ADVERSARY-AWARE PARTIAL LABEL LEARNING WITH LABEL DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

To ensure that the data collected from human subjects is entrusted with a secret, rival labels are introduced to conceal the information provided by the participants on purpose. The corresponding learning task can be formulated as a noisy partial-label learning problem. However, conventional partial-label learning (PLL) methods are still vulnerable to the high ratio of noisy partial labels, especially in a large labelling space. To learn a more robust model, we present Adversary-Aware Partial Label Learning and introduce the *rival*, a set of noisy labels, to the collection of candidate labels for each instance. By introducing the rival label, the predictive distribution of PLL is factorised such that a reasonably good predictive label is achieved with less uncertainty coming from the transition matrix, assuming its generation process is known. Nonetheless, the predictive accuracy is still insufficient to produce an adequately good set of positive samples to minimise the loss function. Moreover, the inclusion of rivals also brings an inconsistency issue for the classifier and risk function due to the intractability of the transition matrix. Consequently, the immature teacher within momentum (ITWM) disambiguation algorithm is proposed to cope with the situation. We utilise the confidence score mapping from the instance space to approximate the intractable term, allowing us to obtain a provably consistent classifier and risk function. Extensive experiments demonstrate that our method achieves promising results on the CIFAR10, CIFAR100 and CUB200 datasets.

1 INTRODUCTION

Deep learning algorithms depend heavily on a large-scale, true annotated training dataset. Nonetheless, the costs of accurately annotating a large volume of true labels to the instances are exorbitant, not to mention the time invested in the labelling procedures. As a result, weakly supervised labels such as partial labels that substitute true labels for learning have proliferated and gained massive popularity in recent years. Partial-label learning (PLL) is a special weakly-supervised learning problem associated with a set of candidate labels $\vec{Y}$ for each instance, in which only one true latent label y is in existence. Nonetheless, without an appropriately designed learning algorithm, the limitations of the partial label are evident since deep neural networks are still vulnerable to the ambiguous issue rooted in the partial label problem because of noisy labels Zhou (2018); Patrini et al. (2017); Han et al. (2018). As a result, there have had many partial label learning works (PLL)Cour et al. (2011); Hüllermeier & Beringer (2006); Feng & An (2019); Feng et al. (2020) successfully solved the ambiguity problem where there is a set of candidate labels for each instance, and only a true label exists. Apart from the general partial label, we have also seen a variety of partial label generations evolved, simulating different real-life scenarios. The independently and uniformly drawing is the one have seen the most Lv et al. (2020); Feng & An (2019). The other problem settings include the instance dependent partial label learning, where each partial label set is generated depending on the instance as well as the true label Xu et al. (2021). Furthermore, Lv et al. (2020) has introduced label specific partial label learning, where the uniform flipping probability of similar instances differs from dissimilar group instances. Overall, the learning objective of the previous works is all about disambiguation. More specifically, the goal is to design a classifier training with partial labels, aiming to correctly label the testing dataset, hoping the classification performance will be as close as the full supervised learning.

On the contrary, there is a lack of discussion on previous works that shed light on the data privacy-enhancing techniques in general partial label learning. The privacy risk is inescapable; thus, privacy-preserving techniques need to be urgently addressed. Recently, we have seen surging data breach cases worldwide. These potential risks posed by the attacker are often overlooked and pose a detrimental threat to society. For instance, it is most likely for the adversary to learn from stolen or leaked partially labelled data for illegal conduct using the previous proposed partial-label learning methods. Subsequently, it has become an inherent privacy concerns in conventional partial label learning. In this paper, the Adversary-Aware partial label learning is proposed to address and mitigate the ramification of the data breach. In a nutshell, we propose an affordable and practical approach to manually corrupt the collected dataset to prevent the adversary from obtaining high-quality, confidential information meanwhile ensure the trustee has full access to the useful information. However, we have observed that adversary-aware partial label learning possesses some intrinsic learnability issues. Firstly, the intractability is raised from the transition matrix. Secondly, the classifier and risk inconsistency problem has been raised. Hence, we propose an immature teacher within momentum (ITWM), adversary-aware loss function, and a new ambiguity condition to counter the issues. In the previous works of partial label generation procedure that there is, only a candidate of the partial label was generated as such.

The Previous Works Label Generation:

$$\begin{aligned}&\sum_{y\in Y}\mathrm{P}(\vec{Y}=\vec{y},Y=y\mid X=x)\\&=\sum_{y\in Y}\mathrm{P}(\vec{Y}=\vec{y}\mid Y=y,X=x)\mathrm{P}(Y=y\mid X=x).\end{aligned}\tag{1}$$

Then we present the difference between the general partial labels and the adversary-aware partial label.

The Adversary-Aware Partial Label Generation:

$$\begin{aligned}&\sum_{y\in Y}\sum_{y'\in Y'}\underbrace{\mathrm{P}(\vec{Y}=\vec{y},Y=y,Y'=y'\mid X=x)}_{\textbf{Adversary Aware Partial Label(Includes the Rival)}}\\&=\sum_{y\in Y}\sum_{y'\in Y'}\underbrace{\mathrm{P}(\vec{Y}=\vec{y}\mid Y=y,X=x,Y'=y')}_{\textbf{Intractable transition matrix}}T_{y,y'}\mathrm{P}(Y=y\mid X=x).\end{aligned}\tag{2}$$

