Dist-PU: Positive-Unlabeled Learning from a Label Distribution Perspective (Supplementary Material)

A. Proof of Proposition 1

A.1. Preliminaries

Definition 1 (Bounded Difference Condition [9])**.** Given m-sample datasets $\boldsymbol{X} = (\boldsymbol{x}_1, .., \boldsymbol{x}_i, .., \boldsymbol{x}_m)$ and $\boldsymbol{X}' = (\boldsymbol{x}_1, .., \boldsymbol{x}'_i, .., \boldsymbol{x}_m)$ with only the i-th sample being different, a function $\phi : \mathbb{R}^{dm} \to \mathbb{R}$ satisfies the bounded difference condition if the following inequality holds:

$$|\phi(\boldsymbol{X}) - \phi(\boldsymbol{X}')| \leq \frac{1}{m}. \tag{24}$$

Definition 2 (Rademacher Variables [11])**.** Rademacher variables $\boldsymbol{\sigma} = (\sigma_1, .., \sigma_m)$ consist of randoml variables in $\{-1, +1\}$ with the same probability $\Pr(\sigma = \pm 1) = 0.5$.

Definition 3 (Rademacher Complexity [11])**.** By introducing the Rademacher variables $\boldsymbol{\sigma}$, the Empirical Rademacher Complexity estimates the richness of a function class $\mathcal{F}$ by measuring the ability to fit to random noise on a m-sample dataset $\boldsymbol{X} = (\boldsymbol{x}_1, .., \boldsymbol{x}_m)$:

$$\hat{\mathfrak{R}}_X(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left[\sum_{i=1}^{m}\sigma_i f(\boldsymbol{x}_i)\right]\right]. \tag{25}$$

The Rademacher complexity is then defined as its expectation w.r.t. the dataset:

$$\mathfrak{R}_m(\mathcal{F}) = \mathbb{E}_{\boldsymbol{X}}\left[\hat{\mathfrak{R}}_X(\mathcal{F})\right]. \tag{26}$$

Theorem 1 (McDiarmid Inequality [9])**.** *Let a function ϕ satisifies the **bounded difference condition**, given a m-sample dataset $\boldsymbol{X} = (\boldsymbol{x}_1, .., \boldsymbol{x}_m)$ consisting of m independent random variables, Then $\forall \epsilon > 0$,*

$$\Pr\left(\phi(\boldsymbol{X}) - \mathbb{E}_{\boldsymbol{X}}\left[\phi(\boldsymbol{X})\right] > \epsilon\right) \leq e^{-2m\epsilon^2}. \tag{27}$$

Lemma 1 (Talagrand's Contraction Lemma [10])**.** *For a k-Lipschitz function $\zeta : \mathbb{R} \to \mathbb{R}$, it holds that*

$$\mathfrak{R}_m(\zeta \circ \mathcal{F}) \leq k \cdot \mathfrak{R}_m(\mathcal{F}). \tag{28}$$

A.2. Bounding the expected risk R

Let the sign of $f(\boldsymbol{x})$ determine its predicted label, since $y \in \{0, 1\}$, the classification error could be reformulated by the zero-one loss $\ell^{0-1}(f(\boldsymbol{x}), y) = \frac{1}{2}(1 - (2y-1)\operatorname{sgn}[f(\boldsymbol{x})])$. Recall that we approximate $\Pr(\hat{y} = 1|\boldsymbol{x})$ through a sigmoid function over $f(\boldsymbol{x})$, it is equivalent with adopting the sigmoid loss as the surrogate of the zero-one loss, which is defined as:

$$\ell^{sig}(f(\boldsymbol{x}), y) = \frac{1}{1 + \exp[(2y-1)f(\boldsymbol{x})]}. \tag{29}$$

Depending on the value of y, it could be rewritten as:

$$\ell^{sig}(f(\boldsymbol{x}), y) = \begin{cases} \dfrac{1}{1+\exp[-f(\boldsymbol{x})]} = s & y = 0, \\ \dfrac{1}{1+\exp[f(\boldsymbol{x})]} = 1 - s & y = 1. \end{cases} \tag{30}$$

Consequently, the expected risk R in the label-distribution-alignment manner using the sigmoid loss is formulated as:

$$\begin{aligned} R^{sig} =& 2\pi_P\left(1 - \mathbb{E}_{\boldsymbol{x}\sim p_P(\boldsymbol{x})}[s]\right) + \left(\mathbb{E}_{\boldsymbol{x}\sim p_U(\boldsymbol{x})}[s] - \pi_P\right) \\ \leq& 2\pi_P \underbrace{\left|\mathbb{E}_{\boldsymbol{x}\sim p_P(\boldsymbol{x})}[s] - 1\right|}_{R_P^{sig}} + \underbrace{\left|\mathbb{E}_{\boldsymbol{x}\sim p_U(\boldsymbol{x})}[s] - \pi_P\right|}_{R_U^{sig}}. \end{aligned} \tag{31}$$

