1 APPENDIX

This file includes supplementary for all proofs and additional experiment details. The proofs for Theorem 1, Theorem 2, Lemma 4, Theorem 3, Lemma 4a, Lemma 4b and Additional Experiment are presented sequentially.

1.1 THE PROOF FOR THEOREM 1

Our goal is to find the optimal classifier, namely

$$h_i^*(X) = \mathrm{P}(Y = y \mid X = x) \forall i \in [c] \tag{1}$$

We can obtained the optimal classifier with modified loss function (equation 17) and assumption 1 when learning from examples with adversary-aware partial labels. The transition matrix of the adversary-aware partial label is defined as $\mathrm{P}(\vec{Y} \mid Y, Y', X)$ and denoted as $Q^* \in \mathbb{R}^{c \times (2^c - 2)}$. The partial label transition matrix $\mathrm{P}(\vec{Y} \mid Y)$ is denotes as $\bar{Q} \in \mathbb{R}^{c \times (2^c - 2)}$. Theoretically, if the true label Y of the vector $\vec{Y}$ is unknown given an instance X, where $\vec{y} \in \vec{Y}$ and there are $2^c - 2$ candidate label sets.The ϵ_x is the instance-dependent rival label noise for each instance where $\epsilon_x \in \mathbb{R}^{1 \times c}$. The class instance-dependent transition matrix is defined as $\bar{T}_{yy'} \in [0,1]^{C \times C}$, in which $\bar{T}_{yy'} = \mathrm{P}(Y' = y' \mid Y = y)$ and we assume $\bar{T}_{yy} = 0$, for $\forall_{yy'} \in [c]$, The inverse problem is to identify a sparse approximation matrix $\boldsymbol{A}$ given $\bar{T}$ to estimate the true posterior probability.

$$\underbrace{P(\vec{Y} \mid X)}_{\textbf{Adversary-aware PLL}} = ([\bar{\boldsymbol{Q}}^{\mathrm{T}} + \boldsymbol{\epsilon}]\bar{\boldsymbol{T}}) \underbrace{P(Y \mid X)}_{\textbf{True Posterior Probability}},$$

$$\bar{\boldsymbol{T}}^{-1}\boldsymbol{A}^{-1} \underbrace{P(\vec{Y} \mid X = x)}_{\textbf{Adversary-aware PLL}} \approx \underbrace{P(Y \mid X = x)}_{\textbf{True Posterior Probability}},$$

which further ensures

$$\underbrace{P(\vec{Y} \mid X)}_{\textbf{Adversary-aware PLL}} = ([\bar{Q}^T + \epsilon]\bar{T}) \underbrace{\mathbf{h}^*(X)}_{\textbf{True Posterior Probability}}. \tag{2}$$

where $Q^* = ([\bar{Q}^T + \epsilon]\bar{T})^T$. If the transition matrix $\bar{\mathbf{T}}$ is full rank and Q^* is identified, then we can define the optimal classifier $h^*(X) = \mathrm{P}(Y = y \mid X = x)$, which guarantees $\hat{f}^* = f^*$. The proof is completed.