Under the adversary-aware partial label problem setting, the rival is added to a candidate set of labels. In order to achieve that, we extend Eq.1 by adding the rival, Y' by factorisation. Then, we decompose the second equation into the rival embedded intractable transition matrix term $\bar{Q}$ and class instance-dependent transition matrix $T_{y,y'}$, which is $\mathrm{P}(Y'=y'\mid Y=y,X=x)$. In our problem setting, $\bar{T}_{y,y'}$, the class instance-independent transition matrix is utilised, which is defined as $\mathrm{P}(Y'=y'\mid Y=y)$, with the assumption the rival is generated depending only on Y but instance X. However, a fundamental problem has been raised, inclusion of the rival implies an inconsistent classifier according to the adversary-aware label generation equation Eq.2. Learning a consistent partial label classifier is vital, but in our problem setting, the consistency classifier may not be obtained due to the intractability of $\bar{Q}$. As a consequence, the immature teacher within momentum (ITWM) is proposed, which is designed to approximate the term $\bar{Q}=\mathrm{P}(\vec{Y}=\vec{y}\mid Y=y,X=x)$. The Moco-style dictionary technique He et al. (2020) and Wang et al. (2022) have inspired us to explore exploiting the the soft label from instance embedding to approximate the $\bar{Q}$ due to the property of informational preservation and tractability. Therefore, a consistent partial label learner is obtained if the uncertainty raised from the transition matrix is reduced greatly. Subsequently, it enables us to re-parameterize the neural network parameters with the adversary-aware matrix. Moreover, there are many labelling strategies. Based on the peculiar property of our problem setting, the immature teacher within momentum(ITWM) method is developed to approximate the $\bar{Q}$. Specifically, we transform the inference of label generation in Adversary-Aware PLL as an approximation for the transition matrix $\bar{Q}$. Ultimately, a tractable solution to the unbiased estimate of $\mathrm{P}(\vec{Y}\mid Y,X)$ can be derived. Lastly, we have rigorously proven that a consistent Adversary-Aware PLL classifier can be obtained if $\mathrm{P}(\vec{Y}\mid Y,Y',X)$ and $\mathrm{P}(Y'\mid Y)$ are approximated accurately. Overall, our proposed method has not only solved the ambiguity problem in Adversary-Aware PLL but also addressed the

potential risks from the data breach by using a rival as the encryption. The **main contributions** of the work are summarized:

- We propose a novel problem setting named adversary-aware partial label learning.
- We propose a novel Adversary-Aware loss function and the immature teacher within momentum (ITWM) disambiguation algorithm.
- A new ambiguity condition Eq.equation 24 for Adversary-Aware Partial Label Learning is derived.
- Theoretically, we proven that the method is a Classifier-Consistent Risk Estimator.

1.1 Related work

Partial Label Learning (PLL) trains an instance associated with a candidate set of labels, in which the true label is included. Many frameworks are designed and proposed to solve the label ambiguation issue in partial label learning. The probabilistic graphical model-based methodsZhang et al. (2016); Wang & Isola (2020); Xu et al. (2019); Lyu et al. (2019) as well as the clustering-based or unsupervised approaches Liu & Dietterich (2012) are proposed by leveraging the graph structure and prior information of feature space to do the label disambiguation. The average-based perspective methods Hüllermeier & Beringer (2006); Cour et al. (2011); Zhang et al. (2016) are designed based on the assumption of uniform treatment of all candidates; however, it is vulnerable to the false positive label, leading to misled prediction. Identification perspective-based methods Jin & Ghahramani (2002) tackle disambiguation by treating the true label as a latent variable. The representative perspective approach uses the maximum margin method Nguyen & Caruana (2008); Wang et al. (2020; 2022) to do the label disambiguation. Most recently, self-training perspective methodsFeng & An (2019); Wen et al. (2021); Feng et al. (2020) have emerged and shown the promising performance.
Contrastive Learning He et al. (2020); Oord et al. (2018) is a self-supervised learning method that uses only augmented input for different learning tasks. The objective is to differentiate the similar and dissimilar parts of the input, maximising the learning of the high-level representations. CL has been studied in unsupervised representation fashionChen et al. (2020); He et al. (2020); Oord et al. (2018), which treats the same classes the positive set to boost the performance. The weakly supervised learning has also borrowed the concepts of CL to tackle the partial label problemWang et al. (2022). The CL has also been applied to semi-supervised learning Li et al. (2020).

1.2 Problem Formulation

In the Adversary-Aware Partial Label case. Given the input space $\mathcal{X} \in \mathbb{R}^d$ and label space is defined as $\mathcal{Y}$ = $[c]$ $\{1 \cdots c\}$ with the number of $c > 2$ classes. Under adversary-aware partial labels, each instance $X \in \mathcal{X}$ has a candidate set of adversary-aware partial labels $\vec{Y} \in \vec{\mathcal{Y}}$. The adversary-aware partial label set has space of $\vec{\mathcal{Y}} := \{\vec{y} \mid \vec{y} \subset \mathcal{Y}\}$=$2^{[c]}$, in which there is total $2^{[c]}$ selection of subsets in $[c]$. The objective is to learn a classifier with the adversary-aware partially labeled sample n which was i.i.d drawn from the $\vec{\mathcal{D}} = \{(X_1, \vec{Y}_1), \ldots, (X_n, \vec{Y}_n)\}$, aiming that it is able to assign the true labels for the testing dataset. Given instance and the adversary-aware partial label $\vec{Y}$ the adversary-aware partial label dataset distribution $\vec{D}$ is defined as $(X, \vec{Y}) \in \mathcal{X} \times \vec{\mathcal{Y}}$. Formally, we assume that each instance X has adversary-aware partial label set $\vec{Y}$, i.e., We assume that the partial label set $\vec{Y}$ is guaranteed to include the true label Wen et al. (2021), for each instance that the partial label set is defined as