Since $\ell^{0-1}(f(\boldsymbol{x}), y) \leq 2\ell^{sig}(f(\boldsymbol{x}), y)$, we have $R \leq 2R^{sig}$. Hence R is bounded by R_P^{sig} and R_U^{sig}:

$$R \leq 2R^{sig} \leq 4\pi_P R_P^{sig} + 2R_U^{sig}. \tag{32}$$

Note that based on Eq.(32), once we obtain the upper bound of R_P^{sig} using $\hat{R}_L$, R_U^{sig} could be bounded with $\hat{R}_U$ analogically. So in the following we mainly show how to bound R_P^{sig} with $\hat{R}_L$.

A.3. Bounding R_P^{sig} with $\hat{R}_L$

In this subsection, we follow the standard process to use **McDiarmid inequality** (Thm. 1) to bound R_P^{sig} with $\hat{R}_L$. Let ζ denote the sigmoid function, then the class of models in $\mathcal{F}$ followed by ζ is defined as $\mathcal{G} = \zeta \circ \mathcal{F}$. Namely, $g(\boldsymbol{x})$ is

s defined in Eq.(14) of the main paper. To achieve our goal, we first bound R_P^{sig} through the supremum of the difference between R_P^{sig} and $\hat{R}_U$ given any $g \in \mathcal{G}$:

$$\phi(\boldsymbol{X_L}) = \sup_{g\in\mathcal{G}}[R_P^{sig} - \hat{R}_L]. \tag{33}$$

Obviously, R_P^{sig} is bounded by the sum of $\hat{R}_L$ and $\phi(\boldsymbol{X_L})$. We then show that ϕ satisfies the **bounded difference condition** (Def.1). Let $\boldsymbol{X_L} = (\boldsymbol{x}_1, .., \boldsymbol{x}_i, .., \boldsymbol{x}_{n_L})$ and $\boldsymbol{X'_L} = (\boldsymbol{x}_1, .., \boldsymbol{x}'_i, .., \boldsymbol{x}_{n_L})$ be n_L-sample datasets with only the i-th sample being different, the following inequality holds:

$$\begin{aligned}|\phi(\boldsymbol{X_L}) - \phi(\boldsymbol{X'_L})| &\leq \left|\hat{R}_L - \hat{R}'_L\right| \\ &\overset{*}{\leq} \left|\mathbb{E}_{\boldsymbol{x}\in\boldsymbol{X_L}}[s] - \mathbb{E}_{\boldsymbol{x}'\in\boldsymbol{X'_L}}[s']\right| \\ &= \left|\frac{s_i - s'_i}{n_P}\right| \leq \frac{1}{n_L},\end{aligned} \tag{34}$$

where $*$ is based on the triangle inequality (*i.e.*, $||a| - |b|| \leq |a - b|$).

According to **McDiarmid inequality**, we then have:

$$\phi(\boldsymbol{X_L}) \leq \mathbb{E}_{\boldsymbol{X_L}}[\phi(\boldsymbol{X_L})] + \sqrt{\frac{\ln 2/\delta}{2n_L}}, \tag{35}$$

with probability at least $1 - \delta/2$. In other words, R_P^{sig} is bounded by $\hat{R}_L$ and $\mathbb{E}_{\boldsymbol{X_L}}[\phi(\boldsymbol{X_L})]$ with probability at least $1 - \delta/2$:

$$R_P^{sig} \leq \hat{R}_L + \mathbb{E}_{\boldsymbol{X_L}}[\phi(\boldsymbol{X_L})] + \sqrt{\frac{\ln 2/\delta}{2n_L}}. \tag{36}$$

Obviously, the problem turns to bound the expectations of ϕ.