1.2 THE PROOF FOR THEOREM 2

for any $x \in \mathcal{X}$, there holds

$$
\begin{aligned}
&\hat{\mathcal{R}}(\vec{\mathcal{L}}, f(X)) \\
=&\mathbb{E}_{\vec{Y}|X}[\vec{\mathcal{L}}(\vec{Y}, f(x)) \mid X = x] \\
=&\sum_{\vec{y} \in 2^{[C]}} \vec{\mathcal{L}}(\vec{y}, f(x)) \mathrm{P}(\vec{Y} = \vec{y} \mid X = x) \\
=&\sum_{\vec{y} \in 2^{[C]}} \vec{\mathcal{L}}(\vec{y}, f(x)) \sum_{y \in Y} \mathrm{P}(\vec{Y} = \vec{y}, Y = y \mid X = x) \\
=&\sum_{\vec{y} \in 2^{[C]}} \vec{\mathcal{L}}(\vec{y}, f(x)) \sum_{y \in Y} \sum_{y' \in Y'} \mathrm{P}(\vec{Y} = \vec{y}, Y = y, Y' = y' \mid X = x) \\
=&\sum_{\vec{y} \in 2^{[C]}} \vec{\mathcal{L}}(\vec{y}, f(x)) \\
&(\sum_{y \in Y} \sum_{y' \in Y'} \mathrm{P}(\vec{Y} = \vec{y} \mid Y = y, Y' = y', X = x) \mathrm{P}(Y' = y' \mid Y = y, X = x) \mathrm{P}(Y = y \mid X = x)) \\
=&\sum_{y=1}^{C} \mathrm{P}(Y = y \mid X = x) \\
&(\sum_{\vec{y} \in 2^{[C]}} \sum_{y' \in Y'} \mathrm{P}(\vec{Y} = \vec{y} \mid Y = y, Y' = y', X = x) \mathrm{P}(Y' = y' \mid Y = y, X = x) \vec{\mathcal{L}}(\vec{y}, f(x))) \\
=&\sum_{y=1}^{C} \mathrm{P}(Y = y \mid X = x) \\
&(\sum_{\vec{y} \in 2^{[C]}} \sum_{y' \in Y'} \mathrm{P}(\vec{Y} = \vec{y} \mid Y = y, Y' = y', X = x) \bar{T}_{yy'} \vec{\mathcal{L}}(\vec{y}, f(x))) \\
=&\sum_{y=1}^{C} \mathrm{P}(Y = y \mid X = x)
\end{aligned}
\tag{3}
$$

and

$$
\begin{aligned}
\mathcal{R}(\mathcal{L}, f(X)) &= \mathbb{E}_{Y|X}[\mathcal{L}(Y, f(x)) \mid X = x] \\
&= \sum_{y=1}^{C} \mathcal{L}(y, f(x)) \mathrm{P}(Y = y \mid X = x).
\end{aligned}
\tag{4}
$$

Since $\mathrm{P}(\vec{Y} = \vec{y} \mid Y = y, X = x) = 0$ for $\vec{y}$ does not have y for the condition that

$$
\begin{aligned}
&\mathcal{L}(y, f(x)) \\
=&\sum_{y=1}^{C} \mathrm{P}(Y = y \mid X = x) \sum_{\vec{y} \in 2^{[C]}} \sum_{y' \in Y'} \mathrm{P}(\vec{Y} = \vec{y} \mid Y = y, Y' = y', X = x) \bar{T}_{yy'} \vec{\mathcal{L}}(\vec{y}, f(x)) \\
&= \sum_{\vec{y} \in \vec{\mathcal{Y}}_y} \sum_{y=1}^{C} \sum_{y' \in Y'} \mathrm{P}(Y = y \mid X = x) \prod_{b' \in \vec{y}, b' \neq y,} p_{b'} \cdot \prod_{t' \notin \vec{y}} (1 - p_{t'}) \, \bar{T}_{yy'} \vec{\mathcal{L}}(\vec{y}, f(x)) \\
&= \sum_{\vec{y} \in \vec{\mathcal{Y}}_y} \prod_{b' \in \vec{y}, b' \neq y,} p_{b'} \cdot \prod_{t' \notin \vec{y}} (1 - p_{t'}) \, \vec{\mathcal{L}}(\vec{y}, f(x)).
\end{aligned}
\tag{5}
$$

1.2.1 THE PROOF FOR LEMMA 4

$$
\mathcal{L}(y, f(x)) = \sum_{\vec{y} \in \vec{\mathcal{Y}}_y} \prod_{b' \in \vec{y}, b' \neq y,} p_{b'} \cdot \prod_{t' \notin \vec{y}} (1 - p_{t'}) \, \vec{\mathcal{L}}(\vec{y}, f(x)) = \vec{\mathcal{L}}(\vec{y}, f(x)),
\tag{6}
$$

Ultimately, we can conclude that

$$
\hat{\mathcal{R}}(\vec{\mathcal{L}}, f(x)) = \mathcal{R}(\mathcal{L}, f(x)).
\tag{7}
$$

The proof is completed.

