbol{x}, y_{\boldsymbol{x}})$ is included in the instance-label level set $I(f, \zeta)$. Hence, we can refine the correct label $y_{\boldsymbol{x}_i}$ for the instance $\boldsymbol{x}_i$ via performing $\tilde{S} \cup \{j\}$ for $(\tilde{S}, j) \in \{(\tilde{S}, j) | (\boldsymbol{x}, \tilde{S}) \in \tilde{\mathcal{D}}, (\boldsymbol{x}, j) \in I(f, \zeta)\}$ to sieve labels from non-candidate labels to candidate labels, recovering the reliability for the candidate label sets which do not contain the correct labels. Naturally, we need a theorem to guide disambiguation, which further purifies the supervison information of the UPLL dataset.

Theorem 2 *(Disambiguation) Under Assumption 1, suppose that for a constant* $\bar{\zeta} = 2(\alpha O(U(\tilde{\mathcal{D}}) + \epsilon)$*, there exists an instance-label level set* $\bar{I}(f, \bar{\zeta}) = \{(\boldsymbol{x}, j) \big| f_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) - f_j(\boldsymbol{x}) \geq \bar{\zeta}, j \in \tilde{S}\}$*. Then for any instance* $\boldsymbol{x}$*, we have* $(\boldsymbol{x}, y_{\boldsymbol{x}}) \notin \bar{I}(f, \bar{\zeta})$*.*

The detailed proof can be found in Appendix A.2. Theorem 2 provides a theoretical guarante for the disambiguation procedure in our framework FREDIS, which suggests that when an appropriate threshold $\bar{\zeta}$ is set, if a candiate label $j \in \tilde{S}_i$ of the instance $\boldsymbol{x}_i$ satisfies the condidtion $f_{y_{f(\boldsymbol{x})}} - f_j(\boldsymbol{x}) \geq \bar{\zeta}$, we can perform $\tilde{S}_i \setminus \{j\}$ to eliminate the incorrect label j from the candidate label set $\tilde{S}_i$.

Base on Theorem 1 and 2, we deduce the following theorem,

Algorithm 1 FREDIS

Input: The UPLL training set $\tilde{\mathcal{D}} = \{(\boldsymbol{x}_i, \tilde{S}_i) | 1 \leq i \leq n\}$, initial refinement and disambiguation thresholds ζ_0 and $\bar{\zeta}_0$, final refinement and disambiguation thresholds ζ_{end}, $\bar{\zeta}_{\text{end}}$, and total rounds R;

Initialize ζ with ζ_0 and $\bar{\zeta}$ with $\bar{\zeta}_0$;
for $r = 1, ..., R$ **do**
  Train the predictive model f on $\tilde{\mathcal{D}}$;
  Obtain instance-label level sets $I(f, \zeta)$ and $\bar{I}(f, \bar{\zeta})$;
  Randomly sample a subset $I'(f, \zeta)$ from $I(f, \zeta)$ and a subset $\bar{I}'(f, \bar{\zeta})$ from $\bar{I}(f, \bar{\zeta})$, which satisfy $\sigma|I'(f, \zeta)| \leq |\bar{I}'(f, \bar{\zeta})|$;
  for each $(\boldsymbol{x}, j)$ in $I'(f, \zeta)$ **do**
    Add the label j to the candidate label set $\tilde{S}$ via performing $\tilde{S} = \tilde{S} \cup \{j\}$;
  end for
  for each $(\boldsymbol{x}, j)$ in $\bar{I}'(f, \bar{\zeta})$ **do**
    Remove the label j from the candidate label set $\tilde{S}$ via performing $\tilde{S} = \tilde{S} \setminus \{j\}$;
  end for
  if $\zeta \leq \zeta_{\text{end}}$, $\bar{\zeta} \geq \bar{\zeta}_{\text{end}}$, and there is no change for all candidate labels **then**
    Increase ζ, Decrease $\bar{\zeta}$;
  end if
end for

Output: The final predictive classifier f.

which is theoretically not heuristically prepared for the fusion of refinement and disambiguation, where the boundary of the gap between f and η will be narrowed as the label ambiguity of the dataset decreases.

Theorem 3 *(One Round Boundary Narrowing) Under Assumption 1, for constants* $\zeta = 2(\alpha O(U(\tilde{\mathcal{D}})) + \epsilon) - \eta_{y_{\boldsymbol{x}}}(\boldsymbol{x}) + \eta_{y_{f(\boldsymbol{x})}}(\boldsymbol{x})$ *and* $\bar{\zeta} = 2(\alpha O(U(\tilde{\mathcal{D}})) + \epsilon)$*, suppose that there exist their corresponding instance-label level sets* $I(f, \zeta) = \{(\boldsymbol{x}, j) \big| f_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) - f_j(\boldsymbol{x}) \leq \zeta, j \notin \tilde{S}\}$ *and* $\bar{I}(f, \bar{\zeta}) = \{(\boldsymbol{x}, j) \big| f_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) - f_j(\boldsymbol{x}) \geq \bar{\zeta}, j \in \tilde{S}\}$ *for an* f *such that* $\sigma|I(f, \zeta)| \leq |\bar{I}(f, \bar{\zeta})|$*, where* $|\cdot|$ *denotes the cardinality. After performing one fusion round of refinement and disambiguation, the boundary of the gap* $|f_j(\boldsymbol{x}) - \eta_j(\boldsymbol{x})|$ *will be reduced at least* $\alpha\varepsilon$*.*

The detailed proof can be found in Appendix A.3. Theorem 3 demonstrates that one fusion round of refinement and disambiguation narrows the boundary of the gap $|f_j(\boldsymbol{x}) - \eta_j(\boldsymbol{x})|$ by at least $\alpha\varepsilon$. Therefore, the performance of the classifier could be guaranteed improved when we simultaneously perform refinement and disambiguation by one round according to the corresponding instance-label sets $I(f, \zeta)$ and $\bar{I}(f, \bar{\zeta})$, which satisfy $\sigma|I(f, \zeta)| \leq |\bar{I}(f, \bar{\zeta})|$, i.e. the disambiguated labels should be at least σ times more than the refined labels.

Besides, from Theorem 3, we immediately obtain the following corollary, which assets that although we add some incorrect labels at the same time as correct labels when we performing the refinement on non-candidate labels, the incorrect labels could be moved out later when we perform the disambiguation due to the continuous decrease of the boundary of the gap $|f_j(\boldsymbol{x}) - \eta_j(\boldsymbol{x})|$.

Corollary 1 *(Mistaken Incorrect Labels Eliminating) Suppose Assumption 1 holds, and assume an instance-label pair* $(\boldsymbol{x}, j) \in I(f, \zeta)$ *with the refinement threshold* $\zeta = 2(\alpha O(U(\tilde{\mathcal{D}})) + \epsilon) - \eta_{y_{\boldsymbol{x}'}}(\boldsymbol{x}') + \eta_{y_{f(\boldsymbol{x}')}}(\boldsymbol{x}')$*, leading the incorrect label* j *of the instance* $\boldsymbol{x}$ *to be mistaken into* $\tilde{S}$ *by the refinement procedure. Simultaneously, the disambiguation procedure removes incorrect labels from candidate labels according to* $\bar{I}(f, \bar{\zeta})$ *with the disambiguation threshold* $\bar{\zeta} = 2(\alpha O(U(\tilde{\mathcal{D}}) + \epsilon)$*. Then after* $R \geq \frac{1}{4\alpha\varepsilon}(2(\min_{(x,j)\in\bar{I}(f,\bar{\zeta})} f_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) - f_j(\boldsymbol{x})) + (\eta_j(\boldsymbol{x}) - \min_j \eta_j(\boldsymbol{x})))$ *rounds, the incorrect label* j *will be moved out from* $\tilde{S}$*.*

The proof is provided in Appendix A.4. Corollary 1 indicates that mistaken incorrect labels in the refinement procedure could be guaranteed to be eliminated in the disambiguation procedure to further purify the candidate label set, as we iteratively perform $R \geq \frac{1}{4\alpha\varepsilon}(2(\min_{(x,j)\in\bar{I}(f,\bar{\zeta})} f_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) - f_j(\boldsymbol{x})) + (\eta_j(\boldsymbol{x}) - \min_j \eta_j(\boldsymbol{x})))$ fusion rounds of refinement and disambiguation. Additionally, Theorem 3 and Corollary 1 encourage us to progressively refine and disambiguate labels. Though we can set the constant $\zeta \geq \max_{\boldsymbol{x}} 2(\alpha O(U(\tilde{\mathcal{D}})) + \epsilon) - \eta_{y_{\boldsymbol{x}}}(\boldsymbol{x}) + \eta_{y_{f(\boldsymbol{x})}}(\boldsymbol{x})$ to directly recover all correct labels for the dataset $\tilde{\mathcal{D}}$ in one round refinement, the supervision information will be impaired a lot by the mistaken incorrect labels, which may not be purified by disambiguation and cannot guarantees the improvement of the classifier f.

