Supplementary Material

Yuanzhuo Xu[1], Xiaoguang Niu[1,4*], Jie Yang[1], Steve Drew[2], Jiayu Zhou[3], Ruizhi Chen[4]

[1]School of Computer Science, Wuhan University, China
[2] Department of Electrical and Software Engineering, University of Calgary, Canada
[3] Department of Computer Science and Engineering, Michigan State University, USA
[4] LIESMARS, Wuhan University, China

Additional Experiment Details

All methods for comparison in the experiments use preset hyperparameters in their literature and repositories. All networks are trained on *MNIST* for 100 epochs and *CIFAR* for 200 epochs. The total number of training epochs on Clothing1M varies and is given according to the literature and repositories. Our algorithm only adds the dropout layer with drop rate 0.25, and other setting remains the same as JoCoR. (Wei et al. 2020).

We construct the CCN, OOD, and IDN datasets as follow.

CCN dataset

We adopt the same scheme as other literature (Wei et al. 2020; Patrini et al. 2017) to construct two types of noise: (1) Symmetric noise (Van Rooyen, Menon, and Williamson 2015); (2) Asymmetric noise (Patrini et al. 2017), which only flips labels for potentially misclassified classes. In order to test the performance of different approaches under low, medium and high noise rates, we constructed a noise rate of [0.2,0.5,0.8] for symmetric noise, and [0.2, 0.45] for asymmetric noise.

OOD dataset

We follow the method used in (Yao et al. 2021) to introduce 20% OOD samples. We construct the OOD dataset with symmetric and pairflip noise, with the range of [0.2,0.5,0.8] for symmetric noise and [0.2, 0.45] for pairflip noise.

IDN dataset

We follow the Algorithm 2 in (Xia et al. 2020) to add instance-dependent-noise to the training set manually for experiment. Label corruption is only performed on easily mislabeled samples with similar feature. We then construct under four noise rate in the range of [0.2,0.3,0.4,0.5].

Reproducibility

All network settings and hyperparameters are listed in previous experiment section. Additionally, we submit the *USDNL* algorithm and dataset generation code in the supplementary material to better demonstrate reproducibility and technical details.

Theory Proof Details

In this section, we provide the detailed proof of Lemma 1, Theorem 1, Theorem 2, and Proposition 1 in main paper.

Proof of Lemma 1

For the closed dataset training, once the linear model well-fit the clean samples in training set, i.e., $E_t\left(\mathcal{P}^{\hat{w}_t}\left(y_c \mid x_c\right)\right) \leq \epsilon$, for the fixed learned clean sets x_c, y_c, we have:

$$E_{t_1,t_2}\left(\left\|\mathcal{P}^{\hat{w}_{t_1}}\left(y_c \mid x_c\right)-\mathcal{P}^{\hat{w}_{t_2}}\left(y_c \mid x_c\right)\right\|\right) \leq c_1\epsilon \tag{1}$$

where c_1 is a finite constant.

Proof: In the classification task, we adopt the cross-entropy loss as the empirical loss, i.e., $E_t\left(\mathcal{P}^{\hat{w}_t}\left(y_c \mid x_c\right)\right) = E_t\left(-\mathbf{y}\log\mathcal{P}^{\hat{w}_t}\left(y_c \mid x_c\right)\right) \leq \epsilon$. Note that the observable $\mathbf{y}$ is the one-hot vector with binary value, we have:

$$\begin{aligned}&E_i\left(KL\left(\mathbf{y}\|\mathcal{P}^{\hat{w}_i}\left(y_c \mid x_c\right)\right)\right)\\&=E_i\left(\mathbf{y}\log\mathbf{y}-\mathbf{y}\log\mathcal{P}^{\hat{w}_i}\left(y_c \mid x_c\right)\right) \leq \epsilon\end{aligned} \tag{2}$$

According to the relation between the KL-divergence and the total variation distance, we have:

$$E_t\left(\left\|\mathbf{y}-\mathcal{P}^{\hat{w}_t}\left(y_c \mid x_c\right)\right\|_1\right) \leq \sqrt{2}\epsilon \tag{3}$$