1.2.2 THE PROOF FOR THEOREM 3

The goal is to design a new loss function that will enable the hypothesis with adversary-aware partial labels to converge to the optimal classifier trained with true labels. We define $\vec{\mathcal{L}}$ as the new proposed loss function for the adversary-aware partial labels learning. Subsequently, the true and empirical loss function regarding the adversary-aware partial labels is stated as $\hat{R}(f) = \mathbb{E}_{(X,\vec{Y}) \sim P_{(X\vec{Y})}}[\vec{\mathcal{L}}(f(X), \vec{Y})]$ and $\hat{R}_{pn}(f) = \frac{1}{n} \sum_{i=1}^{n} \vec{\mathcal{L}}\left(f\left(x_i\right), \vec{y}_i\right)$, correspondingly. Moreover, we have defined $\{(\mathbf{x}_i, \vec{y}_i)\}_{1 \leq i \leq n}$ as the adversary-aware partial label sample space. The functions $\hat{f}^*$ and $\hat{f}_{pn}$ are the optimal classifier with minimum expected risk function $\hat{R}(f)$ and empirical $\hat{R}_{pn}(f)$ risk function respectively. Specifically, the model is formalised as $\hat{f}^* = \arg\min_{f \in \mathcal{F}} \hat{R}(f)$ and $\hat{f}_{pn} = \arg\min_{f \in \mathcal{F}} \hat{R}_{pn}(f)$. The objective of the newly proposed loss function $\vec{\mathcal{L}}$ is to ensure the convergence of the classifier trained with sample adversary-aware partial label to the optimal classifier trained with population dataset with true labels. Formally, the convergence of $\hat{f}_{pn} \xrightarrow{n} f^\star$ is obtained.

Definition. Lets denote $\vec{y}_k$ as kth element of the vector $\vec{y}$ being 1 and others being 0 if $\vec{y}_k \in \vec{y}$. The $\vec{y}$ is a candidate set of the adversary-aware partial label of an instance. Based on Lemma 1 and Theorem 1, the estimation error bound has been proven through

$$
\begin{aligned}
&\hat{R}\left(\hat{f}_{pn}\right) - \min_{f \in F} \hat{R}(f) = \hat{R}\left(\hat{f}_{pn}\right) - \hat{R}\left(\hat{f}^{\star}\right) \\
&= \hat{R}\left(\hat{f}_{pn}\right) - \hat{R}_{pn}(\hat{f}) + \hat{R}_{pn}(\hat{f}) - \hat{R}_{pn}\left(\hat{f}^{\star}\right) + \hat{R}_{pn}\left(\hat{f}^{\star}\right) - \hat{R}\left(\hat{f}^{\star}\right) \\
&\leq \hat{R}\left(\hat{f}_{pn}\right) - \hat{R}_{pn}(\hat{f}) + \hat{R}_{pn}\left(\hat{f}^{\star}\right) - \hat{R}\left(\hat{f}^{\star}\right) \\
&\leq 2 \sup_{f \in \mathcal{F}} \left| \hat{R}(f) - \hat{R}_{pn}(f) \right| \\
&\leq 4\Re\left(\mathcal{F}_v\right) + M\sqrt{\frac{\log \frac{2}{\delta}}{2n}} \\
&\leq 4\sqrt{2}L \sum_{k=1}^{c} \Re_n\left(\mathcal{F}_{\vec{y}_k}\right) + M\sqrt{\frac{\log \frac{2}{\delta}}{2n}}.
\end{aligned}
\tag{8}
$$

Given $\hat{R}_{pn}(\hat{f}) - \hat{R}_{pn}\left(f^{\star}\right) \leq 0$, the first inequality equation is established. The first three equations proof have been shown in Mohri et al. (2018).
The whole proof is based according to Bartlett & Mendelson (2002).

The definition 1 Suppose a space D and a sample distribution D_S are given in which $S = \{s_1, \ldots, s_n\}$ is a set of examples drawn independent, identically distributed from the distribution D_S. In addition, $\mathcal{F}$ is defined as a class of functions $f : S \to \mathbb{R}$. The empirical Rademacher complexity of $\mathcal{F}$ is defined as

$$
\hat{\Re}_n(\mathcal{F}) = \mathbb{E}_{\sigma}\left[\sup_{f \in \mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i f\left(x_i\right)\right)\right].
\tag{9}
$$