4.3. Implementation Details

According to the above theorems, we are able to formulate our fusion framework of refinement and disambiguation FREDIS aimed at UPLL, which simultaneously refines correct labels out of non-candidate labels and disambiguates incorrect labels in the candidate labels in a progressive way. To be specific, in each round, we start with training our predictive model f on the dataset $\tilde{\mathcal{D}}$ using a weighted cross entropy loss with consistency regularization (Wu et al., 2022) until getting a relatively acceptable approximation about η. Then, we can begin to perform the refinement and disambiguation procedure to update candidate labels. Inspired by Theorem 3, we randomly sample subsets from the instance-label level sets of refinement and disambiguation, controlling the disambiguated labels to overwhelm the refined labels. According to the sampled subsets, we add and remove the candidate labels to recover the reliability of

candidate label sets and purify the supervision information.

Certainly, if current thresholds incur no change for any candidate label set $\tilde{S}$ in the dataset $\tilde{\mathcal{D}}$, it is time to slowly increase the refinement threshold ζ and decrease the disambiguation threshold $\bar{\zeta}$ until we can refine correct labels and disambiguate incorrect labels to update the dataset $\tilde{\mathcal{D}}$ again. We repeat the training and updating procedure for R rounds until the classifier will not be improved. Algorithm 1 is the pseudo code depicting our framework.

4.4. Theoretical Analysis

Here, we further analyze the relationship between the classifier f trained by our algorithm FREDIS and the Bayes optimal classifier $\eta^\star$. Before proving our classifier f has a good chance to be consistent with the Bayes optimal classifier $\eta^\star$, the Tsybakov condition (Chaudhuri & Dasgupta, 2014; Belkin et al., 2018; Qiao et al., 2019) is assumed to be hold around the decision boundary of the Bayes optimal classifier $\eta^\star$.

Assumption 2 *(Tsybakov Condition). Let* $s_{\boldsymbol{x}} = \arg\max_{j \neq y_{\boldsymbol{x}}} \eta_j(\boldsymbol{x})$. *Suppose that there exists constants* C, $\lambda > 0$, *and* $t_0 \in (0, 1)$, *such that for all* $t \leq t_0$,

$$\mathbb{P}\left[\eta_{y_{\boldsymbol{x}}}(\boldsymbol{x}) - \eta_{s_{\boldsymbol{x}}}(\boldsymbol{x}) \leq t\right] \leq Ct^{\lambda}, \tag{2}$$

Assumption 2 quantifies how well classes are separated on the decision boundary $\{\boldsymbol{x} : \eta_{y_{\boldsymbol{x}}}(\boldsymbol{x}) = \eta_{s_{\boldsymbol{x}}}(\boldsymbol{x})\}$, and states that the uncertainty of η, denoted by the margin $\eta_{y_{\boldsymbol{x}}}(\boldsymbol{x}) - \eta_{s_{\boldsymbol{x}}}(\boldsymbol{x})$, is bounded. Then we provide a theoretical guarantee for the consistency of the classifier f trained in our framework with respect to the Bayes optimal classifier $\eta^\star$. Empirical experiments conducted in (Zheng et al., 2020) validate the bound, which also indicates the constant C satisfies $C \leq 1$ and the constant $\lambda \geq 1$.

Theorem 4 *(Bayes Consistency) Under Assumption 2 and Suppose* $\mathbb{P}[\boldsymbol{x}, \tilde{S}]$ *is* (α, ϵ, ρ)*-ambiguity bounded, after running* R *fusion rounds of refinement of disambiguation in the algorithm* FREDIS*, we have:*

$$\mathbb{P}[y_{f_{final}(\boldsymbol{x})} = \eta^\star(\boldsymbol{x})] \geq \left(1 - C\left(2\left(\alpha O(U(\tilde{\mathcal{D}})) + \epsilon - \alpha\varepsilon R\right)\right)^{\lambda}\right)(1 - \rho) \tag{3}$$

The proof of Theorem 4 is provided in Appendix A.5. Theorem 4 demonstrates that the classifier f trained in our proposed framework FREDIS could guaranteed to gradually approximate the Bayes optimal classifies $\eta^\star$. Furthermore, the theorem reveals that several factors influence this approximation, including the initial label ambiguity of a UPLL dataset, the maximum fusion rounds, the range of the subset G in (α, ϵ, ρ)-ambiguity bounded distribution, and certain inherent constants.

5. Experiments

5.1. Datasets

In order to validate the effectiveness of our algorithm, we employ four benchmark datasets, which are widely used to be corrupted for validation in deep PLL, including `Kuzushiji-MNIST` (Clanuwat et al., 2018), `Fashion-MNIST` (Xiao et al., 2017), `CIFAR-10` and `CIFAR-100` (Krizhevsky et al., 2009). Each dataset is partitioned into training, validation and test datasets with the proportion of $80\%/10\%/10\%$ respectively.

For each benchmark dataset, we manually corrupt its training dataset into partially labeled versions with unreliable candidate label sets, inspired by the generation process in (Lv et al., 2023). Specifically, the generation process can be divided into three procedures. Firstly, we sample a possibly incorrect label for each instance according to Categorical Distribution, i.e, $\mathbb{P}[\bar{y} = y_{\boldsymbol{x}} | y_{\boldsymbol{x}}] = 1 - \gamma_1$ $(0 < \gamma_1 < 1)$ and $\mathbb{P}[\bar{y} \neq y_{\boldsymbol{x}} | y_{\boldsymbol{x}}] = \frac{\gamma_1}{c-1}$, where $\bar{y}$ represents the possibly incorrect label generated by this procedures and γ_1 denotes the possibility that the sampled label $\bar{y} \neq y_{\boldsymbol{x}}$ given the correct label $y_{\boldsymbol{x}}$. Then we sample the rest candidate labels for each instance according to Bernoulli Distribution. Given the correct label $y_{\boldsymbol{x}}$ of an instance $\boldsymbol{x}$, each incorrect label has the probability γ_2 $(0 < \gamma_2 < 1)$ to be selected into the candidate label set. This procedure will generate a label set $\bar{S}$ without the correct label $y_{\boldsymbol{x}}$. Finally, we create the unreliable candidate label set $\tilde{S}$ via performing $\tilde{S} = \{\bar{y}\} \cup \bar{S}$ for each instance in the benchmark datasets. Compared with the generation process in (Lv et al., 2023), our generation process is experimentally simple and convenient to directly control the unreliable level (overall proportion of unreliable candidate label sets) and partial level (average number of candidate labels) with γ_1 and γ_2 respectively.

5.2. Baselines

In this paper, the proposed framework FREDIS is compared against seven well-established PLL algorithms based on deep neural network. 1) PLCR (Wu et al., 2022), which utilizes consistency regularization regularization by matching the outputs of the classifier on multiple augmentations of each instance to a conformal label distribution. 2) PICO (Wang et al., 2022), which uses an entropy-based regularization item as well as the ensemble technique. 3) CAVL (Zhang et al., 2021), which is a discriminative approach and identifies correct labels from candidate labels by class activation value. 4) LWS (Wen et al., 2021), which introduces a leverage parameter considering the trade-offs between

Table 1. Test accuracy (%) on benchmark datasets `Fashion-MNIST` and `Kuzushiji-MNIST` under different levels of γ_1 and γ_2 when FREDIS is compared with six PLL methods. Average accuracy and standard deviation over 5 trials are reported.