$$P(y \in \vec{Y} \mid Y = y, x) = 1. \tag{3}$$

Under the adversary-aware partial label, the rival label is generated or flipped depending only on true label Y of each instance, for which the relationship can be defined as $T_{y,y'} = \mathrm{P}(Y' = y' \mid Y = y)$, $\bar{T} \in \mathbb{R}^{c \times c}$ and $\bar{T}_{y,y} = 0$, for $\forall_{y,y'} \in [c]$. The $\epsilon_{y'|y,x}$ is the instance dependent label noise for each instance, where $\epsilon \in R^{1 \times c}$. $\bar{Q}_{ij} = p(\vec{Y} = A_j \mid y = i)$ in which $Aj \in \mathcal{A}(j \in [2^c - 2])$. The entries of the transition matrix for each instance is defined as follow

$$p(\vec{Y} \mid Y', Y = y, x) = \begin{cases} \bar{Q}[:, j]^T T_{y,y'} + \epsilon_{y'|y,x} & \text{if } y \in \vec{Y} \\ 0 & \text{other } y \notin \vec{Y}. \end{cases} \tag{4}$$

The formal derivation is shown as

$$
\begin{aligned}
P(\vec{Y}=\vec{y} \mid X=x) &= \sum_{y \in Y} \sum_{y' \in Y'} \mathrm{P}(\vec{Y}=\vec{y}, Y=y, Y'=y' \mid X=x) \\
&= \sum_{y \in Y} \sum_{y' \in Y'} \mathrm{P}(\vec{Y}=\vec{y} \mid Y=y, Y'=y', X=x) \bar{T}_{y,y'} \mathrm{P}(Y=y \mid X=x),
\end{aligned}
\tag{5}
$$

where the conditional distribution of the adversary-aware partial label set $\vec{Y}$ Wen et al. (2021) is denoted as

$$
\mathrm{P}(\vec{Y}=\vec{y} \mid Y=y, Y'=y', X=x) = \prod_{b' \in \vec{y}, b' \neq y} p_{b'} \cdot \prod_{t' \notin \vec{y}} \left(1-p_{t'}\right), \tag{6}
$$

where $p_{t'}$ and $p_{b'}$ are defined as

$$
p_{t'} := \mathrm{P}(t \in \vec{Y} \mid Y=y, Y'=y', X=x) < 1, p_{b'} := \mathrm{P}(b \in \vec{Y} \mid Y=y, Y'=y', X=x) < 1. \tag{7}
$$

We define $\mathrm{P}(\vec{Y} \mid Y, Y', X)$ as $\bar{\boldsymbol{Q}}$, where $\vec{y} \in \vec{Y}\,(\vec{y} \in [2^c])$ is a candidate label set and $\bar{Q} \in \mathbb{R}^{c \times (2^c-2)}$. We summarize the Eq. (5) as a matrix form in Eq.(8). The inverse problem is to learn a sparse approximation matrix $\boldsymbol{A}$ to use Eq.(9) approximate the true posterior distribution.

$$
\underbrace{P(\vec{Y} \mid X=x)}_{\textbf{Adversary-aware PLL}} = (\bar{\boldsymbol{Q}}[:, \boldsymbol{j}]^{\mathrm{T}} \bar{\boldsymbol{T}} + \boldsymbol{\epsilon}) \underbrace{P(Y \mid X=x)}_{\textbf{True posterior probability}}, \tag{8}
$$

$$
\bar{\boldsymbol{T}}^{-1} \boldsymbol{A} \underbrace{P(\vec{Y} \mid X=x)}_{\textbf{Adversary-aware PLL}} \approx \underbrace{P(Y \mid X=x)}_{\textbf{True posterior probability}}. \tag{9}
$$

In reality, due to the computational complexity of the transition matrix, it would be a huge burden to estimate $\bar{Q}$ accurately for each instance. The $2^c - 2$ is an extremely large figure and increases exponentially as the label space increase. Therefore, we no longer wish to obtain the true transition matrix $\mathrm{P}(\vec{Y} \mid Y, Y', X)$ but resort to use instance embedding in an form of an soft label to approximate the adversary-aware partial label transition matrix $\bar{Q}$. Specifically, we proposed to use soft pseudo label with constraints of $[0, 1]$ from the instance embedding (Prototype) to approximate the adversary-aware transition matrix for each instance. This is achieved using the multiplication of the lower embedding of query u and queue v. More specifically, we have used the pseudo label from ITWM as the approximation, in which the confidence is approximated as $\alpha = \mathrm{softmax}_{j \in Y} \boldsymbol{u_i}^{\top} \boldsymbol{v}_j$. The α is transformed as single-scalar parameter with regularization of $\alpha \in [0, 1]$. Since it is the approximation term, thus inevitably the approximation error will be high comparing to using the true transition matrix $Q_{ij}(x)$. Nonetheless, the estimation error is reduced greatly, leading to better classification performance Zhang & Sugiyama (2021). Since the Adversary-aware partial label is influenced by the rival label noise, it is challenging to accurately estimate both the transition matrix $\bar{\boldsymbol{T}}$ and the sparse matrix $\boldsymbol{A}$ simultaneously to estimate the true posterior. Considering that the $\bar{\boldsymbol{T}}$ is private, it is easier for us just to approximate $\boldsymbol{A}$ to estimate the posterior than the adversary.