The expected Rademacher complexity of the function space $\mathcal{F}$ is denoted as

$$
\Re = \mathbb{E}_{D_S}\mathrm{E}_{\sigma}\left[\sup_{f \in \mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}\sigma_i f\left(x_i\right)\right)\right].
\tag{10}
$$

The independent random variables $\sigma_1, \ldots, \sigma_m$ are uniformly selected from $\{-1, 1\}$. We have defined the random variables as Rademacher variables. M is the upper bound of the loss function. Subsequently, for any $\delta > 0$, we will have at least probability $1 - \delta$

$$
\sup_{f \in \mathcal{F}} \left| \hat{R}(f) - \hat{R}_{pn}(f) \right| \leq 2\Re(\vec{\mathcal{L}} \circ \mathcal{F}) + M\sqrt{\frac{\log 1/\delta}{2n}},
\tag{11}
$$

where

$$
\Re(\vec{\mathcal{L}} \circ \mathcal{F}) = \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i \vec{\mathcal{L}}\left(f\left(X_i\right), \vec{Y}_i\right)\right],
\tag{12}
$$

is the function space with the expected Rademacher complexity and $\{\sigma_1, \cdots, \sigma_n\}$ are Rademacher variables which takes with value of positive and negative 1, such as $\{-1, 1\}$ with uniform probabil-

ity. The modified loss function $\vec{\mathcal{L}}$ has been defined in the following equations

$$\vec{\mathcal{L}}(f(X), \vec{Y}) = -\sum_{i=1}^{c} (\bar{q}_i) \log\left(\left(((\bar{\mathbf{T}} + \mathbf{I})^{\top} f(X))_i\right)\right), \tag{13}$$

$$\mathcal{F}_{\mathrm{V}} = \left\{ (X, \vec{Y}) \mapsto \sum_{i=1}^{c} (\bar{q}_i) \log\left(\left(((\bar{\mathbf{T}} + \mathbf{I})^{\top} f(X))_i\right)\right) \mid f \in \mathcal{F} \right\}, \tag{14}$$

$$\sup_{f \in \mathcal{F}} \left| \hat{R}(f) - \hat{R}_{pn}(f) \right| \leq 2\Re(\mathcal{F}_{\mathrm{V}}) + M\sqrt{\frac{\log 1/\delta}{2n}}. \tag{15}$$

According to McDiarmid's inequality McDiarmid et al. (1989), for any $\delta > 0$, with probability at least 1-$\delta/2$ the following equitation holds, namely

$$\sup_{f \in \mathcal{F}} \left| \hat{R}(f) - \hat{R}_{pn}(f) \right| \leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \hat{R}(f) - \hat{R}_{pn}(f) \right| \right] + M\sqrt{\frac{\log 1/\delta}{2n}}, \tag{16}$$

applying the symmetrization property Vapnik (1999) that we can acquire the following

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \hat{R}(f) - \hat{R}_{pn}(f) \right| \right] \leq 2\Re\left(\mathcal{F}_{\mathrm{V}}\right). \tag{17}$$

Assume the loss function $\vec{\mathcal{L}}\left(f(\boldsymbol{X}), \vec{Y}\right)$ has satisfied the L-Lipschitz property with respect to $f(\boldsymbol{X})(0 < L < \infty)$ with all $\vec{y}_k \in \vec{\mathcal{Y}}$ and lastly regarding to the Rademacher vector contraction inequality rule Maurer (2016) the inequality can be held

$$\Re\left(\mathcal{F}_V\right) \leq \sqrt{2}L \sum_{k=1}^{c} \Re_n\left(\mathcal{F}_{\vec{y}_k}\right). \tag{18}$$

The proof is completed.

1.2.3 The Proof for Lemma 4b

Since the loss function has been modified, we will show proof of the modified loss function. The modified loss function consisted of two components, the cross entropy loss function $\bar{\mathcal{L}}$ and a transition matrix and identity matrix. In this section, we introduce the modified loss function $\bar{\mathcal{L}}$ and proven through

$$\begin{aligned} \vec{\mathcal{L}}(f(X), \vec{Y}) &= -\sum_{i=1}^{c} (\bar{q}_i) \log\left(\left(((\bar{\mathbf{T}} + \mathbf{I})^{\top} f(X))_i\right)\right), \\ &= -\sum_{i=1}^{c} \mathbf{1}(\bar{q}_i) \log\left(\frac{\sum_{j=1}^{c} (\bar{T}_{ji}) \exp\left(g_j(X)\right)}{\sum_{k=1}^{c} \exp\left(g_k(X)\right)}\right), \end{aligned} \tag{19}$$