FMNIST	$\gamma_1 = 0.1$ $\gamma_2 = 0.5$	$\gamma_1 = 0.3$ $\gamma_2 = 0.3$	$\gamma_1 = 0.5$ $\gamma_2 = 0.1$	$\gamma_1 = 0.3$ $\gamma_2 = 0.5$	$\gamma_1 = 0.5$ $\gamma_2 = 0.3$
FREDIS	$\mathbf{91.76 \pm 0.09}$	$\mathbf{90.56 \pm 0.11}$	$\mathbf{89.21 \pm 0.15}$	$\mathbf{87.45 \pm 0.12}$	$\mathbf{85.84 \pm 0.16}$
PLCR	$90.97 \pm 0.19\bullet$	$89.78 \pm 0.13\bullet$	$88.18 \pm 0.28\bullet$	$86.29 \pm 0.16\bullet$	$84.41 \pm 0.46\bullet$
PICO	$90.47 \pm 0.11\bullet$	$89.15 \pm 0.22\bullet$	$87.83 \pm 0.38\bullet$	$83.42 \pm 0.64\bullet$	$80.37 \pm 0.89\bullet$
CAVL	$57.89 \pm 8.79\bullet$	$76.98 \pm 2.01\bullet$	$65.86 \pm 1.26\bullet$	$19.75 \pm 3.77\bullet$	$51.58 \pm 7.77\bullet$
LWS	$88.25 \pm 2.01\bullet$	$88.60 \pm 0.19\bullet$	$64.59 \pm 2.13\bullet$	$80.67 \pm 2.76\bullet$	$65.79 \pm 2.14\bullet$
PRODEN	$84.55 \pm 0.20\bullet$	$84.51 \pm 0.49\bullet$	$84.04 \pm 0.20\bullet$	$78.73 \pm 0.62\bullet$	$78.09 \pm 0.43\bullet$
RC	$85.43 \pm 0.41\bullet$	$84.94 \pm 0.20\bullet$	$84.14 \pm 0.40\bullet$	$78.48 \pm 1.10\bullet$	$77.21 \pm 0.97\bullet$
CC	$86.76 \pm 0.54\bullet$	$85.79 \pm 0.15\bullet$	$84.16 \pm 0.46\bullet$	$81.67 \pm 0.27\bullet$	$79.82 \pm 0.74\bullet$
KMNIST	$\gamma_1 = 0.1$ $\gamma_2 = 0.5$	$\gamma_1 = 0.3$ $\gamma_2 = 0.3$	$\gamma_1 = 0.5$ $\gamma_2 = 0.1$	$\gamma_1 = 0.3$ $\gamma_2 = 0.5$	$\gamma_1 = 0.5$ $\gamma_2 = 0.3$
FREDIS	$\mathbf{95.15 \pm 0.05}$	$\mathbf{93.54 \pm 0.28}$	$\mathbf{91.02 \pm 0.15}$	$\mathbf{88.41 \pm 0.58}$	$\mathbf{84.75 \pm 0.39}$
PLCR	$94.07 \pm 0.57\bullet$	$92.83 \pm 0.36\bullet$	$90.06 \pm 0.15\bullet$	$85.78 \pm 0.74\bullet$	$81.81 \pm 0.74\bullet$
PICO	$88.25 \pm 2.87\bullet$	$91.45 \pm 0.37\bullet$	$87.14 \pm 0.67\bullet$	$71.67 \pm 3.39\bullet$	$59.43 \pm 6.60\bullet$
CAVL	$88.27 \pm 2.12\bullet$	$86.83 \pm 1.05\bullet$	$83.91 \pm 1.21\bullet$	$75.60 \pm 2.80\bullet$	$72.73 \pm 2.89\bullet$
LWS	$85.52 \pm 4.51\bullet$	$89.39 \pm 1.05\bullet$	$39.25 \pm 3.06\bullet$	$62.79 \pm 8.29\bullet$	$71.07 \pm 1.64\bullet$
PRODEN	$92.01 \pm 0.47\bullet$	$91.78 \pm 0.24\bullet$	$89.53 \pm 0.61\bullet$	$82.96 \pm 1.06\bullet$	$81.06 \pm 0.69\bullet$
RC	$93.36 \pm 0.32\bullet$	$92.38 \pm 0.52\bullet$	$89.72 \pm 0.71\bullet$	$85.87 \pm 0.82\bullet$	$83.28 \pm 0.55\bullet$
CC	$92.70 \pm 1.00\bullet$	$90.08 \pm 0.48\bullet$	$86.84 \pm 0.34\bullet$	$80.39 \pm 0.40\bullet$	$77.32 \pm 1.58\bullet$

losses on candidate labels and non-candidate labels. 5) PRODEN (Lv et al., 2020), which uses a classifier-consistent risk estimator and updates the label weights in it with the output of the model. 6) RC (Feng et al., 2020), which uses a risk-consistent estimator utilizing the output of the model to calculate the posterior. 7) CC (Feng et al., 2020), which uses a classifier-consistent risk estimator deriving from the transition matrix.

When implementing each algorithm, we employ the same model, optimization, batch size and data augmentation strategy on the same dataset for fairness. For simple benchmark datasets such as `Kuzushiji-MNIST` and `Fashion-MNIST`, we only choose LeNet as our classifier. For relatively complex datasets such as `CIFAR-10` and `CIFAR-100`, we choose ResNet-32 (He et al., 2016) as our backbone. The model is optimized by stochastic gradient decent (SGD) with momentum 0.9. We train each model with the batch size set to 256. The data augmentation strategy is the same as that employed by (Wu et al., 2022). For hyper-parameters like learning rate and weight decay, we select the most appropriate one for each algorithm to ensure the best model parameters according to their performances on the validating datasets. To alleviate overfitting, the training procedure of a model will be early stopped if its performance on the validation dataset does not improve in 50 epochs. Finally, we run 5 trials based on different random seeds for each method to record the performance.

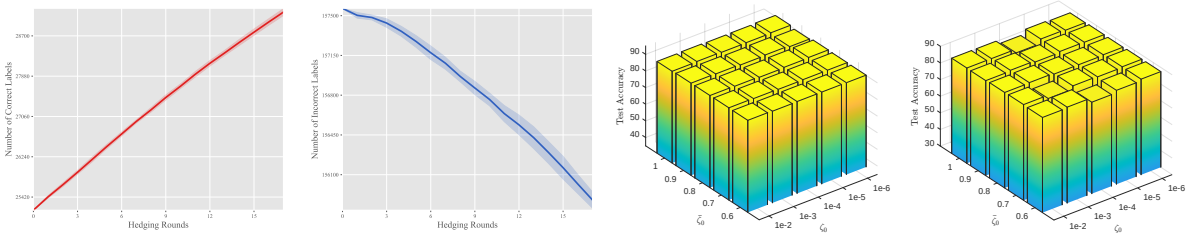
5.3. Experimental Results

Table 1 and 2 summarizes the classification accuracy of each comparison approach on manually corrupted UPLL benchmark datasets. We control the unreliable level and partial level respectively at low, middle and high levels. For the 10-class datasets `Fashion-MNIST`, `Kuzushiji-MNIST` and `CIFAR-10`, γ_1 and γ_2 both take values in $\{0.1, 0.3, 0.5\}$. For the 100-class dataset `CIFAR-100`, γ_1 takes values in $\{0.1, 0.3, 0.5\}$ while γ_2 is controlled in $\{0.01, 0.05, 0.1\}$. The performance at both the high unreliable level and the high partial level is not recorded due to that the approaches do not work in such a case. Due to space constraints, we present a partial set of results in Table 1 and 2, while the remaining results can be found in Appendix A.6. The best results are highlighted in bold. In addition, $\bullet$/$\circ$ indicates whether FREDIS is statistically superior/inferior to the comparing approach on each dataset (pairwise t-test at 0.05 significance level).

From the tables, we can observe that FREDIS outperforms or is at least comparable with all of the other comparative approaches, although the compared approaches PLCR and PICO seem to exhibit some robustness in handling unreliable candidate label sets. For the comparing results on `CIFAR-100` with $\gamma_2 = 0.1$, the superiority of our method

Table 2. Test accuracy (%) on benchmark datasets `CIFAR-10` and `CIFAR-100` under different levels of γ_1 and γ_2 when FREDIS is compared with six PLL methods. Average accuracy and standard deviation over 5 trials are reported.