1.3 Positive Sample Set

We have followed Wang et al. (2022) for the positive sample selection. For illustrative purpose, the generation of a positive sample set is demonstrated as follow. In each mini-batch, we have defined the it as $\vec{D}_b \in \vec{D}$. The $f(x_i)$ is the function of a neural network with a projection head of 128 feature dimensionality. The outputs of D_q and D_k are defined accordingly,

$$
D_q = \{\boldsymbol{u}_i = (\bar{\mathbf{T}} + \mathbf{I}) f\left(\mathrm{Aug}_q\left(\boldsymbol{x}_i\right)\right) \mid \boldsymbol{x}_i \in \vec{D}_b\}, \tag{10}
$$

$$
D_k = \{\boldsymbol{v}_i = (\bar{\mathbf{T}} + \mathbf{I}) f'\left(\mathrm{Aug}_k\left(\boldsymbol{x}_i\right)\right) \mid \boldsymbol{x}_i \in \vec{D}_b\}, \tag{11}
$$

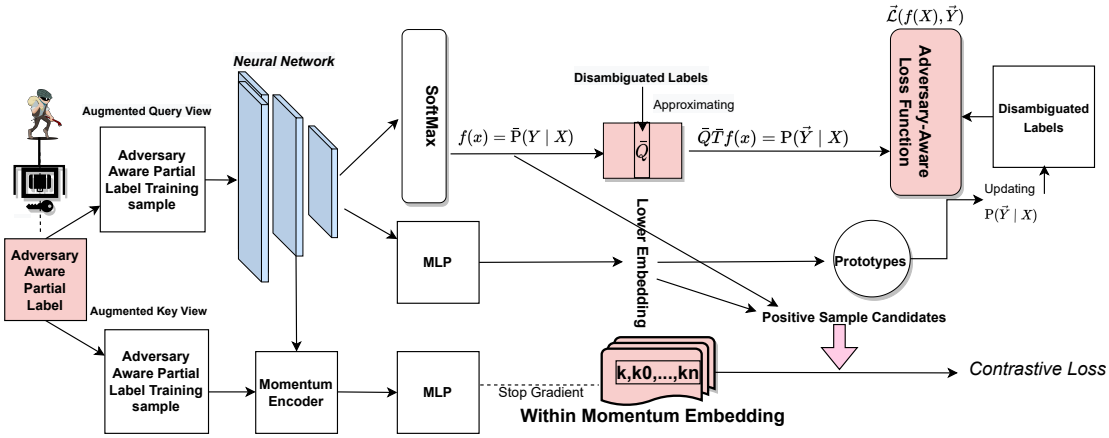


Figure 1: An overview of the proposed method. General partial label can be disclosed to adversary. The initial training is about positive sample selection. Moreover, we have assumed $\bar{T}$ is given. Thus we have applied it directly in the adversary-aware loss function to generate good positive samples and query embedding u. Thereafter, the immature teacher within momentum (ITWM) is used to generate the new pseudo label $\bar{q}$ by $\boldsymbol{u}^\top \boldsymbol{v}$. In the second stage of learning, the $\bar{Q}$ is approximated with immature teacher within momentum (ITWM). The uncertainty of the transition matrix is greatly reduced and $\mathrm{P}(Y = y \mid X = x)$ is obtained given a good approximation of $\bar{Q}$ and adversary-aware transition matrix.

where $\bar{S}(\boldsymbol{x})$ is the sample set excluding the query set q and is defined as $\bar{S}(\boldsymbol{x}) = \bar{C} \backslash \{\boldsymbol{q}\}$, in which $\bar{\mathcal{C}} = D_q \cup D_k \cup$ queue . Positive set selection plays an instrumental role in our method since we have used it to approximate the transition matrix $P(\vec{Y} \mid Y', Y, X)$. The instances from the current mini-batch with the prediction label $\bar{y}'$ equal to $(\hat{y}_i = c)$ from the $\bar{\mathcal{S}}(x)$. is chosen to be the positive sample set. Finally, the $N(\boldsymbol{x})$ is obtained, and we can define it as

$$N_+(\boldsymbol{x}_i) = \left\{ \boldsymbol{v}' \mid \boldsymbol{v}' \in \bar{\mathcal{S}}\left(\boldsymbol{x}_i\right), \bar{y}' = (\hat{y}_i = c) \right\}. \tag{12}$$

2 Methodology

2.1 The immature teacher within momentum (ITWM)

The key task of partial label learning is label disambiguation, which targets to identify the true label among candidate label set. One of the common approach is to search for the maximum margin by finding a classifier f which maximizes the predictive difference between $F(\boldsymbol{x}, y)$ and $\max_{y \neq y'} F(\boldsymbol{x}, y')$Koltchinskii & Panchenko (2002). The v denotes as the feature representation. Formally, the margin of a sample $(\boldsymbol{x}, y)$ is defined as

$$\gamma(\boldsymbol{x}, y) = F(\boldsymbol{x}, y) - \max_{y \neq y'} F(\boldsymbol{x}, y') = \boldsymbol{w}_y^\top \boldsymbol{v} - \max_{y \neq y'} \boldsymbol{w}_{y'}^\top \boldsymbol{v}.$$