in which $((\mathbf{T}+\mathbf{I})^{\top} f(X))_i$ is defined as the i-th row of $(\mathbf{T}+\mathbf{I})^{\top} f; h : \mathcal{X} \to \mathbb{R}^c, f_i(X) \in \mathcal{H}, \forall i \in [c]$; In addition $f_i(X) = \frac{\exp(g_i(X))}{\sum_{k=1}^{c} \exp(g_k(X))}$.
The proof is completed.

1.2.4 Additional Experimental Details

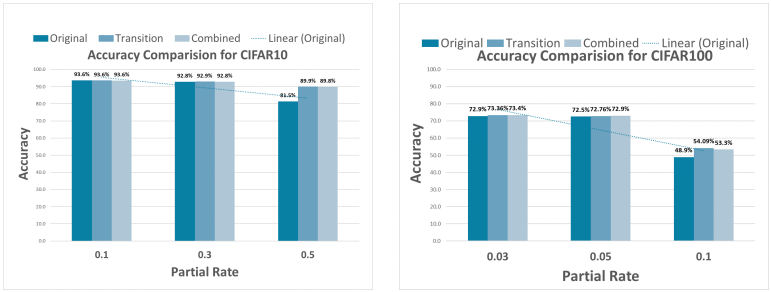
We have compared with most recent partial label learning algorithms, which are PICO Wang et al. (2022), LWSWen et al. (2021), and PRODENLv et al. (2020) on CIFAR-10Krizhevsky et al. (2009), CIFAR-100Krizhevsky et al. (2009) and CUB200Wah et al. (2011). The negative and rival labels of adversary-aware partial labels datasets are generated according to the probability $q^*_{b,l} := \mathrm{P}(b, l \in \vec{Y} \mid Y = y, Y' = l, X = x)$ with $b \neq y$. The class instance-dependent partial labels are manually generated. We have used the 0.02 proportion of the output $\Delta(f_i(X))$ corresponding to each instance after the softmax layer from the pre-trained classifier Resent18 He et al. (2016a). More specifically, we have defined all $C-1$ negative label where $\bar{y} \neq y$ with a uniform probability to be flipped to false positive. Finally, the probability can be defined as $q^*_{b,l} \pm 0.02$. The projection head of the contrastive network has 128-dimensional embedding with a 2-layer MLP. The data augmentation modules are following the previous work Wang et al. (2022). The queue size is fixed at 8192, 8192 and 4192 for the CIFAR-10, CIFAR-100 and CUB200 correspondingly. The momentum coefficients are 0.999 for the contrastive network update. The α is the hyperparameter of the immature teacher within momentum (ITWM), controlling the proportion of prototype updates. The $\alpha = 0.1$ and $\beta = 0.01$ are selected for the immature teacher within momentum (ITWM) without adversary-aware loss and the immature teacher within momentum (ITWM). The optimizer SGD with a momentum of 0.9 and 256 batch size are used to train the model for 299 epochs with a cosine learning rate schedule. Except for the total epochs, others are identical to the previous work Wang et al. (2022). For the temperature parameter τ, we have set it to 0.07. The loss weighting factors are set to $\lambda = \{0.5\}$. The partial label rate at $q \in \{0.1, 0.3, 0.5\}$ have been implemented for CIFAR-10 and $q \in \{0.03, 0.05, 0.1\}$ for CIFAR-100 and CUB200. The adversary partial label rate at $q^* \in \{0.1 \pm 0.02, 0.3 \pm 0.02, 0.5 \pm 0.02\}$ have been implemented for CIFAR-10 and $q^* \in \{0.03 \pm 0.02, 0.05 \pm 0.02, 0.1 \pm 0.02\}$ for CIFAR-100 and CUB200. Training without contrastive learning for CIFAR-10 is 1 epoch for all the partial rates with respect to clean partial labels. For CIFAR10 adversary-aware partial labels, the setting of 50 epochs training without contrastive learning is applied for q =$\{0.1, 0.3, 0.5\}$. We have trained without contrastive learning for the clean partial label with q = $\{0.01, 0.05, 0.1\}$ for epochs of $\{20, 20, 100\}$ on CIFAR-100 and CUB200. Moreover, the epochs of $\{20, 100, 100\}$ is set for the adversary-aware partial rate at q^*=$\{0.03 \pm 0.02, 0.05 \pm 0.02, 0.1 \pm 0.02\}$ of adversary-aware partial labels learning problem on CIFAR-100 and CUB200.