CIFAR10	$\gamma_1 = 0.1$ $\gamma_2 = 0.5$	$\gamma_1 = 0.3$ $\gamma_2 = 0.3$	$\gamma_1 = 0.5$ $\gamma_2 = 0.1$	$\gamma_1 = 0.3$ $\gamma_2 = 0.5$	$\gamma_1 = 0.5$ $\gamma_2 = 0.3$
FREDIS	$\mathbf{87.42 \pm 0.21}$	81.02 ± 0.60	$\mathbf{75.80 \pm 0.24}$	$\mathbf{65.15 \pm 0.13}$	$\mathbf{51.29 \pm 0.25}$
PLCR	84.64 ± 0.35•	77.95 ± 0.67•	68.59 ± 0.65•	46.13 ± 0.92•	37.93 ± 2.49•
PICO	86.90 ± 0.16•	$\mathbf{81.59 \pm 0.50}$	74.33 ± 0.30•	61.39 ± 2.64•	30.32 ± 7.58•
CAVL	39.14 ± 6.25•	48.90 ± 1.90•	50.96 ± 0.49•	14.31 ± 2.08•	12.65 ± 2.72•
LWS	18.19 ± 0.71•	18.64 ± 1.63•	15.13 ± 0.93•	12.49 ± 2.34•	18.81 ± 1.06•
PRODEN	76.23 ± 0.58•	73.95 ± 0.78•	68.08 ± 1.06•	58.77 ± 1.52•	39.39 ± 2.85•
RC	79.17 ± 0.73•	76.36 ± 1.11•	70.98 ± 0.60•	63.82 ± 0.86•	50.88 ± 4.00•
CC	72.06 ± 0.38•	68.66 ± 0.98•	64.09 ± 1.45•	43.57 ± 2.36•	33.16 ± 4.00•
CIFAR100	$\gamma_1 = 0.1$ $\gamma_2 = 0.1$	$\gamma_1 = 0.3$ $\gamma_2 = 0.05$	$\gamma_1 = 0.5$ $\gamma_2 = 0.01$	$\gamma_1 = 0.3$ $\gamma_2 = 0.1$	$\gamma_1 = 0.5$ $\gamma_2 = 0.05$
FREDIS	$\mathbf{60.31 \pm 0.10}$	$\mathbf{56.15 \pm 0.18}$	$\mathbf{50.72 \pm 0.18}$	$\mathbf{50.14 \pm 0.48}$	$\mathbf{48.05 \pm 0.20}$
PLCR	51.02 ± 0.81•	51.87 ± 0.25•	48.27 ± 0.50•	27.99 ± 0.48•	35.44 ± 2.14•
PICO	43.52 ± 1.60•	53.98 ± 0.57•	47.76 ± 0.71•	31.62 ± 2.42•	40.00 ± 0.91•
CAVL	22.14 ± 0.84•	31.07 ± 2.03•	29.55 ± 2.99•	17.29 ± 1.46•	23.29 ± 1.30•
LWS	10.06 ± 1.63•	5.59 ± 0.24•	5.05 ± 0.55•	6.28 ± 0.80•	8.01 ± 0.63•
PRODEN	42.96 ± 0.45•	49.39 ± 0.76•	47.05 ± 0.29•	26.29 ± 0.47•	35.79 ± 1.24•
RC	51.86 ± 0.32•	52.49 ± 0.40•	49.07 ± 0.57•	37.15 ± 0.59•	40.65 ± 0.58•
CC	52.99 ± 0.36•	50.04 ± 0.80•	47.07 ± 0.75•	44.51 ± 0.63•	40.11 ± 0.64•



(a) `Correct Labels` (b) `Incorrect Labels` (c) `Vary` ζ_0 `and` $\bar{\zeta}_0$ (d) `Vary` ζ_0 `and` $\bar{\zeta}_0$

Figure 1. Further analysis of FREDIS on `Fashion-MNIST` and `CIFAR-10`.

is more significant when compared to that on previous simpler datasets, owing to that more incorrect labels in the corrupted dataset could be used by the disambiguation procedure to progressively alleviate the risk to supervision information brought from the mistaken incorrect labels in the refinement procedure.

5.4. Further Analysis

Figure 1(a) and 1(b) illustrates the variation curves of our method FREDIS on `CIFAR-10` with $\gamma_1 = 0.5$ and $\gamma_2 = 0.3$. More details under other cases such as the variation curves on `Fashion-MNIST` with $\gamma_1 = 0.5$ and $\gamma_2 = 0.1$ can be referred to in Appendix A.6. We can see that correct labels in candidate label sets increase and incorrect candidate labels decrease after several fusion rounds, of which the changes are stable. This means that the reliability of candidate label sets is being recovered by refinement and the supervision information is being purified by dismabiguation during the training and fusion updating process.

In addition, we conduct the sensitivity analysis about the threshold ζ_0 and $\bar{\zeta}_0$ in FREDIS on `Fashion-MNIST` with $\gamma_1 = 0.3, \gamma_2 = 0.3$ and `CIFAR-10` with $\gamma_1 = 0.1, \gamma_2 = 0.5$, which is illustrated in Figure 1(c) and 1(d), respectively. We select ζ_0 from $\{1e-2, 1e-3, 1e-4, 1e-5, 1e-6\}$ and $\bar{\zeta}_0$ from $\{1, 0.9, 0.8, 0.7, 0.6\}$ according to validation datasets, and fix $\zeta_{\text{end}} = 0.9$ and $\bar{\zeta}_{\text{end}} = 0.1$. We can find that the performance of the proposed FREDIS is relatively stable over a range of threshold values, which indicates the

robustness and is desirable for algorithm design.

6. Conclusion

In this paper, we propose a novel framework FREDIS aimed at solving the problem of unreliable partial label learning problem. Different from partial label learning, we consider not only a disambiguation procedure but also a refinement procedure, and propose a theoretically-guaranteed framework, which could fuse refinement and disambiguation and train the classifier with an iteratively updated dataset with less and less label ambiguity and is theoretically guaranteed to eventually have a good chance to be consistent with the Bayes optimal classifier under mild assumptions for UPLL. Extensive experiments on widely used benchmark datasets validate the effectiveness of the proposed method.

Acknowledgments

This research was supported by the National Key Research & Development Plan of China (2018AAA0100104), the National Science Foundation of China (62206050, 62125602, and 62076063), China Postdoctoral Science Foundation (2021M700023), Jiangsu Province Science Foundation for Youths (BK20210220), Young Elite Scientists Sponsorship Program of Jiangsu Association for Science and Technology (TJ-2022-078), and the Big Data Computing Center of Southeast University.

References

Belkin, M., Hsu, D. J., and Mitra, P. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *Advances in neural information processing systems*, 31, 2018.

Chaudhuri, K. and Dasgupta, S. Rates of convergence for nearest neighbor classification. *Advances in Neural Information Processing Systems*, 27, 2014.

Chen, C.-H., Patel, V. M., and Chellappa, R. Learning from ambiguously labeled face images. *IEEE transactions on pattern analysis and machine intelligence*, 40(7):1653–1667, 2017.

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

Cour, T., Sapp, B., and Taskar, B. Learning from partial labels. *The Journal of Machine Learning Research*, 12: 1501–1536, 2011.

CROWDSOURCING, H. J. Why the power of the crowd is driving the future of business. *The International Achievement institute*, 2008.

Feng, L. and An, B. Partial label learning with self-guided retraining. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 3542–3549, 2019.

Feng, L., Lv, J., Han, B., Xu, M., Niu, G., Geng, X., An, B., and Sugiyama, M. Provably consistent partial-label learning. *arXiv preprint arXiv:2007.08929*, 2020.

Gao, Y. and Zhang, M.-L. Discriminative complementary-label learning with weighted loss. In *International Conference on Machine Learning*, pp. 3587–3597. PMLR, 2021.

Gong, C., Shi, H., Liu, T., Zhang, C., Yang, J., and Tao, D. Loss decomposition and centroid estimation for positive and unlabeled learning. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):918–932, 2019.

Gong, X., Yuan, D., and Bao, W. Understanding partial multi-label learning via mutual information. *Advances in Neural Information Processing Systems*, 34:4147–4156, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hüllermeier, E. and Beringer, J. Learning from ambiguously labeled examples. *Intelligent Data Analysis*, 10(5):419–439, 2006.