Thus, we have presented the immature teacher within momentum (ITWM) to first maximise the inter-class separability using the the immature teacher soft labelling and then minimise the intra class separability using the within momentum embedding. Subsequently, the more discriminative features of the prototypes is learnt and the concentration of the feature is more attached with respect to each prototype. In a nutshell, the key to our momentum updating rule is to separate the margin between each prototype representation on hyper-sphere as large as possible. The margin between prototype vector $\boldsymbol{v}_i \in \mathbb{S}^{d-1}$ and prototype vector $\boldsymbol{v}_j \in \mathbb{S}^{d-1}$ is defined as

$$m_{ij} = \exp\left(-\boldsymbol{v}_i^\top \boldsymbol{v}_j\right). \tag{13}$$

For prototype $\boldsymbol{v}_i$, we define the normalized margin between $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ as

$$\bar{m}_{ij} = \frac{\exp\left(-\boldsymbol{v}_i^\top \boldsymbol{v}_j\right)}{\sum_{j \neq i} \exp\left(-\boldsymbol{v}_i^\top \boldsymbol{v}_j\right)}. \tag{14}$$

For each $\boldsymbol{v}_i, i \in \{1, \cdots, K\}$, we perform momentum updating with the normalized margin between $\boldsymbol{v}_j$ and $\boldsymbol{v}_i$ for all $j \neq i$ as an regularization. The resulted new update rule is given as

$$\boldsymbol{v}_i^{t+1} = \sqrt{1-\alpha^2}\boldsymbol{v}_i^t + \alpha \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}, \tag{15}$$

where the gradient $\boldsymbol{g}$ is given as

$$\boldsymbol{g} = \boldsymbol{u} - \beta \sum_{j \neq i} \bar{m}_{ij}^t \boldsymbol{v}_j^t, \tag{16}$$

where $\boldsymbol{u}$ is the query embedding whose prediction is class i, $\bar{m}_{ij}^t$ is the normalized margin between prototype vectors at step t (i.e., $\boldsymbol{v}_j^t, j \neq i.$). The $\alpha \in (0,1)$ is a positive number. Since the $\bar{Q}$ is intractable, we have used the immature teacher within momentum (ITWM) to infer the term. Therefore, in the application, we have not used it directly; instead, we have used the hard label generated from the (ITWM) as the true label.

$$\bar{\boldsymbol{q}} = \phi \bar{\boldsymbol{q}} + (1-\phi)\boldsymbol{v}, \quad v_c = \begin{cases} 1 & \text{if } c = \arg\max_{j \in Y} \boldsymbol{u}^\top \boldsymbol{v} \\ 0 & \text{otherwise} \end{cases}. \tag{17}$$

2.1.1 Adversary Aware Loss Function.

The goal is to build a risk consistent loss function, hoping it can achieve the same generalization error as the supervised classification risk $R(f)$ with the same classifier f. To train the classifier, we minimize the following modified loss function estimator by leveraging the updated pseudo label from the immature teacher distillation method and transition plus identity matrix, $I_{i,j} \in [0,1]^{c \times c}$, $I_{i,i} = 1$, for $\forall_{i=j} \in [c]$, $I_{i,j} = 0$, for $\forall_{i \neq j} \in [c]$: where $f(\boldsymbol{X}) \in \mathbb{R}^{|c|}$,

$$\vec{\mathcal{L}}(f(X), \vec{Y}) = -\sum_{i=1}^{c} (\bar{q}_i) \log\left(((\bar{\mathbf{T}} + \mathbf{I}) f(X))_i\right), \tag{18}$$

where $\bar{q}$ is the target prediction and it was initialised as the uniform probability $\bar{\boldsymbol{q}} = \frac{1}{|c|}1$ and updated accordingly to the Eq.equation 17. The ϕ is the hyper-parameter controlling for the updating of $\bar{\boldsymbol{q}}$. The proof for the modified loss function is shown in the appendix lemma 4. In our case, the new contrastive learning loss is utilised incorporate with E.q equation 18 to approximate the transition matrix of adversary-aware partial label learning and is defined as

$$\begin{aligned} &\mathcal{L}(f(x), \tau, C) \\ &= \frac{1}{|D_q|} \sum_{\boldsymbol{u} \in D_q} \{-\frac{1}{N_+(x)} \sum_{\boldsymbol{v}_+ \in N_+(x)} \log \frac{\exp(\boldsymbol{u}^\top \boldsymbol{v}/\tau)}{\sum_{\boldsymbol{v}' \in \bar{C}(\boldsymbol{x})} \exp(\boldsymbol{u}^\top \boldsymbol{v}/\tau)}\}. \end{aligned} \tag{19}$$

Finally, we have the Adversary-Aware Loss expressed as

$$\textbf{Adversary-Aware Loss} = \lambda \mathcal{L}(f(x_i), \tau, C) + \vec{\mathcal{L}}(f(X), \vec{Y}). \tag{20}$$

3 Theoretical Analysis

The section introduces the concepts of classifier consistency and risk consistency Xia et al. (2019)Zhang (2004), which are crucial in weakly supervised learning. Risk consistency is achieved if the risk function of weak supervised learning is the same as the risk of fully supervised learning with the same hypothesis. The risk consistency implies classifier consistency, meaning classifier trained with partial labels is consistent as the optimal classifier of the fully supervised learning.

Classifier-Consistent Risk Estimator Learning with True labels. Lets denote $f(X) = (g_1(x), \ldots, g_K(x))$ as the classifier, in which $g_c(x)$ is the classifier for label $c \in [K]$. The prediction of the classifier $f_c(x)$ is $P(Y = c \mid x)$. We want to obtain a classifier $f(X)$ =$\arg\max_{i \in [K]} g_i(x)$. The loss function is to measure the loss given classifier $f(X)$. To this end, the true risk can be denoted as

$$R(f) = \mathbb{E}_{(X,Y)}[\mathcal{L}(f(X), Y)]. \tag{21}$$

The ultimate goal is to learn the optimal classifier f^*=$\arg\min_{f\in\mathcal{F}} R(f)$ for all loss functions, for instance to enable the empirical risk $\bar{R}_{pn}(f)$ to be converged to true risk $R(h)$. To obtain the optimal classifier, we need to prove that the modified loss function is risk consistent as if it can converge to the true loss function.