Due to the arbitrariness of i, we finally have:

$$\begin{aligned}&E_{t_1,t_2}\left(\left\|\mathcal{P}^{\hat{w}_{t_1}}\left(y_c \mid x_c\right)-\mathcal{P}^{\hat{w}_{t_2}}\left(y_c \mid x_c\right)\right\|_1\right)\\&=E_c\left(\left\|\mathcal{P}^{\hat{w}_{t_1}}\left(y_c \mid x_c\right)-\mathbf{y}+\mathbf{y}-\mathcal{P}^{\hat{w}_{t_2}}\left(y_c \mid x_c\right)\right\|_1\right)\\&\leq E_c\left(\left\|\mathcal{P}^{\hat{w}_{t_1}}\left(y_c \mid x_c\right)-\mathbf{y}\right\|_1+\left\|\mathbf{y}-\mathcal{P}^{\hat{w}_{t_2}}\left(y_c \mid x_c\right)\right\|_1\right)\\&\leq 2\sqrt{2}\epsilon\end{aligned} \tag{4}$$

Let $c_1 = 2\sqrt{2}$, and we can get the result of Lemma 1.

Proof of Theorem 1

Define a linear model $\mathcal{P}^w(y|x)$ and its sub-model $\mathcal{P}^{\hat{w}}(y|x)$ with w and Bernoulli sampling $\hat{w}$, respectively. We denote $\mathcal{H}$ as the truncated entropy loss function of the sub-model with $\hat{w}$. For fixed learned clean sets x_c, y_c, we have:

$$\begin{aligned}&\mathbb{E}_c\left(\left|\mathcal{H}\left(\mathbb{E}_l\left(\mathcal{P}^{\hat{w}_l}\left(y_c \mid x_c\right)\right)\right)-\mathcal{H}\left(\mathbb{E}_k\left(\mathcal{P}^{\hat{w}_k}\left(y_c \mid x_c\right)\right)\right)\right|\right)\\&\leq c_2\epsilon,\end{aligned} \tag{5}$$

[*] The corresponding author.

where l is a finite integer (at least 1), and c_2 is the Lipschitz constant satisfying the empirical loss function.

Proof: Here, the truncated entropy loss means that we perform minima threshold truncation on the values of each dimension of the probability vector before computing the entropy, which does not affect the prediction results, but allows the entropy loss to satisfy the Lipschitz condition. We have:

$$\begin{aligned} &\mathbb{E}_c\left(\left|\mathcal{H}\left(\mathbb{E}_l\left(\mathcal{P}^{\hat{w}_l}\left(y_c \mid x_c\right)\right)\right)-\mathcal{H}\left(\mathbb{E}_k\left(\mathcal{P}^{\hat{w}_k}\left(y_c \mid x_c\right)\right)\right)\right|\right) \\ &\leq c_{l_1} \mathbb{E}_c\left(\|\mathbb{E}_l\left(\mathcal{P}^{\hat{w}_l}\left(y_c \mid x_c\right)\right)-\mathbb{E}_k\left(\mathcal{P}^{\hat{w}_k}\left(y_c \mid x_c\right)\right)\|_1\right), \end{aligned} \tag{6}$$

where c_{l_1} is the Lipschitz constant. Furthermore, we have:

$$\begin{aligned} &\|\mathbb{E}_l\left(\mathcal{P}^{\hat{w}_l}\left(y_c \mid x_c\right)\right)-\mathbb{E}_k\left(\mathcal{P}^{\hat{w}_k}\left(y_c \mid x_c\right)\right)\|_1 \\ &=|\frac{1}{l} \sum_{i=1}^{l} \mathcal{P}^{\hat{w}_i}\left(y_c \mid x_c\right)-\frac{1}{k} \sum_{j=1}^{k} \mathcal{P}^{\hat{w}_j}\left(y_c \mid x_c\right)| \\ &=\mathbb{E}_{i,j}\left(\left|\mathcal{P}^{\hat{w}_i}\left(y_c \mid x_c\right)-\mathcal{P}^{\hat{w}_j}\left(y_c \mid x_c\right)\right|\right). \end{aligned} \tag{7}$$

Combined with the Lemma 1, we have:

$$\begin{aligned} &\mathbb{E}_c\left(\left|\mathcal{H}\left(\mathbb{E}_l\left(\mathcal{P}^{\hat{w}_l}\left(y_c \mid x_c\right)\right)\right)-\mathcal{H}\left(\mathbb{E}_k\left(\mathcal{P}^{\hat{w}_k}\left(y_c \mid x_c\right)\right)\right)\right|\right) \\ &\leq 2\sqrt{2} c_{l_1} \epsilon \end{aligned} \tag{8}$$

Let $c_2 = 2\sqrt{2} c_{l_1} \epsilon$, the result is proved.

