

A PROOFS

A.1 RADEMACHER COMPLEXITY

We define the Rademacher complexity over the hypothesis space \mathcal{F} to facilitate the theorem proof.

Definition A.1 (Rademacher Complexity). *Let $\mathcal{D}_m = \{x_1, \dots, x_m\}$ be a set of identically independently distributed instances that are drawn from target distribution $P(X)$. The empirical Rademacher complexity of \mathcal{F} with respect to \mathcal{D}_m is defined as*

$$\widehat{\mathfrak{R}}(\mathcal{F}; \mathcal{D}_m) := \frac{1}{m} \mathbb{E}_{\varepsilon} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^m \sum_{j \in \mathcal{Y}} \varepsilon_{i,j} f^j(x_i) \right] \quad (6)$$

where $f^j(x_i)$ is the j -th model prediction of $x_i \in \mathcal{D}_m$; $\varepsilon_{i,j}$ is the independent Rademacher random variables uniformly sampled from $\{-1, +1\}$. The Rademacher complexity is the expectation of this empirical complexity:

$$\mathfrak{R}(\mathcal{F}) := \mathbb{E}_{\mathcal{D}_m \sim P(X)} [\widehat{\mathfrak{R}}(\mathcal{F}; \mathcal{D}_m)] \quad (7)$$

A.2 LIPSCHITZNESS

Lemma A.2 (Contraction lemma (Lemma 5 from Cortes et al. (2016))). *Let N and m be two positive integers. Let also \mathcal{H} be a set of functions that map \mathcal{X} to \mathbb{R}^N . Suppose that for each $i \in [m]$, function $\Psi_i : \mathbb{R}^N \rightarrow \mathbb{R}$ is μ_i -Lipschitz with the 2-norm, i.e.,*

$$|\Psi_i(v') - \Psi_i(v)| \leq \mu_i \|v' - v\|_2 \quad \forall v, v' \in \mathbb{R}^N \quad (8)$$

Then, for any set of m points $x_1, \dots, x_m \in \mathcal{X}$, the following holds:

$$\frac{1}{m} \mathbb{E}_{\sigma} \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^m \sigma_i \Phi_i(h(x_i)) \right] \leq \frac{\sqrt{2}}{m} \mathbb{E}_{\epsilon} \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^m \sum_{j=1}^N \epsilon_{ij} \mu_{ij} h_j(x_i) \right]$$

where the σ_i 's and the ϵ_{ij} 's are independent Rademacher variables uniformly distributed over $\{-1, +1\}$.

In the following lemma, we consider the Lipschitzness of the SAT loss.

Lemma A.3 (Lipschitzness). *Given a classifier $f : \mathcal{X} \rightarrow \mathcal{Y}$, with data sample $\forall (x, s) \in \mathcal{D}$, let $f(x)$ be the prediction output. For $\forall f, f' \in \mathcal{F}$, we have:*

$$|\ell^{SAT}(f(x), s) - \ell^{SAT}(f'(x), s)| \leq \sqrt{|s|} |f(x_i) - f'(x_i)|_2 \quad (9)$$

where s is the corresponding candidate label set; $|s|$ denotes the number of the candidate labels.

Proof. Given (x, s) sampled from a partial label dataset \mathcal{D} , we use s to represent the candidate label set and $f^i(x)$, $f^j(x)$ to indicate the prediction probabilities of class i and class j by a classifier f . We first find the gradient of the SAT loss function ℓ^{SAT} with respect to the classifier prediction $f(x)$.

$$\begin{aligned} \frac{\partial \ell^{SAT}}{\partial f^i(x)} &= \frac{\partial}{\partial f^i(x)} \left[-\log \left(1 - \prod_{j \in s} (1 - f^j(x)) \right) \right] \\ &= -\frac{1}{1 - \prod_{j \in s} (1 - f^j(x))} \cdot \frac{\partial}{\partial f^i} \left[\prod_{j \in s} (1 - f^j(x)) \right] \end{aligned}$$

Now, the derivative of the product inside:

$$\frac{\partial}{\partial f^i(x)} \left[\prod_{j \in s} (1 - f^j(x)) \right] = \prod_{j \in s, j \neq i} (1 - f^j(x)) \cdot (-1) \quad (10)$$

So, combining these:

$$\frac{\partial \ell^{SAT}}{\partial f^i(x)} = \frac{\prod_{j \in s, j \neq i} (1 - f^j(x))}{1 - \prod_{j \in s} (1 - f^j(x))} \quad (11)$$