Learning with adversary-aware Partial Label. An input $X \in \mathcal{X}$ has a candidate set of $\vec{Y} \in \vec{\mathcal{Y}}$ but a only true label $Y \in \vec{\mathcal{Y}}$. Given the adversary-aware partial label $\vec{Y} \in \vec{\mathcal{Y}}$ and instance $X \in \mathcal{X}$ that the objective of the loss function is denoted as

$$\hat{R}(f) = \mathbb{E}_{(X,\vec{Y})}\vec{\mathcal{L}}\left(f\left(X\right), \vec{Y}\right). \tag{22}$$

Since the true adversary-aware partial label distribution $\bar{\mathcal{D}}$ is unknown, our goal is approximate the optimal classifier with sample distribution $\bar{D}_{pn}$ by minimising the empirical risk function, namely

$$\hat{R}_{pn}(f) = \frac{1}{n}\sum_{i=1}^{n}\vec{\mathcal{L}}\left(f\left(\boldsymbol{x}_i\right), \vec{y}_i\right). \tag{23}$$

The following conditions are important for the classifier to be consistent under the adversary-aware partial label learning problem. Lemma1 is the new ERM learnability condition, Lemma2 is the convergence of optimal classifier, Theorem2 is the convergence of the loss function, and Theorem 3 is the generalisation error bound. According to Cour et al. (2011) there needs to be certain degrees of ambiguity for the partial label learning.

Lemma 1. We have proposed a new condition specifically for the newly proposed problem setting as

$$P_{b,l} := \mathrm{P}(b, l \in \vec{Y} \mid Y = y, Y' = y', X = x). \tag{24}$$

It has to be met to ensure the Adversary-Aware PLL problem is learnable with $b \neq y$, l=y' and $l \neq y$ the condition ensures the ERM learnability Liu & Dietterich (2014) of the adversary-aware PLL problem if there is small ambiguity degree condition. In our case which is that, $P_{b,y'} < 1$ and $P_y = 1$. The y is the true label corresponding to each x the instance and y' , which is the rival or noisy label.

Assumption 1. According to Yu et al. (2018) that the minimization of the expected risk $R(f)$ given clean true population implies that the optimal classifier is able to do the mapping of $f_i^*(X) = P(Y = i \mid X)$, $\forall i \in [c]$. Given the assumption 2, we are able to draw conclusion that $\hat{f}^* = f^*$ applying the theorem 2 in the following.

Theorem 1. *Assume that the Adversary-Aware matrix $T_{y,y'}$ is fully ranked and the Assumption 2 is met, the the minimizer of $\hat{f}^*$ of $\hat{R}(f)$ will be converged to f^* of $R(f)$, meaning $\hat{f}^* = f^*$.*

Remark. If the $\bar{Q}$ and $T_{y,y'}$ is estimated correctly the empirical risk of the designed algorithm trained with adversary-aware partial label will converge to the expected risk of the optimal classifier trained with the true label. If the number of sample is reaching infinitely large that given the adversary-aware partial labels, $\hat{f}_n$ is going to converged to $\hat{f}^*$ theoretically. Subsequently, $\hat{f}_n$ will converge to the optimal classifier f^* as claimed in the theorem 1. With the new generation procedure, the loss function risk consistency theorems are introduced.

Theorem 2. *The adversary-aware loss function proposed is risk consistent estimator if it can asymptotically converge to the expected risk given sufficiently good approximate of $\bar{Q}$ and the adversary-aware matrix.*

$$\begin{aligned}\mathcal{L}(y, f(x)) = &\sum_{\vec{y}\in\vec{\mathcal{Y}}_y}\sum_{y=1}^{C}\sum_{y'\in Y'}(\mathrm{P}(Y = y \mid X = x) \\ &\prod_{b'\in\vec{y}} p_{b'} \cdot \prod_{t'\notin\vec{y}}\left(1 - p_{t'}\right)\bar{T}_{yy'}\vec{\mathcal{L}}(\vec{y}, f(x))) \\ &= \vec{\mathcal{L}}(\vec{y}, f(x)).\end{aligned} \tag{25}$$

The proof is in appendix lemma 4.

3.0.1 GENERALISATION ERROR

Define $\hat{R}$ and $\hat{R}_{pn}$ as the true risk the empirical risk respectively given the adversary-aware partial label dataset. The empirical loss classifier is obtained as $\hat{f}_{pn} = \arg\min_{f\in\mathcal{F}} \hat{R}_{pn}(f)$. Suppose a set of real hypothesis $\mathcal{F}_{\vec{y}_k}$ with $f_i(X) \in \mathcal{F}, \forall i \in [c]$. Also, assume it's loss function $\vec{\mathcal{L}}(\boldsymbol{f}(X), \vec{Y})$ is L-Lipschitz continuous with respect to $f(X)$ for all $\vec{y}_k \in \vec{\mathcal{Y}}$ and upper-bounded by M, i.e., $M = \sup_{x\in\mathcal{X}, f\in\mathcal{F}, y_k\in\vec{Y}} \vec{\mathcal{L}}(f(x), \vec{y}_k)$. The expected Rademacher complexity of $\mathcal{F}_k$ is denoted as $\Re_n(\mathcal{F}_{\vec{y}_k})$
Bartlett & Mendelson (2002)
Theorem 3. *For any $\delta > 0$, with probability at least $1-\delta$,*

$$\hat{R}\left(\hat{f}_{pn}\right) - \hat{R}\left(\hat{f}^{\star}\right) \leq 4\sqrt{2}L \sum_{k=1}^{c} \Re_n\left(\mathcal{F}_{\vec{y}_k}\right) + M\sqrt{\frac{\log\frac{2}{\delta}}{2n}}. \tag{26}$$