A.4. Bounding with Rademacher complexity

In this subsection, we bound $\mathbb{E}_{\boldsymbol{X_L}}[\phi(\boldsymbol{X_L})]$ using **Rademacher complexity** (Def.3). Firstly, we have:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X_L}}[\phi(\boldsymbol{X_L})] \overset{*}{\leq}& \mathbb{E}_{\boldsymbol{X_L}}\left[\sup_{g\in\mathcal{G}}[\mathbb{E}_{\boldsymbol{X'_L}}\left[\hat{R}'_L\right] - \hat{R}_L]\right] \\ \leq& \mathbb{E}_{\boldsymbol{X_L}}\mathbb{E}_{\boldsymbol{X'_L}}\left[\sup_{g\in\mathcal{G}}[\hat{R}_L - \hat{R}'_L]\right] \\ \leq& \mathbb{E}_{\boldsymbol{X_L}}\mathbb{E}_{\boldsymbol{X'_L}}\left[\sup_{g\in\mathcal{G}}\left|\mathbb{E}_{\boldsymbol{x}\in\boldsymbol{X_L}}[s] - \mathbb{E}_{\boldsymbol{x}'\in\boldsymbol{X'_L}}[s']\right|\right] \\ =& \mathbb{E}_{\boldsymbol{X_L}}\mathbb{E}_{\boldsymbol{X'_L}}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{n_L}\sum_{i=1}^{n_L}(s_i - s'_i)\right|\right],\end{aligned} \tag{37}$$

where * is based on the inequality that $R_P^{sig} \leq \mathbb{E}_{\boldsymbol{X'_L}}\left[\hat{R}'_L\right]$; Then we introduce the Rademacher variables $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_i, \ldots, \sigma_{n_L})$ into the supremum in Eq.(37):

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{X_L}}\mathbb{E}_{\boldsymbol{X'_L}}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{n_L}\sum_{i=1}^{n_L}(s_i - s'_i)\right|\right] \\ =&\mathbb{E}_{\boldsymbol{X_L}}\mathbb{E}_{\boldsymbol{X'_L}}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{n_L}\sum_{i=1}^{n_L}\sigma_i(s_i - s'_i)\right|\right].\end{aligned} \tag{38}$$

Assuming that $\forall \boldsymbol{x} \in \mathcal{X}$, for any $f \in \mathcal{F}$, there exists another $\tilde{f} \in \mathcal{F}$ such that $\tilde{f}(\boldsymbol{x}) = -f(\boldsymbol{x})$, then the absolute symbol in Eq.(38) could be eliminated. By this, we further have:

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{X_L}}\mathbb{E}_{\boldsymbol{X'_L}}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{1}{n_L}\sum_{i=1}^{n_L}\sigma_i(s_i - s'_i)\right] \\ \leq& 2\mathbb{E}_{\boldsymbol{X_L}}\mathbb{E}_{\boldsymbol{X'_L}}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\frac{1}{n_L}\sum_{i=1}^{n_L}\sigma_i s_i\right] = 2\mathfrak{R}_{n_L}(\mathcal{G}).\end{aligned} \tag{39}$$

According to Eq.(3.14) in [11], with probability at least $1 - \delta/2$:

$$\mathfrak{R}_{n_L}(\mathcal{G}) \leq \hat{\mathfrak{R}}_{X_L}(\mathcal{G}) + \sqrt{\frac{\ln 2/\delta}{2n_L}}. \tag{40}$$

Therefore, $\mathbb{E}_{\boldsymbol{X_L}}[\phi(\boldsymbol{X_L})]$ can be bound by the empirical Rademacher complexity with probability at least $1 - \delta/2$:

$$\mathbb{E}_{\boldsymbol{X_L}}[\phi(\boldsymbol{X_L})] \leq 2\mathfrak{R}_{n_L}(\mathcal{G}) \leq 2\hat{\mathfrak{R}}_{X_L}(\mathcal{G}) + 2\sqrt{\frac{\ln 2/\delta}{2n_L}}. \tag{41}$$

Combined with Eq.(36), R_P^{sig} can be bounded by the empirical Rademacher complexity with probability at least $1 - \delta$:

$$R_P^{sig} \leq \hat{R}_L + 2\mathfrak{R}_{n_L}(\mathcal{G}) + 3\sqrt{\frac{\ln 2/\delta}{2n_L}}. \tag{42}$$

Similarly, we have the bound for R_U^{sig} with probability at least $1 - \delta$:

$$R_U^{sig} \leq \hat{R}_U + 2\mathfrak{R}_{n_U}(\mathcal{G}) + 3\sqrt{\frac{\ln 2/\delta}{2n_U}}. \tag{43}$$

In conclusion, by gathering Eq.(32,42,43), we finally deduce the bound of R using $\hat{R}_{lab}$ and Rademacher complexity with probability at least $1 - \delta$:

$$\begin{aligned}R \leq& 2\hat{R}_{lab} + 8\pi_P\hat{\mathfrak{R}}_{X_L}(\mathcal{G}) + 12\pi_P\sqrt{\frac{\ln 4/\delta}{2n_L}} \\ &+ 4\hat{\mathfrak{R}}_{X_U}(\mathcal{G}) + 6\sqrt{\frac{\ln 4/\delta}{2n_U}}.\end{aligned} \tag{44}$$