By Equation 11, we have the derivative of the SAT loss as:

$$\begin{aligned} \frac{\partial \ell^{SAT}(x, s)}{\partial f^j(x)} &= \frac{\prod_{k \in s \setminus \{j\}} (1 - f^k(x))}{1 - \prod_{j \in s} (1 - f^j(x))} \\ &= \frac{1}{\prod_{k \in s \setminus \{j\}} (1 - f^k(x)) - (1 - f^j(x))} \leq 1 \end{aligned} \quad (12)$$

Known that a function f is **Lipschitz continuous** if there exists a constant L such that for all x, y in the domain of function h :

$$|h(x) - h(y)| \leq L \|x - y\|$$

where $\|x - y\|$ denotes the norm of the vector difference between x and y .

By *mean value theorem*, we have:

$$h(x) - h(y) = \nabla h(c) \cdot (x - y) \quad (13)$$

for some c in the line segment between x and y . Given the bound on the derivatives, this can be written as:

$$|h(x) - h(y)| \leq \|\nabla h(c)\| \|x - y\|$$

Since $\left| \frac{\partial \ell^{SAT}}{\partial f^j(x)} \right| \leq 1$ for all j , the norm of the gradient $\nabla \ell^{SAT}$ at any point is bounded by the square root of the number of components j in s , each component being bounded by 1. Thus:

$$\|\nabla \ell^{SAT}(c)\| \leq \sqrt{|s|} \quad (14)$$

Let $|s|$ represents the number of candidate classes for which the probabilities $f^j(x)$ are being estimated, the Lipschitz constant L can be set as $\sqrt{|s|}$, which effectively bounds the change in the loss function ℓ^{SAT} in terms of the change in the probabilities. Hence:

$$|\ell^{SAT}(f(x), s) - \ell^{SAT}(f'(x), s)| \leq \sqrt{|s|} \|f(x) - f'(x)\|_2 \quad (15)$$

This ends the proof of Lemma A.3 □

A.3 PROOF OF PROPOSITIONS

Proposition 5.3. [Low entropy preference] Let $H(f(x); s) := -\sum_{j \in s} f^j(x) \log f^j(x)$ denote the entropy of the PLL sample (x, s) subject to f . Then $\ell^{SAT}(f(x), s) \propto H(f(x); s)$.

To prove the proposition stated, we need to demonstrate that as the maximum prediction probability among the candidate labels increases, the SAT loss decreases, and this loss is proportional to the entropy of the candidate label set s .

Proof. ℓ^{SAT} and Maximum Prediction Probability

Let $f^i(x) = \max_{j \in s} f^j(x)$. To analyze the behavior of ℓ^{SAT} , consider the function $\prod_{j \in s} (1 - f^j(x))$. As $f^i(x)$ increases, at least one term $(1 - f^j(x))$ decreases, thereby decreasing the entire product $\prod_{j \in s} (1 - f^j(x))$. Consequently, the term $1 - \prod_{j \in s} (1 - f^j(x))$ increases. Applying the negative logarithm, which is a monotonically decreasing function, we observe that ℓ^{SAT} decreases. Thus, an increase in $f^i(x)$ leads to a decrease in ℓ^{SAT} .

Relationship between ℓ^{SAT} and Entropy $H(f(x); s)$

To establish the relationship with entropy, consider the expression for entropy $H(f(x); s)$. High entropy indicates a more uniform distribution of probabilities across the labels $j \in s$. As entropy decreases, indicating less uniformity, there tends to be a dominant label j with a high $f^j(x)$, which is close to $f^i(x)$.

From Part 1, we know that an increase in $f^i(x)$ (associated with low entropy) leads to a decrease in ℓ^{SAT} . Hence, we can infer that ℓ^{SAT} is inversely related to the entropy of the candidate label set s . Specifically, lower entropy (more certainty and less uniformity among label probabilities) results in a lower ℓ^{SAT} .