Proof of Theorem 2

(Finite dropout on cross-entropy loss) (Finite dropouts on empirical loss) *Define a linear model $\mathcal{P}^{w}(y|x)$ and its sub-model $\mathcal{P}^{\hat{w}}(y|x)$ with w and Bernoulli sampling $\hat{w}$, respectively. We denote $\mathcal{L}(\mathcal{P}(\tilde{y}|x)) = -\log(\mathcal{P}(\tilde{y}|x))$ as the cross-entropy loss function of the sub-model with dropout applied on w. For the fixed learned clean sets $\{x_c, y_c\}$, we have*

$$\begin{aligned} &\mathbb{E}_c\left(\left|\mathbb{E}_l\left(\mathcal{L}\left(\mathcal{P}^{\hat{w}_l}\left(\tilde{y}_c \mid x_c\right)\right)\right)-\mathbb{E}_k\left(\mathcal{L}\left(\mathcal{P}^{\hat{w}_k}\left(\tilde{y}_c \mid x_c\right)\right)\right)\right|\right) \\ &\leq c_3 \epsilon, \end{aligned} \tag{9}$$

where l is a finite integer (at least 1), and c_3 is the Lipschitz constant satisfying the empirical loss function.

Proof: With the same truncated probability vector, we have:

$$\begin{aligned} &\left|\mathbb{E}_l\left(\mathcal{L}\left(\mathcal{P}^{\hat{w}_l}\left(\tilde{y}_c \mid x_c\right)\right)\right)-\mathbb{E}_k\left(\mathcal{L}\left(\mathcal{P}^{\hat{w}_k}\left(\tilde{y}_c \mid x_c\right)\right)\right)\right| \\ &=|\frac{1}{l} \sum_{i=1}^{l} \tilde{y}_c \log \mathcal{P}^{\hat{w}_i}\left(y_c \mid x_c\right)-\frac{1}{k} \sum_{j=1}^{k} \tilde{y}_c \log \mathcal{P}^{\hat{w}_j}\left(y_c \mid x_c\right)| \\ &\leq |\frac{1}{lk} \sum_{i,j}\left(\log \mathcal{P}^{\hat{w}_i}\left(y_c \mid x_c\right)-\log \mathcal{P}^{\hat{w}_j}\left(y_c \mid x_c\right)\right)| \\ &\leq \frac{c_{l_2}}{lk} \sum_{i,j}\left|\mathcal{P}^{\hat{w}_i}\left(y_c \mid x_c\right)-\mathcal{P}^{\hat{w}_j}\left(y_c \mid x_c\right)\right| \end{aligned} \tag{10}$$

Combined with the Lemma 1, we have:

$$\begin{aligned} &\mathbb{E}_c\left(\left|\mathbb{E}_l\left(\mathcal{L}\left(\mathcal{P}^{\hat{w}_l}\left(y_c \mid x_c\right)\right)\right)-\mathbb{E}_k\left(\mathcal{L}\left(\mathcal{P}^{\hat{w}_k}\left(y_c \mid x_c\right)\right)\right)\right|\right) \\ &\leq 2\sqrt{2} c_{l_2} \epsilon \end{aligned} \tag{11}$$

where c_{l_2} is the Lipischitz constant. Let $c_3 = 2\sqrt{2} c_{l_2}$, we then get the result.

Analysis of Proposition 1

In a selection task of clean samples $(x_c, \hat{y}_c)$, for a well-trained linear model $\mathcal{P}^{w}(y_c|x_c)$, the empirical loss $\mathcal{L}\left(\mathcal{P}^{\hat{w}}\left(\tilde{y_c} \mid x_c\right)\right)$ and the epistemic uncertainty $H(\mathbf{p})$ with single dropout sampling are positively correlated on clean label samples in the learned distribution space.