A.5. Bounding with VC dimension

According to **Talagrand's contraction lemma** (Lem.1), since the sigmoid loss ζ is 1-Lipschitz,

$$\begin{aligned}\hat{\mathfrak{R}}_{X_L}(\mathcal{G}) \leq \hat{\mathfrak{R}}_{X_L}(\mathcal{F}),\\ \hat{\mathfrak{R}}_{X_U}(\mathcal{G}) \leq \hat{\mathfrak{R}}_{X_U}(\mathcal{F}).\end{aligned} \tag{45}$$

By substituting $\hat{\mathfrak{R}}_{X_L}(\mathcal{G})$ and $\hat{\mathfrak{R}}_{X_U}(\mathcal{G})$ into Eq.(44), with probability at least $1-\delta$, we have:

$$\begin{aligned}R \leq & 2\hat{R}_{lab} + 8\pi_P \hat{\mathfrak{R}}_{X_L}(\mathcal{F}) + 12\pi_P \sqrt{\frac{\ln 4/\delta}{2n_L}} \\ & + 4\hat{\mathfrak{R}}_{X_U}(\mathcal{F}) + 6\sqrt{\frac{\ln 4/\delta}{2n_U}}.\end{aligned} \tag{46}$$

According to Lemma 13, 16, and 17 in [7], for a class of b-uniformly bounded functions $\mathcal{F}$ and a universal constant C, we have:

$$\hat{\mathfrak{R}}_{X_L}(\mathcal{F}) \leq C\sqrt{\frac{\mathcal{V}}{n_L}}, \tag{47}$$

$$\hat{\mathfrak{R}}_{X_U}(\mathcal{F}) \leq C\sqrt{\frac{\mathcal{V}}{n_U}}, \tag{48}$$

where $\mathcal{V}$ denotes the VC-dimension of $\mathcal{F}$.

By incorporating Eq.(47,48) into Eq.(46), with probability at least $1-\delta$, the following inequality holds:

$$\begin{aligned}R \leq & 2\hat{R}_{lab} + 8\pi_P \cdot C\sqrt{\frac{\mathcal{V}}{n_L}} + 12\pi_P \sqrt{\frac{\ln 4/\delta}{2n_L}} \\ & + 4C\sqrt{\frac{\mathcal{V}}{n_U}} + 6\sqrt{\frac{\ln 4/\delta}{2n_U}}.\end{aligned} \tag{49}$$

This completes the proof. □

B. Description of Competitors

In this section, we descibe the 10 competitive PU algorithms as follows:

- **naive** constructs the negative class with randomly sampled unlabeled data. Some underlying positives are likely to be included, thus resulting in training label noise.
- **uPU** [3] proposes a general unbiased risk estimator for PU learning.
- **nnPU** [8] improves uPU by forcing the estimated risk of negative class to be non-negative, increasing model robustness against overfitting.
- **RP** [12] treats labeled data as clean positives and unlabeled data as noisy negatives. It ranks the training data by confidence and selects the most confident samples as positives or negatives. Then traditional supervised learning could work on the chosen data.
- **PUSB** [6] focuses on selection bias during the labeling process. It aims to learn a score function that maintains the order-preserving property and proposes a threshold estimating algorithm for classification.
- **PUbN** [5] is a two-step approach that firstly estimates the class posterior probability of $\boldsymbol{x}$ to partition the data into confident positives, confident negatives, and samples unsure of their labels. It then minimizes a risk approximated by the above three partitions.
- **Self-PU** [2] incorporates a self-paced training strategy, self-calibration of a mentor-net-like manner, and self-distillation with several teacher-student networks to exert the learning capability of the deep model itself.
- **aPU** [4] deals with an arbitrary positive shift between source and target distributions.
- **VPU** [1] designs an optimization objective without class prior by introducing a variational principle.
- **ImbPU** [13] adapts the nnPU loss to enable learning from imbalanced data. It is equivalent to oversampling the minority class for the balance of the data.

C. Sensitivity Analysis

We study the impact of misspecified class prior on our model in Fig.(7). Dist-PU appears more stable than other competitors, especially when the prior is underestimated.

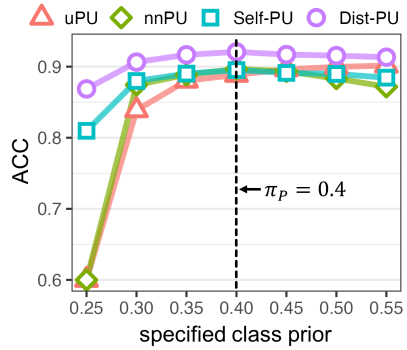

Figure 7. Sensitivity to prior (CIFAR-10).