As the number of samples reaches to infinity $n \to \infty, \Re_n\left(\mathcal{F}_{\vec{y}_k}\right) \to 0$ with a bounded norm. Subsequently, $\bar{R}(\hat{f}) \to \bar{R}\left(\hat{f}^{\star}\right)$ as the number of training data reach to infinitely large. The proof is given in Appendix Theorem 3.

4 EXPERIMENTS

Datasets We evaluate the proposed method on three benchmarks-CIFAR10, CIFAR100 Krizhevsky et al. (2009), and fine-grained CUB200 Wah et al. (2011) with general partial label and adversary-aware partial label datasets.

Table 1: Benchmark datasets for accuracy comparisons. Superior results are indicated in bold. Our proposed methods have shown comparable results to fully supervised learning and outperform previous methods in a more challenging learning scenario, such as the partial rate at 0.5(CIFAR10) and 0.1(CIFAR100, CUB200). The hyper-parameter α is set to 0.1 for our method. (The symbol $*$ indicates Adversary-Aware partial label dataset).

Dataset	Method	$q = 0.01$	$q = 0.05$	$q = 0.1$
CIFAR100	(ITWM)(Without T)(Our)	**73.43** ±0.11	72.63±0.27	**72.35**±0.22
	PiCO	73.28 ±0.24	**72.90** ±0.27	71.77±0.14
	LWS	65.78 ±0.02	59.56 ±0.33	53.53 ±0.08
	PRODEN	62.60±0.02	60.73±0.03	56.80±0.29
	Full Supervised		73.56 ±0.10	
Dataset	Method	$q = 0.03 \pm 0.02$	$q = 0.05 \pm 0.02$	$q = 0.1 \pm 0.02$
CIFAR100	(ITWM)(Our)*	73.36 ±0.32	72.76±0.14	**54.09 ±1.88**
	PiCO*	72.87 ±0.26	72.53±0.37	48.03±3.32
	LWS*	46.8±0.06	24.82±0.17	4.53±0.47
	PRODEN*	59.33±0.48	41.20±0.27	13.44±0.41

Dataset	Method	$q = 0.01$	$q = 0.05$	$q = 0.1$
CUB200	(ITWM) (Without T)(Our)	**74.43**±0.876	**72.30**±0.521	**66.87**±0.98
	PiCO	74.11±0.37	71.75±0.56	66.12±0.99
	LWS	73.74±0.23	39.74±0.47	12.30±0.77
	PRODEN	72.34±0.04	62.56±0.10	35.89±0.05
	Full Supervised		76.02±0.19	
Dataset	Method	$q = 0.03 \pm 0.02$	$q = 0.05 \pm 0.02$	$q = 0.1 \pm 0.02$
CUB200	(ITWM) (Our)∗	**72.22**±1.36	**72.43**±0.86	**56.26**±0.70
	PiCO*	71.85 ±0.53	71.15 ±0.41	50.31 ±1.01
	LWS*	9.6 ±0.62	4.02 ±0.03	1.44 ±0.06
	PRODEN*	18.71±0.45	17.63 ±0.89	17.99 ±0.62

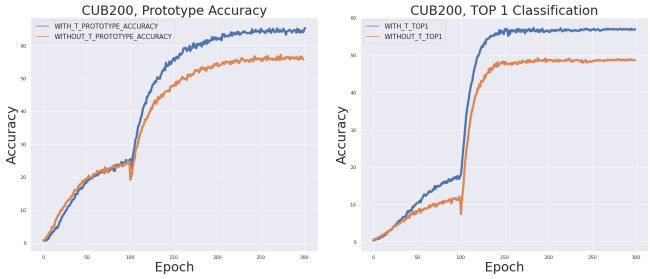
Dataset	Method	$q = 0.1$	$q = 0.3$	$q = 0.5$
CIFAR10	(ITWM)(Without T)(Our)	93.57±0.16	93.17±0.09	92.22±0.40
	PiCO	**93.74**±0.24	**93.25**±0.32	**92.46**±0.38
	LWS	90.30 ±0.60	88.99 ±1.43	86.16 ±0.85
	PRODEN	90.24±0.32	89.38±0.31	87.78±0.07
	Full Supervised		94.91±0.07	
Dataset	Method	$q = 0.1 \pm 0.02$	$q = 0.3 \pm 0.02$	$q = 0.5 \pm 0.02$
CIFAR10	(ITWM) (Our) *	93.52 ±0.11	**92.98**±0.51	**89.62**±0.79
	PiCO*	**93.64**±0.24	92.85±0.43	81.45±0.57
	LWS*	87.34±0.87	39.9±0.72	9.89±0.55
	PRODEN*	88.80±0.14	81.88±0.51	20.32±3.43

Main Empirical Results for CIFAR10. All the classification accuracy is shown in Table 1. We have compared classification results on CIFAR-10 with previous works Wang et al. (2022); Lv et al. (2020); Wen et al. (2021) using the immature teacher within momentum (ITWM). The method has shown consistently superior results in all learning scenarios where $q = \{0.3, 0.5\}$ for the adversary-aware partial label learning. More specifically, the proposed method achieves **8.17**$\%$ superior classification performance at a 0.5 partial rate than the previous state of art work Wang et al. (2022). Moreover, our proposed method has achieved comparable results at 0.1 and 0.3 partial rates. The experiments for CIFAR-10 have been repeated four times with four random seeds.