1.2.5 Additional Experiment for CIFAR-10

We have verified our method on an additional synthetic dataset, CIFAR-10. The implementation setting is mainly identical to Wang et al. (2022). For CIFAR-10 clean partial label learning, we have implemented the experiments according to each baseline's implementation details, and the best results were replicated from the baseline worksWang et al. (2022). The CIFAR-10 adversary-aware partial label problem has used the ResNet18 neural networkHe et al. (2016b) as the backbone. The $\alpha = 0.1$ and $\beta = 0.01$ are chosen for the immature teacher within momentum (ITWM) without (Clean partial label) and immature teacher within momentum (ITWM) method. The learning rate is 0.01, and the weight decay is $1e-3$. The ResNet-18 is used for training. For the clean partial label, q at $\{0.1, 0.3, 0.5\}$ is used for the experiments. The adversary-aware partial label is set to q^* = { 0.1±0.02, 0.3±0.02, 0.5±0.02 } for experiments. We have trained the model without contrastive loss for the epochs of $\{1, 1, 1\}$ with the clean partial label at partial rate of q = $\{0.1, 0.3, 0.5\}$. We have trained the model without contrastive learning for epochs of $\{50, 50, 50\}$ for the adversary-aware partial labels with partial rates of q^* = { 0.1±0.02, 0.3±0.02, 0.5±0.02 }.

1.2.6 The Classification Accuracy Comparisons

Our proposed methods have consistently outperformed the previous works for the most challenging scenarios q =$\{$0.5,0.5,0.1$\}$ on CIFAR-10, CIFAR-100 and CUB200.



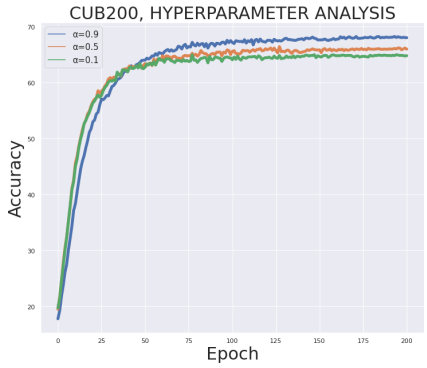
(a) (b)

(c)

Figure 1: The Classification Accuracy Comparisons

1.2.7 The Hyperparameter Comparisons

We have also conducted a comparative analysis on the impact of hyperparameter α on the final classification performance. The larger the hyperparameter, the better the classification performance. Our proposed method has compared the hyperparameter α at $\{0.1,0.5,0.9\}$ for all dataset. The $\alpha = 0.1$ has been chosen throughout the experiments.

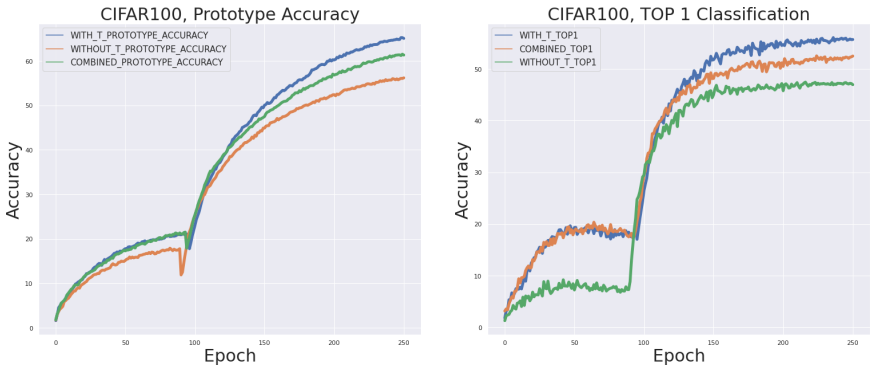


(a)

Figure 2: The Classification Accuracy of our proposed method using α **=[0.1,0.5,0.9]** for CUB200

1.2.8 Adversary-Aware Loss Comparison.

Figure 3 shows the experimental result comparisons for CIFAR100 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 CIFAR100 in Top-1 classification accuracy. 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 3: The Top 1 and Prototype Accuracy of the Proposed Method and the Method in Wang et al. (2022) PiCO on CIFAR100.