Jin, R. and Ghahramani, Z. Learning with multiple labels. In *NIPS*, volume 2, pp. 897–904. Citeseer, 2002.

Kamnitsas, K., Castro, D., Le Folgoc, L., Walker, I., Tanno, R., Rueckert, D., Glocker, B., Criminisi, A., and Nori, A. Semi-supervised learning via compact latent space clustering. In *International Conference on Machine Learning*, pp. 2459–2468. PMLR, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lian, Z., Xu, M., Chen, L., Sun, L., Liu, B., and Tao, J. Irnet: Iterative refinement network for noisy partial label learning, 2023.

Liu, L. and Dietterich, T. G. A conditional multinomial mixture model for superset label learning. In *Advances in neural information processing systems*, pp. 548–556. Citeseer, 2012.

Lu, N., Niu, G., Menon, A. K., and Sugiyama, M. On the minimal supervision for training any binary classifier from only unlabeled data. *arXiv preprint arXiv:1808.10585*, 2018.

Luo, J. and Orabona, F. Learning from candidate labeling sets. Technical report, MIT Press, 2010.

Lv, J., Xu, M., Feng, L., Niu, G., Geng, X., and Sugiyama, M. Progressive identification of true labels for partial-label learning. In *International Conference on Machine Learning*, pp. 6500–6510. PMLR, 2020.

Lv, J., Liu, B., Feng, L., Xu, N., Xu, M., An, B., Niu, G., Geng, X., and Sugiyama, M. On the robustness of average losses for partial-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. in press, 2023.

Nguyen, N. and Caruana, R. Classification with partial labels. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 551–559, 2008.

Ni, P., Zhao, S.-Y., Dai, Z.-G., Chen, H., and Li, C.-P. Partial label learning via conditional-label-aware disambiguation. *Journal of Computer Science and Technology*, 36(3):590–605, 2021.

Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1944–1952, 2017.

Qiao, X., Duan, J., and Cheng, G. Rates of convergence for large-scale nearest neighbor classification. *Advances in neural information processing systems*, 32, 2019.

Robbins, H. and Monro, S. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Wang, D.-B., Zhang, M.-L., and Li, L. Adaptive graph guided disambiguation for partial label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Wang, H., Xiao, R., Li, Y., Feng, L., Niu, G., Chen, G., and Zhao, J. Pico: Contrastive label disambiguation for partial label learning. *arXiv preprint arXiv:2201.08984*, 2022.

Wen, H., Cui, J., Hang, H., Liu, J., Wang, Y., and Lin, Z. Leveraged weighted loss for partial label learning. In *International Conference on Machine Learning*, pp. 11091–11100. PMLR, 2021.

Wu, D.-D., Wang, D.-B., and Zhang, M.-L. Revisiting consistency regularization for deep partial label learning. In *International Conference on Machine Learning*, pp. 24212–24225. PMLR, 2022.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

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

Xu, N., Qiao, C., Geng, X., and Zhang, M.-L. Instance-dependent partial label learning. *Advances in Neural Information Processing Systems*, 34, 2021b.

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

Yao, Y., Deng, J., Chen, X., Gong, C., Wu, J., and Yang, J. Deep discriminative cnn with temporal ensembling for ambiguously-labeled image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12669–12676, 2020a.

Yao, Y., Gong, C., Deng, J., and Yang, J. Network cooperation with progressive disambiguation for partial label learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 471–488. Springer, 2020b.

Yu, F. and Zhang, M.-L. Maximum margin partial label learning. In *Asian conference on machine learning*, pp. 96–111. PMLR, 2016.

Yu, X., Liu, T., Gong, M., and Tao, D. Learning with biased complementary labels. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 68–83, 2018.

Zeng, Z., Xiao, S., Jia, K., Chan, T.-H., Gao, S., Xu, D., and Ma, Y. Learning by associating ambiguously labeled images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 708–715, 2013.

Zhang, F., Feng, L., Han, B., Liu, T., Niu, G., Qin, T., and Sugiyama, M. Exploiting class activation value for partial-label learning. In *International Conference on Learning Representations*, 2021.

Zhang, M.-L. and Yu, F. Solving the partial label learning problem: An instance-based approach. In *Twenty-fourth international joint conference on artificial intelligence*, 2015.

Zhang, M.-L., Zhou, B.-B., and Liu, X.-Y. Partial label learning via feature-aware disambiguation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1335–1344, 2016.

Zhang, M.-L., Yu, F., and Tang, C.-Z. Disambiguation-free partial label learning. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2155–2167, 2017.

Zheng, S., Wu, P., Goswami, A., Goswami, M., Metaxas, D., and Chen, C. Error-bounded correction of noisy labels. In *International Conference on Machine Learning*, pp. 11447–11457. PMLR, 2020.

Zhou, Z.-H. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.

A. Appendix

A.1. Proof of Theorem 1

As for Theorem 1, we provide a proof by contradiction to carry it out in a simple-minded pattern.

Proof. Suppose that $(\boldsymbol{x}, y_{\boldsymbol{x}}) \notin I(f, \zeta)$. Then we have $f_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) - f_{y_{\boldsymbol{x}}}(\boldsymbol{x}) > \zeta = 2(\alpha O(U(\tilde{\mathcal{D}})) + \epsilon) - \eta_{y_{\boldsymbol{x}}}(\boldsymbol{x}) + \eta_{y_{f(\boldsymbol{x})}}(\boldsymbol{x})$, i.e.,

$$[f_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) - \eta_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) - (\alpha O(U(\tilde{\mathcal{D}})) + \epsilon)] - [f_{y_{\boldsymbol{x}}}(\boldsymbol{x}) - \eta_{y_{\boldsymbol{x}}}(\boldsymbol{x}) + (\alpha O(U(\tilde{\mathcal{D}})) + \epsilon)] > 0. \tag{4}$$

Due to that Assumption 1 holds, we have $|f_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) - \eta_{y_{f(\boldsymbol{x})}}(\boldsymbol{x})| \leq \alpha O(U(\tilde{\mathcal{D}})) + \epsilon$ and $|f_{y_{\boldsymbol{x}}}(\boldsymbol{x}) - \eta_{y_{\boldsymbol{x}}}(\boldsymbol{x})| \leq \alpha O(U(\tilde{\mathcal{D}})) + \epsilon$, which means that

$$-(\alpha O(U(\tilde{\mathcal{D}})) + \epsilon) \leq f_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) - \eta_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) \leq (\alpha O(U(\tilde{\mathcal{D}})) + \epsilon),$$

$$f_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) - \eta_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) - (\alpha O(U(\tilde{\mathcal{D}})) + \epsilon) \leq 0, \tag{5}$$

and

$$-(\alpha O(U(\tilde{\mathcal{D}})) + \epsilon) \leq f_{y_{\boldsymbol{x}}}(\boldsymbol{x}) - \eta_{y_{\boldsymbol{x}}}(\boldsymbol{x}) \leq (\alpha O(U(\tilde{\mathcal{D}})) + \epsilon),$$

$$f_{y_{\boldsymbol{x}}}(\boldsymbol{x}) - \eta_{y_{\boldsymbol{x}}}(\boldsymbol{x}) + (\alpha O(U(\tilde{\mathcal{D}})) + \epsilon) \geq 0. \tag{6}$$

Hence, according to Eq.(5) and Eq.(6),

$$[f_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) - \eta_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) - (\alpha O(U(\tilde{\mathcal{D}})) + \epsilon)] - [f_{y_{\boldsymbol{x}}}(\boldsymbol{x}) - \eta_{y_{\boldsymbol{x}}}(\boldsymbol{x}) + (\alpha O(U(\tilde{\mathcal{D}})) + \epsilon)] \leq 0 \tag{7}$$

This contradiction between Eq.(4) and Eq.(7) suggests that $(x, y_{\boldsymbol{x}}) \in I(f, \zeta)$ and proves Theorem 1.