Proof: In the task of clean samples selection, the low uncertainty is a necessary but not sufficient condition for a correct annotation of a sample (i.e., clean label). The uncertainty $H(\mathbf{p}) = -\sum_{c=1}^{C} p_c \log p_c$ is low if and only if the probability p_l of any class l is high, but the sample is not selected if its label does not match the prediction class l. The prerequisite for the sample to be selected is that its predicted class is consistent with the label, i.e., the cross-entropy loss between the prediction and the label is small.

Once the prediction is consistent with the label, if the probability p_c is higher, the epistemic uncertainty and cross-entropy loss will then be both lower, and vice versa. This is because they exhibit the same monotonicity when p_c fluctuates. Therefore, we can state that the epistemic uncertainty and cross-entropy are positively correlated in the selection of clean samples. The positive correlation between them justifies the use of single cross-entropy instead of weighted sum.

Additional Experimental Results

We supplement the result of the test accuracy and label precision not shown in the main paper. The results include the comparison between *USDNL* and other state-of-art methods on CCN-CIFAR-100 (Fig. 1), CIFAR80N-O (Fig. 2), IDN-CIFAR-10 (Fig. 3) and IDN-CIFAR-100 (Fig. 4) at various noise rates. The results show that we have optimal performance at almost the full noise rate setting.

References

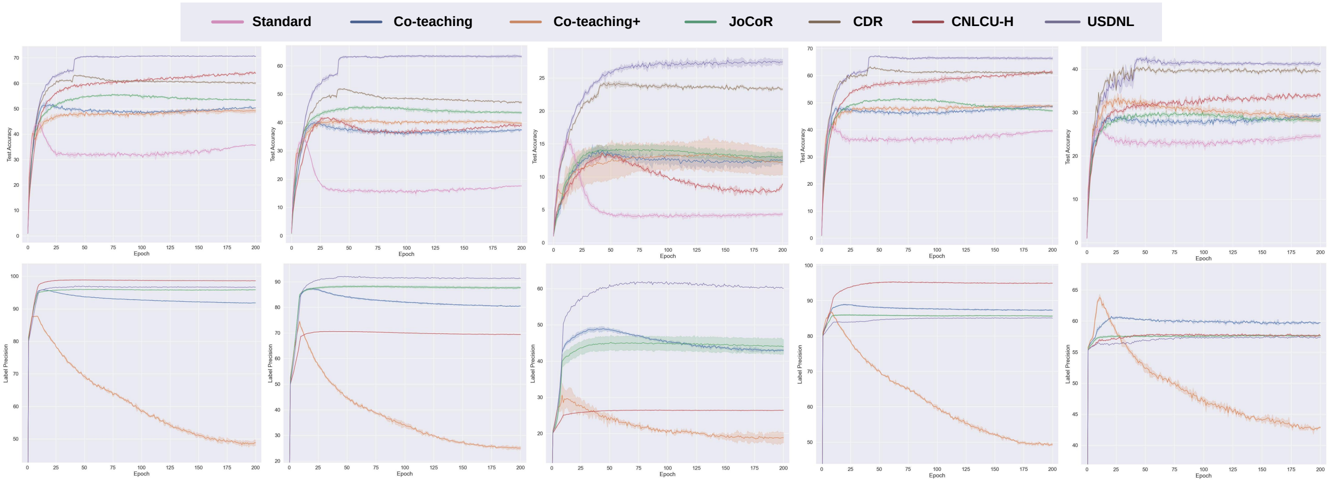
Patrini, G.; Rozza, A.; Krishna Menon, A.; Nock, R.; and Qu, L. 2017. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1944–1952.

Van Rooyen, B.; Menon, A.; and Williamson, R. C. 2015. Learning with symmetric label noise: The importance of being unhinged. *Advances in neural information processing systems*, 28.

Wei, H.; Feng, L.; Chen, X.; and An, B. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13726–13735.