To summarize, $\ell^{SAT}(f(x_i); s) \propto H(f(x); s)$ indicates that as the entropy of the probability distribution of the labels in s decreases (implying increasing certainty or predictability among the labels), the SAT loss decreases, validating the proposed relationship. \square

A.4 PROOF OF ERROR BOUND UNDER SMALL AMBIGUITY DEGREE

Proposition A.4 (Partial loss bound via ambiguity degree). *(Proposition 1 in [Cour et al. \(2011b\)](#))*
For any classifier $f \in \mathcal{F}$, with a partial label dataset \mathcal{D}_N and small ambiguity degree γ , we have:

$$\mathcal{R}^{01}(f; \mathcal{D}_N) \leq \frac{1}{1 - \gamma} \mathcal{R}_P^{01}(f; \mathcal{D}_N) \quad (16)$$

where ℓ^{01} and ℓ_P^{01} are the zero-one loss and partial zero-one loss separately.

Lemma A.5. For $\forall f \in \mathcal{F}$, the partial labeled data instance $(x, s) \in \mathcal{D}$, its candidate label set s , we have

$$\ell_P^{01}(f(x), s) \leq \ell^{SAT}(f(x), s) \leq \ell_P^{CE}(f(x), s) \quad (17)$$

Proof. Let $F_1 = \prod_{j \in s} (1 - f^j(x))$ and $F_2 = \prod_{j \in s} f^j(x)$, where $f^j(x)$ is the prediction probability of class j that always smaller than 1. We have $-\log(1 - F_1) \leq -\log(F_2)$, Thus:

$$\ell^{SAT}(f(x), s) \leq \ell_P^{CE}(f(x), s) \quad (18)$$

The SAT loss lower bounds the Partial Cross Entropy Loss.

Since the zero-one loss ℓ_P^{01} is either 0 or 1 we have:

$$\ell_P^{01}(f(x), s) \leq \ell^{SAT}(f(x), s) \quad (19)$$

This concludes the proof of Lemma A.5 \square

Theorem 5.4. [Error bound under small ambiguity degree] Given partial labeled dataset \mathcal{D}_N with small ambiguity degree $\gamma \in (0, 1)$, $\forall \epsilon, \delta \in (0, 1)$, with probability at least $1 - \delta$, we have:

$$\mathcal{R}^{01}(f; \mathcal{D}_N) \leq \frac{1}{1 - \gamma} \left(\hat{\mathcal{R}}^{SAT}(f; \mathcal{D}_N) + 2\sqrt{(C - 1)\gamma + 1} \mathfrak{R}_N(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2N}} \right). \quad (5)$$

$\hat{\mathcal{R}}^{SAT}(f; \mathcal{D}_N)$ is the empirical risk of SAT loss ℓ^{SAT} over an arbitrary partial labeled dataset \mathcal{D}_N ; $\mathfrak{R}(\mathcal{F})$ denotes the Rademacher complexity (Definition A.1) of hypothesis space \mathcal{F} .

Proof. For an arbitrary partial labeled dataset $\mathcal{D}_N = (x, s)_{i=1}^N$, we have the below inequality holds with probability at least $1 - \delta$, where $\delta \in (0, 1)$, according to the standard Rademacher complexity bounds.

$$\mathcal{R}^{SAT}(f; \mathcal{D}_N) \leq \widehat{\mathcal{R}}^{SAT}(f; \mathcal{D}_N) + 2\mathfrak{R}_N(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2N}} \quad (20)$$

Here $\mathcal{F} = \{(x, s) \mapsto \ell^{SAT}(f(x); s) : f \in \mathcal{F}\}$.

Combined with Lemma A.2 and Lemma A.3, we have:

$$\begin{aligned} \mathfrak{R}_N(\mathcal{F}) &= \frac{1}{N} \mathbb{E}_{\mathcal{D}_N} \mathbb{E}_{\varepsilon} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^N \sum_{j \in \mathcal{Y}} \sqrt{|s|} \varepsilon_{ij} f^j(x_i) \right] \\ &= \frac{\sqrt{|s|}}{N} \mathbb{E}_{\mathcal{D}_N} \mathbb{E}_{\varepsilon} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^N \sum_{j \in \mathcal{Y}} \varepsilon_{ij} f^j(x_i) \right] \\ &= \sqrt{|s|} \mathfrak{R}_N(\mathcal{F}) \end{aligned} \quad (21)$$

From Proposition A.4, we have

$$\mathcal{R}^{01}(f; \mathcal{D}_N) \leq \frac{1}{1 - \gamma} (\mathcal{R}_P^{01}(f; \mathcal{D}_N)) \quad (22)$$

where γ is the small ambiguity level within range $[0, 1)$.