A.2. Proof of Theorem 2

Proof. To prove Theorem 2, we need to prove that $f_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) - f_{y_{\boldsymbol{x}}}(\boldsymbol{x}) \leq \bar{\zeta} = 2(\alpha O(U(\tilde{\mathcal{D}})) + \epsilon)$ under Assumption 1. Then $(\boldsymbol{x}, y_{\boldsymbol{x}}) \notin \bar{I}(f, \bar{\zeta})$ will hold. Since the posterior probability of the correct label is larger than that of the rest, i.e., $\forall j \in \mathcal{Y}, \eta_{y_{\boldsymbol{x}}}(\boldsymbol{x}) - \eta_j(\boldsymbol{x}) \geq 0$, we have $\eta_{y_{\boldsymbol{x}}}(\boldsymbol{x}) - \eta_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) \geq 0$.

Hence, for an instance $\boldsymbol{x}$ and the classifier f,

$$\begin{aligned} f_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) - f_{y_{\boldsymbol{x}}}(\boldsymbol{x}) &\leq f_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) - f_{y_{\boldsymbol{x}}}(\boldsymbol{x}) + (\eta_{y_{\boldsymbol{x}}}(\boldsymbol{x}) - \eta_{y_{f(\boldsymbol{x})}}(\boldsymbol{x})) \\ &= (f_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) - \eta_{y_{f(\boldsymbol{x})}}(\boldsymbol{x})) + (\eta_{y_{\boldsymbol{x}}}(\boldsymbol{x}) - f_{y_{\boldsymbol{x}}}(\boldsymbol{x})) \\ &\leq |f_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) - \eta_{y_{f(\boldsymbol{x})}}(\boldsymbol{x})| + |f_{y_{\boldsymbol{x}}}(\boldsymbol{x}) - \eta_{y_{\boldsymbol{x}}}(\boldsymbol{x})| \\ &\leq 2(\alpha O(U(\tilde{\mathcal{D}})) + \epsilon). \end{aligned} \tag{8}$$

The proof has been completed.

A.3. Proof of Theorem 3

Proof. Let us review the definition of the whole label ambiguity for a dataset and the corresponding scoring function: $U(\tilde{\mathcal{D}}) = \sum_{(\boldsymbol{x}, \tilde{S}) \in \tilde{\mathcal{D}}} \sum_j \mathbb{I}_{\{j = y_{\boldsymbol{x}}, j \notin \tilde{S}\}} + \mathbb{I}_{\{j \neq y_{\boldsymbol{x}}, j \in \tilde{S}\}}$ represent the whole ambiguity of the dataset $\tilde{\mathcal{D}}$ and a scoring function $O(U(\tilde{\mathcal{D}})) : \mathbb{R}^+ \mapsto [0, 1]$ depict the concrete score of the ambiguity. As we mention before, the scoring function has two characteristics, which can help us understand the refinement and disambiguation procedure:

1) Suppose an data point $(\boldsymbol{x}, \tilde{S})$ in the dataset $\tilde{\mathcal{D}}$, which satisfies $y_{\boldsymbol{x}} \notin \tilde{S}$, is replaced by another point $(\boldsymbol{x}, \tilde{S}')$, which satisfies $\tilde{S}' = \tilde{S} \cup \{y_{\boldsymbol{x}}\}$, to form a new dataset $\tilde{\mathcal{D}}'$. Then $O(U(\tilde{\mathcal{D}})) - O(U(\tilde{\mathcal{D}}')) > \sigma\varepsilon$ will hold.

2) Suppose an data point $(\boldsymbol{x}, \tilde{S})$ in the dataset $\tilde{\mathcal{D}}$, which satisfies $j \in \tilde{S}$ and $j \neq y_{\boldsymbol{x}}$, is replaced by another point $(\boldsymbol{x}, \tilde{S}')$, which satisfies $\tilde{S}' = \tilde{S} \backslash \{j\}$, to form a new dataset $\tilde{\mathcal{D}}'$. Then $\varepsilon \leq O(U(\tilde{\mathcal{D}})) - O(U(\tilde{\mathcal{D}}')) \leq \sigma\varepsilon$ will holds.

Let the dataset changed by the refinement be $\tilde{\mathcal{D}}^+_{|I(f,\zeta)|}$, where $\tilde{\mathcal{D}}^+_k$ denotes $\tilde{\mathcal{D}}$ has added k candidate labels. Then according to the characteristic (1), the difference of the score between the original dataset $\tilde{\mathcal{D}}$ and the refined dataset $\tilde{\mathcal{D}}^+_{|I(f,\zeta)|}$ can be

formulated as:

$$\begin{aligned} O(U(\tilde{\mathcal{D}})) - O(U(\tilde{\mathcal{D}}^+_{|I(f,\zeta)|})) &= [O(U(\tilde{\mathcal{D}})) - O(U(\tilde{\mathcal{D}}^+_1))] + [O(U(\tilde{\mathcal{D}}^+_1)) - O(U(\tilde{\mathcal{D}}^+_2))] + \cdots \\ &\quad + [O(U(\tilde{\mathcal{D}}^+_{|I(f,\zeta)|-1})) - O(U(\tilde{\mathcal{D}}^+_{|I(f,\zeta)|}))] \\ &\geq \sigma\varepsilon - \sigma(|I(f,\zeta)| - 1)\varepsilon \\ &= (2\sigma - \sigma|I(f,\zeta)|)\varepsilon. \end{aligned} \tag{9}$$

Similarly, let the dataset changed by the disambiguation be $\tilde{\mathcal{D}}^-_{|\bar{I}(f,\bar{\zeta})|}$, where $\tilde{\mathcal{D}}^-_k$ denotes $\mathcal{D}^+{}_{|I(f,\zeta)|}$ has removed k candidate labels. Then the difference of the score between the refined dataset $\tilde{\mathcal{D}}^+_{|I(f,\zeta)|}$ and the disambiguated dataset $\tilde{\mathcal{D}}^-_{|\bar{I}(f,\bar{\zeta})|}$ can be formulated as:

$$\begin{aligned} O(U(\tilde{\mathcal{D}}^+_{|I(f,\zeta)|})) - O(U(\tilde{\mathcal{D}}^-_{|\bar{I}(f,\bar{\zeta})|})) &= [O(U(\tilde{\mathcal{D}}^+_{|I(f,\zeta)|})) - O(U(\tilde{\mathcal{D}}^-_1))] + [O(U(\tilde{\mathcal{D}}^-_1)) - O(U(\tilde{\mathcal{D}}^-_2))] + \cdots \\ &\quad + [O(U(\tilde{\mathcal{D}}^-_{|\bar{I}(f,\bar{\zeta})|-1})) - O(U(\tilde{\mathcal{D}}^-_{|\bar{I}(f,\bar{\zeta})|}))] \\ &\geq |\bar{I}(f,\bar{\zeta})|\varepsilon. \end{aligned} \tag{10}$$

Hence, after one fusion round of refinement and disambiguation, the boundary difference of the gap $|f_j(\boldsymbol{x}) - \eta_j(\boldsymbol{x})|$ between the original dataset $\tilde{\mathcal{D}}$ and the refined dataset can be formulated as:

$$\begin{aligned} (\alpha O(U(\tilde{\mathcal{D}}) + \epsilon) - (\alpha O(U(\tilde{\mathcal{D}}^-_{|\bar{I}(f,\bar{\zeta})|})) + \epsilon) &= \alpha(O(U(\tilde{\mathcal{D}}) - O(U(\tilde{\mathcal{D}}^-_{|\bar{I}(f,\bar{\zeta})|}))) \\ &= \alpha(O(U(\tilde{\mathcal{D}})) - O(U(\tilde{\mathcal{D}}^+_{|I(f,\zeta)|})) + O(U(\tilde{\mathcal{D}}^+_{|I(f,\zeta)|})) - O(U(\tilde{\mathcal{D}}^-_{|\bar{I}(f,\bar{\zeta})|}))) \\ &\geq \alpha(2\sigma - \sigma|I(f,\zeta)| + |\bar{I}(f,\bar{\zeta})|)\varepsilon \\ &\geq 2\alpha\sigma\varepsilon \\ &\geq \alpha\varepsilon \end{aligned} \tag{11}$$

The proof has been completed.