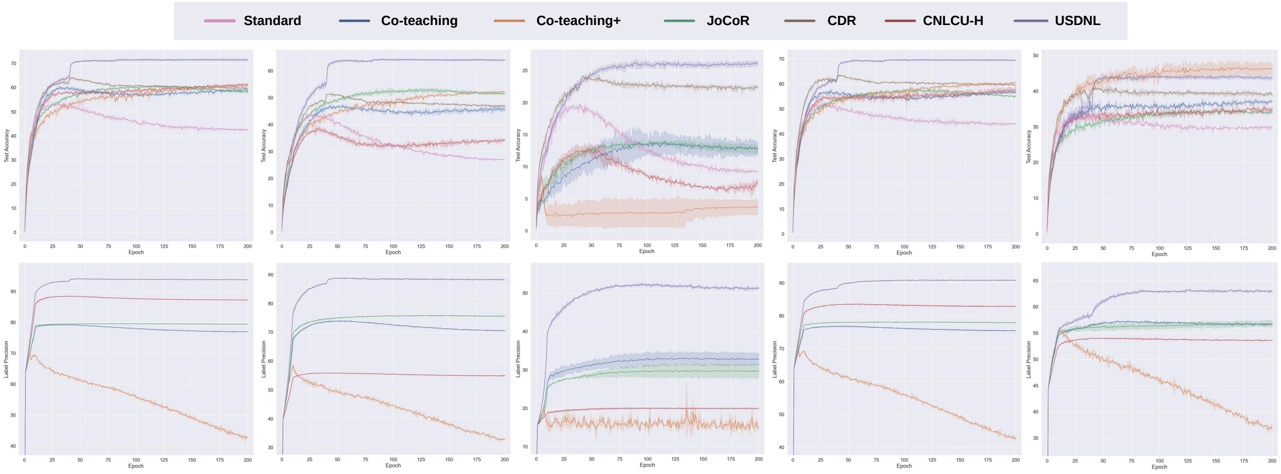
Xia, X.; Liu, T.; Han, B.; Wang, N.; Gong, M.; Liu, H.; Niu, G.; Tao, D.; and Sugiyama, M. 2020. Part-dependent label noise: Towards instance-dependent label noise. *Advances in Neural Information Processing Systems*, 33: 7597–7610.

Yao, Y.; Sun, Z.; Zhang, C.; Shen, F.; Wu, Q.; Zhang, J.; and Tang, Z. 2021. Jo-src: A contrastive approach for combating noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5192–5201.



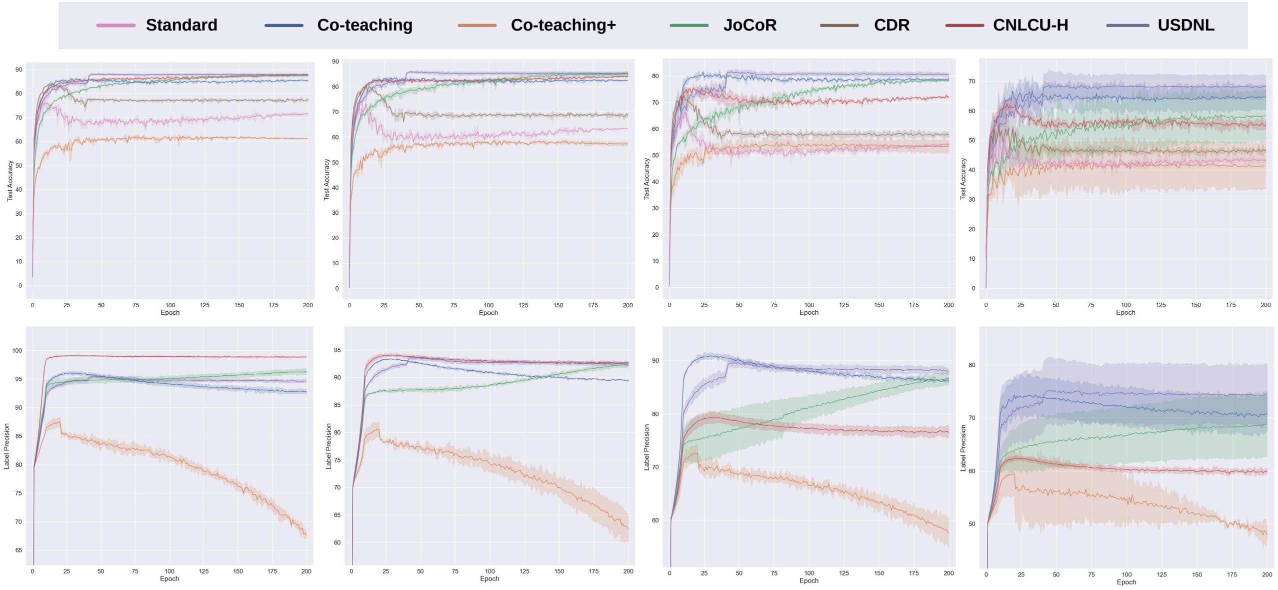
(a) Symmetry-20% (b) Symmetry-50% (c) Symmetry-80% (d) Asymmetry-20% (e) Asymmetry-45%