Main Empirical Results for CUB200 and CIFAR100. The proposed method has shown superior results for the Adversary-Aware Partial Label, especially in more challenging learning tasks like the 0.1 partial rate of the dataset cub200 and CIFAR100, respectively. On the cub200 dataset, we have shown **5.95%** improvement at partial rates 0.1 and 1.281% and 0.37% where the partial rate is at 0.05 and 0.03. On the CIFAR100 dataset, the method has shown **6.06%** and 0.4181%, 0.5414% higher classification margin at partial rate 0.1, 0.05 and 0.03. The experiments have been repeated five times with five random seeds.

4.1 Ablation Study

Adversary-Aware Loss Comparison. Figure 3 shows the experimental result comparisons for CUB200 between the modified loss function and cross-entropy loss function before and after the momentum updating strategy. Our method achieves SOTA performance. The adversary-aware matrix plays an indispensable role. In the first stage, the divergence becomes more apparent as the epoch reaches 100 epochs for CUB200 in Top-1 classification accuracy. The difference becomes more deviated in the second stage as the epoch reaches 200. The comparison demonstrated that the modified loss function works consistently throughout the whole stage of learning, especially for the more challenging learning scenario where the partial rate is at 0.1.



(a) (b)

Figure 2: The Top1 and Prototype Accuracy of the Proposed Method and the Method inWang et al. (2022) PiCO on CUB200.

5 Conclusion and Future works

This paper introduces a novel Adversary-Aware partial label learning problem. The new problem setting has taken local data privacy protection into account. The novel adversary-aware loss function, together with an immature teacher within momentum disambiguation algorithm, has achieved state-of-the-art performance and proven to be a provable classifier. Specifically, we have added the rival to the partial label candidate set as encryption for the dataset. Nonetheless, the generation process has made the intractable transition matrix even more complicated, leading to an inconsistency issue. As a result, the adversary-aware loss function is proposed, incorporating an adversary-aware matrix into the cross-entropy loss function to obtain a consistent classifier. The immature teacher with within momentum is proposed to solve intractable terms. Future work will use variational inference methods to approximate the intractable transition matrix.

REFERENCES

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Brian Chen, Bo Wu, Alireza Zareian, Hanwang Zhang, and Shih-Fu Chang. General partial label learning via dual bipartite graph autoencoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10502–10509, 2020.

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

Lei Feng and Bo An. Partial label learning with self-guided retraining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3542–3549, 2019.

Lei Feng, Jiaqi Lv, Bo Han, Miao Xu, Gang Niu, Xin Geng, Bo An, and Masashi Sugiyama. Provably consistent partial-label learning. *Advances in Neural Information Processing Systems*, 33: 10948–10960, 2020.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

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

Rong Jin and Zoubin Ghahramani. Learning with multiple labels. *Advances in neural information processing systems*, 15, 2002.

Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

Liping Liu and Thomas Dietterich. A conditional multinomial mixture model for superset label learning. *Advances in neural information processing systems*, 25, 2012.

Liping Liu and Thomas Dietterich. Learnability of the superset label learning problem. In *International Conference on Machine Learning*, pp. 1629–1637. PMLR, 2014.

Jiaqi Lv, Miao Xu, Lei Feng, Gang Niu, Xin Geng, and Masashi Sugiyama. Progressive identification of true labels for partial-label learning. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6500–6510. PMLR, 13–18 Jul 2020.

Gengyu Lyu, Songhe Feng, Tao Wang, Congyan Lang, and Yidong Li. Gm-pll: graph matching based partial label learning. *IEEE Transactions on Knowledge and Data Engineering*, 33(2): 521–535, 2019.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. 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.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-ucsd birds-200-2011. Technical report, 2011.

Haobo Wang, Yuzhou Qiang, Chen Chen, Weiwei Liu, Tianlei Hu, Zhao Li, and Gang Chen. Online partial label learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 455–470. Springer, 2020.

Haobo Wang, Ruixuan Xiao, Yixuan Li, Lei Feng, Gang Niu, Gang Chen, and Junbo Zhao. Pico: Contrastive label disambiguation for partial label learning. *ICLR*, 2022.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Hongwei Wen, Jingyi Cui, Hanyuan Hang, Jiabin Liu, Yisen Wang, and Zhouchen Lin. Leveraged weighted loss for partial label learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11091–11100. PMLR, 18–24 Jul 2021.

Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? *Advances in Neural Information Processing Systems*, 32:6838–6849, 2019.

Ning Xu, Jiaqi Lv, and Xin Geng. Partial label learning via label enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5557–5564, 2019.

Ning Xu, Congyu Qiao, Xin Geng, and Min-Ling Zhang. Instance-dependent partial label learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Xiyu Yu, Tongliang Liu, Mingming Gong, and Dacheng Tao. Learning with biased complementary labels. In *ECCV*, pp. 68–83, 2018.

Min-Ling Zhang, Bin-Bin Zhou, and Xu-Ying Liu. 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.

Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5(Oct):1225–1251, 2004.

Yivan Zhang and Masashi Sugiyama. Approximating instance-dependent noise via instance-confidence embedding. *arXiv preprint arXiv:2103.13569*, 2021.

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