A.4. Proof of Corollary 1

Since the disambiguation procedure removes incorrect labels from candidate labels according to $\bar{I}(f,\bar{\zeta})$ with the disambiguation threshold $\bar{\zeta} = 2(\alpha O(U(\tilde{\mathcal{D}})) + \epsilon)$, we have the following constraint:

$$\min_{(\boldsymbol{x},j)\in\bar{I}(f,\bar{\zeta})} f_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) - f_j(\boldsymbol{x}) \geq 2(\alpha O(U(\tilde{\mathcal{D}}) + \epsilon)). \tag{12}$$

To eliminate the mistaken incorrect label j from candidate labels, the incorrect label j for the instance $\boldsymbol{x}$ need to satisfy the following constraint:

$$f'_{y_{f'(\boldsymbol{x})}}(\boldsymbol{x}) - f'_j(\boldsymbol{x}) \geq 2(\alpha O(U(\tilde{\mathcal{D}}')) + \epsilon), \tag{13}$$

where $\tilde{\mathcal{D}}'$ and f' are the new dataset and classifier respectively.

According to Theorem 3, after R fusion rounds, we have

$$\alpha O(U(\tilde{\mathcal{D}}')) + \epsilon \leq \alpha O(U(\tilde{\mathcal{D}})) + \epsilon - R\alpha\varepsilon \tag{14}$$

Hence, if $f'_{y_{f'(\boldsymbol{x})}}(\boldsymbol{x}) - f'_j(\boldsymbol{x}) \geq 2(\alpha O(U(\tilde{\mathcal{D}})) + \epsilon - R\alpha\varepsilon)$, the incorrect label j will be disambiguated.

According to Assumption 1, we have

$$\begin{aligned} &f'_j(\boldsymbol{x}) - f'_{y_{f'(\boldsymbol{x})}}(\boldsymbol{x}) \leq (\eta_j(\boldsymbol{x}) + (\alpha O(U(\tilde{\mathcal{D}}')) + \epsilon)) - (\eta_{y_{f'(\boldsymbol{x})}}(\boldsymbol{x}) - (\alpha O(U(\tilde{\mathcal{D}}')) + \epsilon)) \\ &= \eta_j(\boldsymbol{x}) - \eta_{y_{f'(\boldsymbol{x})}}(\boldsymbol{x}) + 2(\alpha O(U(\tilde{\mathcal{D}}')) + \epsilon) \\ &\leq \eta_j(\boldsymbol{x}) - \eta_{y_{f'(\boldsymbol{x})}}(\boldsymbol{x}) + 2(\alpha O(U(\tilde{\mathcal{D}})) + \epsilon - R\alpha\varepsilon) \\ &= \eta_j(\boldsymbol{x}) - \eta_{y_{f'(\boldsymbol{x})}}(\boldsymbol{x}) + 2(\alpha O(U(\tilde{\mathcal{D}})) + \epsilon) - 2R\alpha\varepsilon \\ &\leq \eta_j(\boldsymbol{x}) - \min_j \eta_j(\boldsymbol{x}) + 2(\alpha O(U(\tilde{\mathcal{D}})) + \epsilon) - 2R\alpha\varepsilon \end{aligned} \tag{15}$$

Hence, when

$$-(\eta_j(\boldsymbol{x}) - \min_j \eta_j(\boldsymbol{x}) + 2(\alpha O(U(\tilde{\mathcal{D}})) + \epsilon) - 2R\alpha\varepsilon) \geq 2(\alpha O(U(\tilde{\mathcal{D}})) + \epsilon - R\alpha\varepsilon),$$

Eq.(13) will be satisfied, i.e.,

$$R \geq \frac{1}{4\alpha\varepsilon}(2(\min_{(\boldsymbol{x},j)\in\bar{I}(f,\bar{\zeta})} f_{y_{f(\boldsymbol{x})}}(\boldsymbol{x}) - f_j(\boldsymbol{x})) + (\eta_j(\boldsymbol{x}) - \min_j \eta_j(\boldsymbol{x}))), \tag{16}$$

the mistaken label j will be removed from candidate label. The proof has finished.

A.5. Proof of Theorem 4

Under Assumption 2, for (α, ϵ, ρ)-ambiguity bounded $\mathbb{P}[\boldsymbol{x}, \tilde{S}]$, after running R rounds of refinement of disambiguation in the algorithm FREDIS, we have

$$\begin{aligned}
\mathbb{P}\left[y_{f_{\text{final}}(\boldsymbol{x})} = \eta^\star(\boldsymbol{x})\right] &= \mathbb{P}\left[y_{f_{\text{final}}(\boldsymbol{x})} = \eta^\star(\boldsymbol{x}) | (\boldsymbol{x}, S) \in G\right] \mathbb{P}\left[(\boldsymbol{x}, S) \in G\right] + \mathbb{P}\left[y_{f_{\text{final}}(\boldsymbol{x})} = \eta^\star(\boldsymbol{x}) | (\boldsymbol{x}, S) \notin G\right] \mathbb{P}\left[(\boldsymbol{x}, S) \notin G\right] \\
&= \mathbb{P}\left[y_{f_{\text{final}}(\boldsymbol{x})} = \eta^\star(\boldsymbol{x}) | (\boldsymbol{x}, S) \in G\right] (1-\rho) + \mathbb{P}\left[y_{f_{\text{final}}(\boldsymbol{x})} = \eta^\star(\boldsymbol{x}) | (\boldsymbol{x}, S) \notin G\right] \rho \\
&\geq \mathbb{P}\left[y_{f_{\text{final}}(\boldsymbol{x})} = \eta^\star(\boldsymbol{x}) | (\boldsymbol{x}, S) \in G\right] (1-\rho) \\
&\geq \mathbb{P}\left[\eta_{y_{\boldsymbol{x}}}(\boldsymbol{x}) - \eta_{s_{\boldsymbol{x}}}(\boldsymbol{x}) \geq 2(\alpha O(U(\tilde{\mathcal{D}})) + \epsilon - \alpha\varepsilon R) | (\boldsymbol{x}, S) \in G\right] (1-\rho) \\
&= (1 - \mathbb{P}\left[\eta_{y_{\boldsymbol{x}}}(\boldsymbol{x}) - \eta_{s_{\boldsymbol{x}}}(\boldsymbol{x}) \leq 2(\alpha O(U(\tilde{\mathcal{D}})) + \epsilon - \alpha\varepsilon R) | (\boldsymbol{x}, S) \in G\right])(1-\rho) \\
&\geq (1 - C(2(\alpha O(U(\tilde{\mathcal{D}})) + \epsilon - \alpha\varepsilon R))^\lambda)(1-\rho).
\end{aligned} \tag{17}$$

A.6. Details for experiments

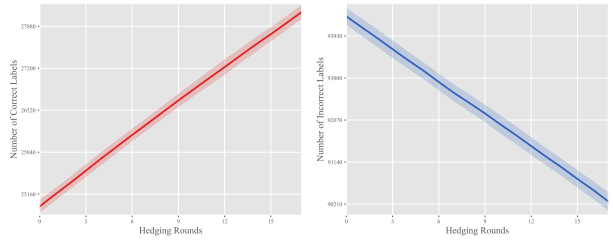
Table 3 and 4 summarize the classification accuracy of each comparison approach on manually corrupted UPLL benchmark datasets with the rest pairs of γ_1 and γ_2. Figure 2 illustrates the change of the number about correct labels and incorrect labels on benchmark datasets with various pairs of γ_1 and γ_2.

Table 3. Test accuracy (%) on benchmark datasets `Fashion-MNIST` and `Kuzushiji-MNIST` under different levels of γ_1 and γ_2 when FREDIS is compared with six PLL methods. Average accuracy and standard deviation over 5 trials are reported.