Figure 1: CCN results on CIFAR-100 dataset. Top: test accuracy(%) v.s. epochs; bottom: label precision(%) v.s. epochs.



(a) Symmetry-20% (b) Symmetry-50% (c) Symmetry-80% (d) Asymmetry-20% (e) Asymmetry-45%

Figure 2: OOD results on CIFAR80N-O dataset. Top: test accuracy(%) v.s. epochs; bottom: label precision(%) v.s. epochs.



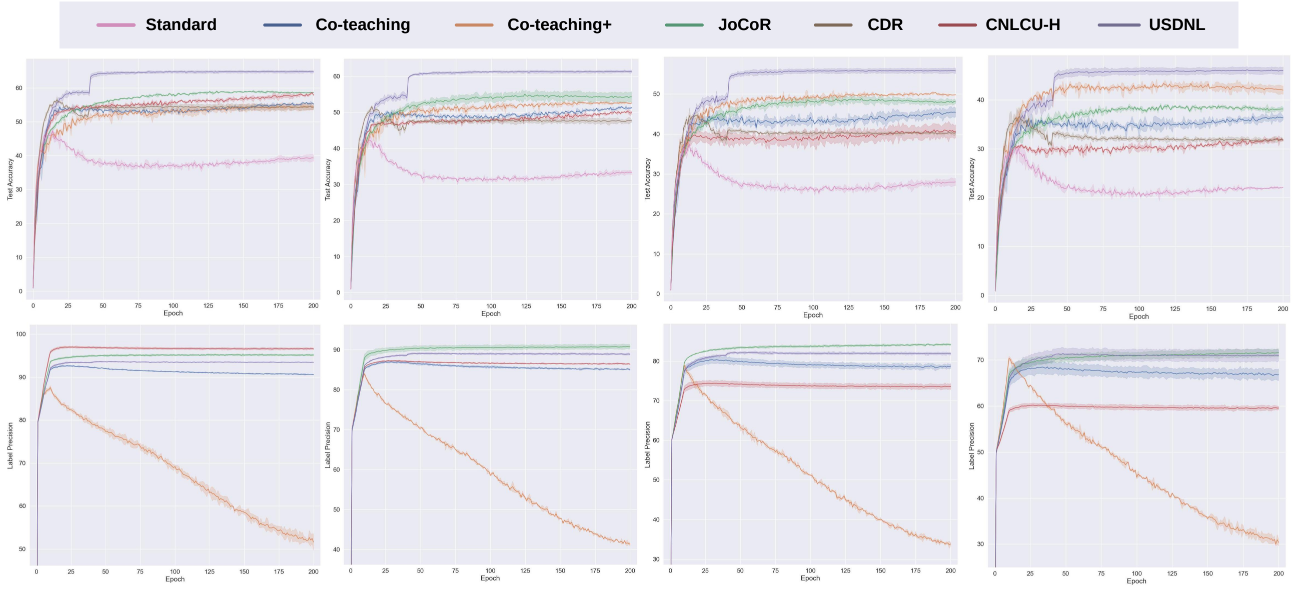
(a) Instance-20%

(b) Instance-30%

(c) Instance-40%

(d) Instance-50%

Figure 3: IDN resultd on CIFAR-10 dataset. Top: test accuracy(%) v.s. epochs; bottom: label precision(%) v.s. epochs.



(a) Instance-20%

(b) Instance-30%

(c) Instance-40%

(d) Instance-50%

Figure 4: IDN results on CIFAR-100 dataset. Top: test accuracy(%) v.s. epochs; bottom: label precision(%) v.s. epochs.