By Lemma A.5, we have $\ell_P^{01}(f(x), s) \leq \ell^{SAT}(f(x), s)$. As the risk is the expectation of the loss function, we can get:

$$\mathcal{R}_P^{01}(f; \mathcal{D}_N) \leq \mathcal{R}^{SAT}(f; \mathcal{D}_N) \quad (23)$$

In addition, we have the risk bound of partial cross-entropy loss as

$$\widehat{\mathcal{R}}^{SAT}(f; \mathcal{D}_N) \leq \widehat{\mathcal{R}}_P^{CE}(f; \mathcal{D}_N) \quad (24)$$

By integrating the above inequalities, we conclude the proof of Theorem 5.4. \square

B ADDITIONAL EXPERIMENTS

B.1 EMPIRICAL RESULT ON NOISY PARTIAL LABEL DATA

We evaluate the performance of SAT loss and SAT-integration methods on noisy partial labels, where the ground truth label may not always be in the candidate label set. The noise rate η represents the probability that the true label is not included in the candidate label set. we set the uniform partial label rate for CIFAR-100 and CUB-200 as $\mathbf{p} = 0.2$ and $\mathbf{p} = 0.05$ separately, and change the noisy rate η from 0.05 to 0.2.

Table 3 demonstrates the superiority of ℓ^{SAT} under varying noise conditions. Specifically, the integration of ℓ^{SAT} into PiCO⁺ improves the accuracy by 13.06% on CIFAR-100 for $\eta = 0.2$. For CIFAR-100, PAPI_{SAT} outperforms all other configurations, achieving the best results under all noise rates, which highlights ℓ^{SAT} 's resilience to noise. Conversely, on CUB-200, PiCO⁺_{SAT} provides the most significant improvements, particularly at higher noise rates.

B.2 EMPIRICAL RESULT ON REAL-WORLD DATASET

The Pascal VOC 2007 dataset(Everingham et al.) is a widely used benchmark in visual object classification and detection tasks. It contains 9,963 images with 20 object categories, including animals,

Table 3: Mean classification accuracy on CIFAR-100 and CUB-200 for noisy and uniform partial labels. **Best results** across all baselines are in red.

METHODS	CIFAR-100 ($q = 0.2$)			CUB-200 ($q = 0.05$)		
	$\eta = 0.05$	$\eta = 0.1$	$\eta = 0.2$	$\eta = 0.05$	$\eta = 0.1$	$\eta = 0.2$
CC	57.20%	50.73%	37.86%	43.94%	41.82%	30.48%
RC	50.01%	44.18%	32.88%	53.05%	47.45%	36.64%
SAT	67.84%	60.08%	39.75%	65.21%	60.00%	53.15%
PAPi	72.77%	72.26%	71.06%	34.79%	27.38%	17.85%
PAPi _{SAT}	74.89%	74.10%	73.66%	60.92%	46.87%	28.96%
PiCO ⁺	54.01%	49.33%	45.07%	65.87%	63.95%	60.04%
PiCO ⁺ _{SAT}	66.61%	62.92%	58.13%	68.43%	65.88%	63.19%

Table 4: Mean classification accuracy on PASCAL VOC 2007.

METHODS	PASCAL VOC 2007		
	$q = 0.1$	$q = 0.3$	$q = 0.5$
PAPi	85.05%	66.21%	30.04%
PAPi _{SAT}	85.46%	73.22%	46.39%
PiCO ⁺	65.60%	33.31%	26.43%
PiCO ⁺ _{SAT}	68.54%	42.88%	35.02%

vehicles, and household items, making it diverse and representative of real-world environments. The dataset is challenging due to the presence of multiple objects in varying scales, occlusions, and cluttered backgrounds. Its inherent ambiguity in object labels and annotations makes it an ideal testbed for evaluating PLL methods under real-world conditions, where accurate and unambiguous labeling is often impractical. We provide some image examples in Figure 5. In the experiment, we consider partial label rate $q \in \{0.1, 0.3, 0.5\}$.

In Table 4, the empirical results on Pascal VOC 2007 (Everingham et al.), are presented to demonstrate the efficacy of the proposed SAT loss in handling PLL under real-world conditions. The table summarizing the results reveals that the integration of SAT loss into existing PLL methods significantly improves classification accuracy across different ambiguity levels. In particular, PAPi_{SAT} consistently outperformed other methods, achieving the highest accuracy at each partial label rate, with an improvement of over 15% when $q = 0.5$ compared to other methods.



Figure 5: Exemplar images from Pascal VOC 2007 dataset