1.3 Implementation Details

Adversary-Aware Matrix. The transition matrix is a common tool for building statistically consistent classifiers in noise label and complementary label problems Yu et al. (2018); Huang et al. (2006); Liu & Dietterich (2014). In this paper that we have introduced the adversary-aware matrix for building statistically consistent classifiers for the adversary-aware partial label problem. The Adversary-Aware Matrix $\in \mathbb{R}^{c\times c}$ is constructed as $T_{y,y'} = \bar{T} + I$. We set the diagonal element of $\tilde{T}$ to one to ensure $y \in \vec{y}$.

New rival Label. The rival label is generated according to the label noise transition matrix $\bar{T}$. We have defined ordinary partial label generation B. The B is defined as $\mathrm{P}(\vec{Y} \mid X)$, general partial label generation, is defined accordingly as E.q 2. In application, we can randomly give out proportional of survey with rival and other without the rival according to the adversary aware matrix. This will ensure that adversary will not be able to retrieve the insightful data by purposely enquiry a participant to reveal given out answers. By formulating the rival as $\boldsymbol{R} = \mathrm{P}(\vec{Y} \mid X)$, which equal to $\min\{1, B^{(2^c-2)\times c}\bar{T}^{c,c}\}$ and $R_{i,j} \in [0,1]^{(2^c-2)\times c}$, for $\forall_{i,j} \in [c]$. We now have the adversary aware partial label.

1.4 Ablation Study for $\bar{T}$

In the following, we have shown how the classification performance is impacted if the entries of the class instance-dependent transition matrix $\bar{T}_{\textbf{Original}}$ is updated to $\bar{T}_{\textbf{New}}$ to show the robustness of our proposed method. For instance if the number of class is equal to 10, then the entries of $\bar{T}$ is defined as below. In our problem setting, each row has five entries equal to 0.2 in the original $\bar{T}$ and has five entries equal to 0.3 each row for new $\bar{T}$.

$$\bar{T}_{\textbf{Original}} = \begin{bmatrix} 0 & 0.2 & 0 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0 & 0.2 & 0.2 & 0 & 0.2 \\ 0.2 & 0.2 & 0 & 0.2 & 0 & 0.2 \\ 0.2 & 0.2 & 0 & 0 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0 & 0.2 & 0 & 0.2 \\ 0.2 & 0 & 0.2 & 0.2 & 0.2 & 0 \end{bmatrix} \bar{T}_{\textbf{New}} = \begin{bmatrix} 0 & 0.3 & 0 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0 & 0.3 & 0.3 & 0 & 0.3 \\ 0.3 & 0.3 & 0 & 0.3 & 0 & 0.3 \\ 0.3 & 0.3 & 0 & 0 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0 & 0.3 & 0 & 0.3 \\ 0.3 & 0 & 0.3 & 0.3 & 0.3 & 0 \end{bmatrix}$$

Data	Method	q^*=0.1
CIFAR100	PiCOWang et al. (2022)	20.941(24.015)%
Data	Method	q^*=0.1
CIFAR100	ATM	**54.156(0.066)**%

Data	Method	q^*=0.1
CUB200	PiCOWang et al. (2022)	21.22(-25.155)%
Data	Method	q^*=0.1
CUB200	ATM	**48.62(-7.64)**%