D. Stability Analysis

A model shows the ability to rectify the negative-prediction preference if it tends to maintain a relatively good Precision-Recall balance during the training since the number of predicted positives will be around that of ground-truth ones when the predicted prior is close to π_P. Such a trend can be captured exactly by our $\hat{R}_{lab}$ in Fig.8. By contrast, other models are with increasing Prec and decreasing Rec because their negative-prediction preference causes a reduction in predicted positives. Besides, all metrics of our $\hat{R}_{lab}$ are more stable.

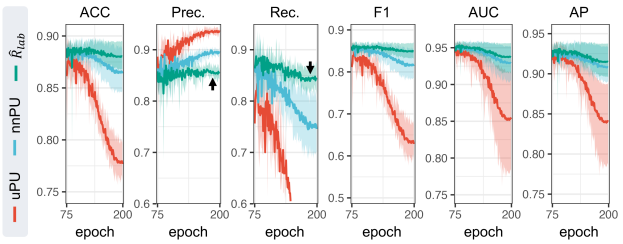

Figure 8. Stability analysis on CIFAR-10.

E. Implementation Details of Fig.1.

The batch size is 256. We use Adam as the optimizer with a cosine annealing scheduler, where the initial learning rate is set as 5×10^{-4}; weight decay is set as 5×10^{-3}.

F. Comparison with uPU and nnPU

Our essential differences are two-fold: **a)** To adopt the class prior as learning supervision, we need to reformulate the 0-1 risk with L1 distances between $\mathbb{E}[\hat{y}]$ and $\mathbb{E}[y]$ over $\boldsymbol{X_P}$ and $\boldsymbol{X_N}$. In this way, alignments based on class priors are established, with $\mathbb{E}[\hat{y}]$ being the predicted prior and $\mathbb{E}[y]$ being the ground-truth one. **b)** Without such a translation in **a)**, uPU[3] and nnPU[8] lack an absolute function on $\mathbb{E}_{\boldsymbol{x}_U}[\hat{y}] - \pi_P$, directly leading to the negative-prediction preference. In contrast, our formulation pursues the label distribution consistency on unlabeled data by $|\mathbb{E}_{\boldsymbol{x}_U}[\hat{y}] - \pi_P|$, and thus rectifies the negative-prediction preference shown in Fig.(8). These key points essentially set our method apart from the other two.

References

[1] Hui Chen, Fangqing Liu, Yin Wang, Liyue Zhao, and Hao Wu. A variational approach for learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems*, 2020. 3

[2] Xuxi Chen, Wuyang Chen, Tianlong Chen, Ye Yuan, Chen Gong, Kewei Chen, and Zhangyang Wang. Self-pu: Self boosted and calibrated positive-unlabeled training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 1510–1519, 2020. 3

[3] Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems*, pages 703–711, 2014. 3, 4

[4] Zayd Hammoudeh and Daniel Lowd. Learning from positive and unlabeled data with arbitrary positive shift. In *Advances in Neural Information Processing Systems*, 2020. 3

[5] Yu-Guan Hsieh, Gang Niu, and Masashi Sugiyama. Classification from positive, unlabeled and biased negative data. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 2820–2829, 2019. 3

[6] Masahiro Kato, Takeshi Teshima, and Junya Honda. Learning from positive and unlabeled data with a selection bias. In *Preceedings of the 7th International Conference on Learning Representations*, 2019. 3

[7] Justin Khim, Liu Leqi, Adarsh Prasad, and Pradeep Ravikumar. Uniform convergence of rank-weighted learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 5254–5263, 2020. 3

[8] Ryuichi Kiryo, Gang Niu, Marthinus Christoffel du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems*, pages 1675–1685, 2017. 3, 4

[9] Colin McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998. 1

[10] Mehryar Mohri and Andres Munoz Medina. Learning theory and algorithms for revenue optimization in second price auctions with reserve. In *International Conference on Machine Learning*, pages 262–270. PMLR, 2014. 1

[11] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT Press, 2018. 1, 2

[12] Curtis G. Northcutt, Tailin Wu, and Isaac L. Chuang. Learning with confident examples: Rank pruning for robust classification with noisy labels. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, 2017. 3

[13] Guangxin Su, Weitong Chen, and Miao Xu. Positive-unlabeled learning from imbalanced data. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 2995–3001, 2021. 3