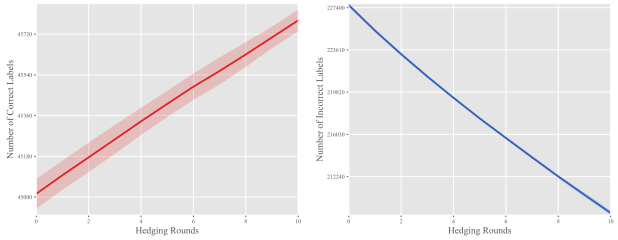
FMNIST	$\gamma_1 = 0.1$ $\gamma_2 = 0.1$	$\gamma_1 = 0.1$ $\gamma_2 = 0.3$	$\gamma_1 = 0.3$ $\gamma_2 = 0.1$
FREDIS	$\mathbf{93.10 \pm 0.10}$	$\mathbf{92.59 \pm 0.07}$	$\mathbf{91.41 \pm 0.11}$
PLCR	92.48 ± 0.19•	92.03 ± 0.12•	91.06 ± 0.06•
PICO	92.11 ± 0.22•	91.45 ± 0.15•	90.60 ± 0.27•
CAVL	88.04 ± 0.28•	80.91 ± 3.79•	82.19 ± 5.35•
LWS	91.13 ± 0.10•	90.70 ± 0.14•	89.70 ± 0.24•
PRODEN	88.01 ± 0.24•	87.16 ± 0.16•	86.68 ± 0.20•
RC	88.34 ± 0.33•	87.27 ± 0.11•	86.59 ± 0.35•
CC	88.51 ± 0.29•	87.92 ± 0.38•	86.86 ± 0.17•
KMNIST	$\gamma_1 = 0.1$ $\gamma_2 = 0.1$	$\gamma_1 = 0.1$ $\gamma_2 = 0.3$	$\gamma_1 = 0.3$ $\gamma_2 = 0.1$
FREDIS	$\mathbf{96.68 \pm 0.12}$	$\mathbf{96.22 \pm 0.09}$	94.55 ± 0.12
PLCR	96.17 ± 0.24•	95.67 ± 0.12•	94.21 ± 0.36•
PICO	96.38 ± 0.16•	95.70 ± 0.17•	$\mathbf{94.73 \pm 0.36}$
CAVL	93.76 ± 0.11•	93.03 ± 0.40•	91.70 ± 0.25•
LWS	93.96 ± 0.71•	92.22 ± 0.31•	91.01 ± 0.78•
PRODEN	95.81 ± 0.15•	94.91 ± 0.29•	93.97 ± 0.61•
RC	95.51 ± 0.46•	95.17 ± 0.46•	93.70 ± 0.28•
CC	95.66 ± 0.49•	93.62 ± 2.17•	93.29 ± 0.20•

Table 4. Test accuracy (%) on benchmark datasets `CIFAR-10` and `CIFAR-100` under different levels of γ_1 and γ_2 when FREDIS is compared with six PLL methods. Average accuracy and standard deviation over 5 trials are reported.

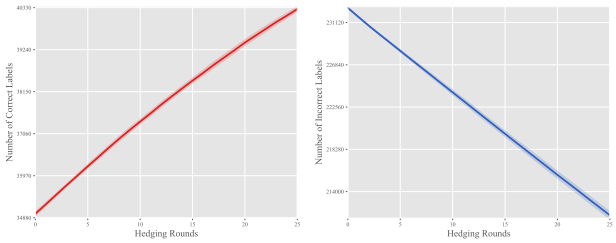
CIFAR10	$\gamma_1 = 0.1$ $\gamma_2 = 0.1$	$\gamma_1 = 0.1$ $\gamma_2 = 0.3$	$\gamma_1 = 0.3$ $\gamma_2 = 0.1$
FREDIS	$\mathbf{90.57 \pm 0.23}$	$\mathbf{89.02 \pm 0.15}$	$\mathbf{84.35 \pm 0.20}$
PLCR	89.38 ± 0.41•	88.17 ± 0.17•	82.76 ± 0.47•
PICO	89.52 ± 0.24•	88.49 ± 0.31•	83.86 ± 0.34•
CAVL	81.43 ± 3.85•	58.45 ± 9.29•	68.56 ± 5.94•
LWS	17.48 ± 0.72•	67.10 ± 0.88•	16.75 ± 0.57•
PRODEN	81.86 ± 0.52•	80.11 ± 0.98•	77.29 ± 1.11•
RC	83.44 ± 0.88•	82.00 ± 0.36•	78.54 ± 0.58•
CC	81.60 ± 0.93•	78.80 ± 0.85•	74.56 ± 1.19•
CIFAR100	$\gamma_1 = 0.1$ $\gamma_2 = 0.01$	$\gamma_1 = 0.1$ $\gamma_2 = 0.05$	$\gamma_1 = 0.3$ $\gamma_2 = 0.01$
FREDIS	$\mathbf{64.73 \pm 0.28}$	$\mathbf{66.43 \pm 0.22}$	$\mathbf{59.42 \pm 0.18}$
PLCR	62.79 ± 0.24•	61.07 ± 0.58•	57.11 ± 0.38•
PICO	64.05 ± 1.53	64.41 ± 0.66•	55.64 ± 0.54•
CAVL	44.73 ± 2.54•	36.01 ± 1.15•	38.19 ± 2.79•
LWS	5.09 ± 0.56•	5.29 ± 0.56•	5.27 ± 1.09•
PRODEN	58.91 ± 0.76•	55.96 ± 0.18•	54.09 ± 0.77•
RC	59.66 ± 0.68•	57.98 ± 0.51•	55.71 ± 0.56•
CC	58.41 ± 0.76•	56.41 ± 0.45•	53.79 ± 0.40•



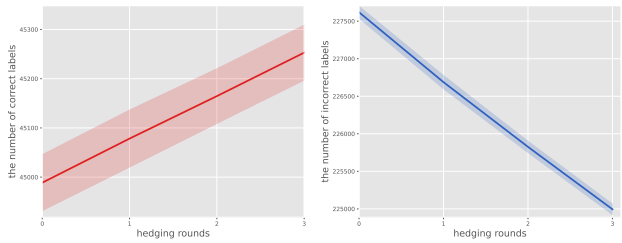
(a) Fashion-MNIST, $\gamma_1 = 0.5$ and $\gamma_2 = 0.1$



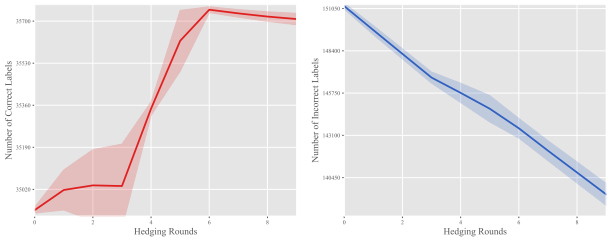
(b) Fashion-MNIST, $\gamma_1 = 0.1$ and $\gamma_2 = 0.5$



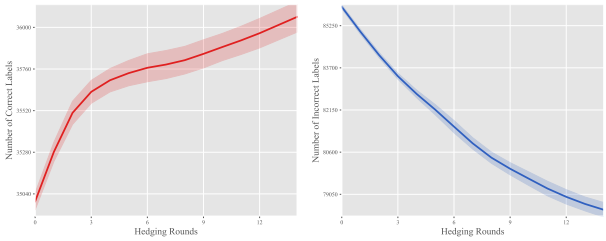
(c) Kuzushiji-MNIST, $\gamma_1 = 0.3$ and $\gamma_2 = 0.5$



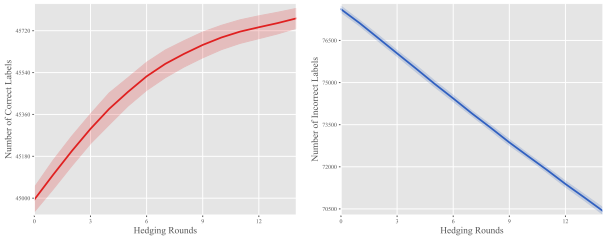
(d) Kuzushiji-MNIST, $\gamma_1 = 0.5$ and $\gamma_2 = 0.3$



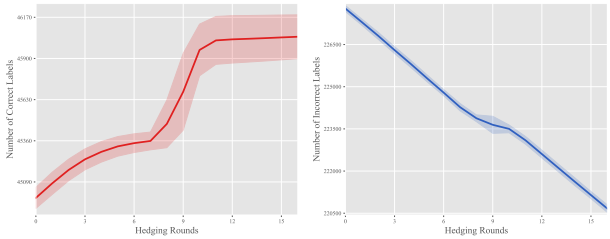
(e) CIFAR-10, $\gamma_1 = 0.3$ and $\gamma_2 = 0.3$



(f) CIFAR-10, $\gamma_1 = 0.3$ and $\gamma_2 = 0.1$



(g) CIFAR-100, $\gamma_1 = 0.1$ and $\gamma_2 = 0.01$



(h) CIFAR-100, $\gamma_1 = 0.1$ and $\gamma_2 = 0.05$

Figure 2. The number of correct labels (Left) and incorrect labels (Right) on benchmark datasets with various pairs of γ_1 and γ_2