1.5 WHY ADVERSARY AWARE PARTIAL LABEL LEARNING IS A MORE CHALLENGING PROBLEM

By adding the rival, the partial label generation process compare with the new label generation is easier:

$$A = \begin{bmatrix} 0 & 0.25 & 0 & 0.25 & 0.25 & 0.25 \\ 0.2 & 0 & 0.25 & 0.25 & 0 & 0.25 \\ 0.2 & 0.25 & 0 & 0.25 & 0 & 0.25 \\ 0.25 & 0.25 & 0 & 0 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0 & 0.25 & 0 & 0.25 \\ 0.25 & 0 & 0.25 & 0.25 & 0.25 & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} 1 & 0.5 \pm 0.02 & 0.5 \pm 0.02 & 0.5 \pm 0.02 & 0.5 \pm 0.02 & 0.5 \pm 0.02 \\ 0.5 \pm 0.02 & 1 & 0.5 \pm 0.02 & 0.5 \pm 0.02 & 0 & 0.5 \pm 0.02 \\ 0.5 \pm 0.02 & 0.5 \pm 0.02 & 1 & 0.5 \pm 0.02 & 0 & 0.5 \pm 0.02 \\ 0.5 \pm 0.02 & 0.5 \pm 0.02 & 0 & 1 & 0.5 \pm 0.02 & 0.5 \pm 0.02 \\ 0.5 \pm 0.02 & 0.5 \pm 0.02 & 0 & 0.5 \pm 0.02 & 1 & 0.5 \pm 0.02 \\ 0.5 \pm 0.02 & 0 & 0.5 \pm 0.02 & 0.5 \pm 0.02 & 0.5 \pm 0.02 & 1 \end{bmatrix}$$

By adding the rival using the label noise transition matrix as such $A \times B$ and we can conclude that $[A \times B]_{ij} > [B]_{ij}$. Even though noises added to the partial label noise has made the problem more challenging, unless the adversary-aware transition matrix is given, it will greatly help us to reduce the uncertainty from the transition matrix.

1.5.1 ALGORITHM TABLE

Algorithm 1 Adversary Aware Partial label learning

Goal: Minimise the Total loss function λ **Input:** The Adversary-Aware PLL $\bar{\mathcal{D}}$ and Batch size Samples $\bar{\mathcal{D}}_b$. **Output:** The optimal W of the Total Loss Function.

for `x_i` $\in$ `Total Epochs` **do**
  $\vec{D}_b \in \vec{D}$
  $D_q = \{\boldsymbol{u}_i = f\left(\mathrm{Aug}_q\left(\boldsymbol{x}_i\right)\right) \mid \boldsymbol{x}_i \in \vec{D_b}\}$
  $D_k = \{\boldsymbol{z}_i = f'\left(\mathrm{Aug}_k\left(\boldsymbol{x}_i\right)\right) \mid \boldsymbol{x}_i \in \vec{D_b}\}$
  $\bar{\mathcal{C}} = D_q \cup D_k \cup$ queue
  for $\mathbf{x}_i \in \vec{D}$ **do**
    $\hat{y}_i = \arg\max_{c \in Y_i} f^c\left(\mathrm{Aug}_q\left(\boldsymbol{x}_i\right)\right)$
    $\boldsymbol{v}_i^{t+1} = \sqrt{1-\alpha^2}\boldsymbol{v}_i^t + \alpha\frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2}$
    $N_+(\boldsymbol{x}_i) = \left\{\boldsymbol{z}' \mid \boldsymbol{z}' \in \bar{\mathcal{C}}\left(\boldsymbol{x}_i\right), \bar{y}' = (\hat{y}_i = c)\right\})$
  end for
  for $\mathbf{u}_i \in D_q$ **do**

$$r_c = \begin{cases} 1 & \text{if } c = \arg\max_{j \in Y} \boldsymbol{u_i}^\top \boldsymbol{v}_j \\ 0 & \text{otherwise} \end{cases},$$

    $\bar{\boldsymbol{q}} = \phi\bar{\boldsymbol{q}} + (1-\phi)\boldsymbol{r_c}$
  end for
  $\mathcal{L}$=$\lambda\mathcal{L}(f(x_i), \tau, C)$+$\vec{\mathcal{L}}(f(x_i), \vec{Y})$ ▷ Equation 18 + Equation 17
  $\mathcal{L}$=$\lambda\mathcal{L} + \vec{\mathcal{L}}$ ▷ Total Loss
end for

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016b.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19, 2006.

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

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.

Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pp. 3–17. Springer, 2016.

Colin McDiarmid et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1): 148–188, 1989.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

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

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.

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.

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