Probabilistic Instance-dependent Label Refinement for Noisy Label Learning

Hao-Yuan He, Yu Liu, Ren-Biao Liu, Zheng Xie and Ming Li

Learning from Noisy Examples

For dataset $\{(\boldsymbol{x}_i, \tilde{\boldsymbol{y}}_i)\}_{i=1}^N$, the given label $\tilde{\boldsymbol{y}}$ could be noisy.

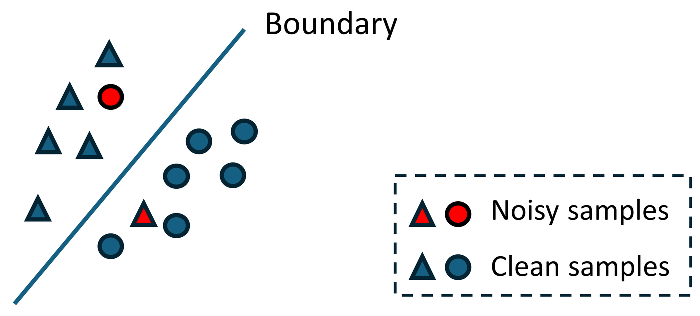

Figure 1. Classification with noisy supervision.

The previous researchers assume that the noise is class-dependent, i.e., there exists a matrix T that models the transition between noisy and clean labels [1]. However, this is not realistic in the real world.

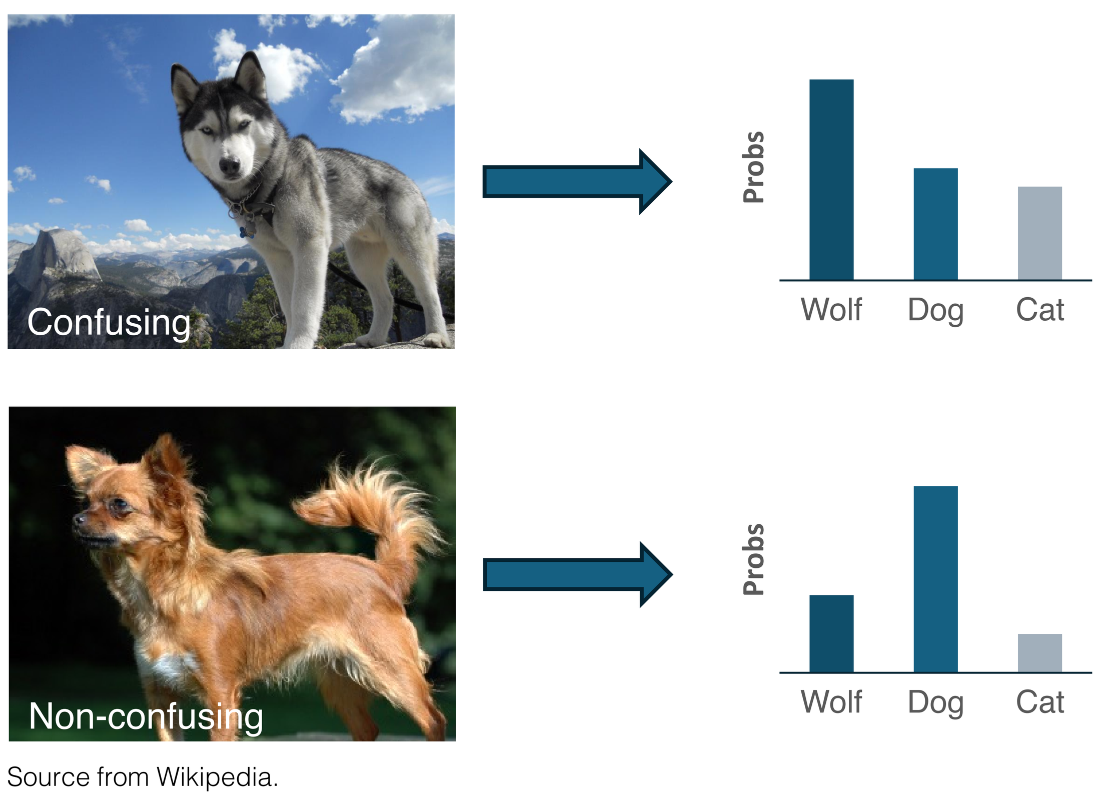

Figure 2. **Realistic label noise**. *Confusing* instances are more likely to be misclassified.

Main idea

We model the refined label $\boldsymbol{q}_i$ as:

$$\boldsymbol{q}_i = \boldsymbol{v}_i \cdot (\eta_i \cdot \hat{\boldsymbol{y}}_i + (1 - \eta_i) \cdot \tilde{\boldsymbol{y}}_i), \tag{1}$$

where $\hat{\boldsymbol{y}}_i$ and $\tilde{\boldsymbol{y}}_i$ is the model's prediction and the noisy label, $\boldsymbol{v}_i \in \mathbb{R}^c$ is instance transition ratio which reflects the shift of class distribution of label noise, and η_i is the confusing probability of the i-th instance.

Instance-dependent Noise (IDN) Modeling

Setup. Consider a dataset of training samples $\{(\boldsymbol{x}_i, \tilde{\boldsymbol{y}}_i)\}_{i=1}^N$, where each sample is assoicated with a true label $\boldsymbol{y}_i$. The label space is $\{z^\top z = 1 \mid z \in \{0,1\}^c\}$.

Modeling. The estimated true label $\boldsymbol{q}_i = [\Pr(\boldsymbol{y}_i^1 = 1 \mid \boldsymbol{x}_i), ..., \Pr(\boldsymbol{y}_i^c = 1 \mid \boldsymbol{x}_i)]^\top$. We first consider $\boldsymbol{q}_i^j = \Pr(\boldsymbol{y}_i^j = 1 \mid \boldsymbol{x}_i)$, by Bayes formula:

$$\Pr\left(\boldsymbol{y}_i^j = 1 \,\middle|\, \boldsymbol{x}_i\right) = \frac{\Pr\left(\tilde{\boldsymbol{y}}_i, \boldsymbol{y}_i^j = 1 \,\middle|\, \boldsymbol{x}_i\right)}{\Pr\left(\tilde{\boldsymbol{y}}_i \,\middle|\, \boldsymbol{y}_i^j = 1, \boldsymbol{x}_i\right)} = \underbrace{\Pr(\tilde{\boldsymbol{y}}_i \mid \boldsymbol{x}_i)}_{\text{denote as } \psi_i} \cdot \frac{\Pr\left(\tilde{\boldsymbol{y}}_i^j = 1 \,\middle|\, \boldsymbol{y}_i, \boldsymbol{x}_i\right)}{\Pr\left(\tilde{\boldsymbol{y}}_i \,\middle|\, \boldsymbol{y}_i^j = 1, \boldsymbol{x}_i\right)}. \tag{2}$$

Using the concept of confusing probability, expand $\Pr\left(\tilde{\boldsymbol{y}}_i^j = 1 \,\middle|\, \boldsymbol{y}_i, \boldsymbol{x}_i\right)$ as follows:

$$\Pr\left(\boldsymbol{y}_i^j = 1 \,\middle|\, s_i = 0, \tilde{\boldsymbol{y}}_i, \boldsymbol{x}_i\right) \cdot (1 - \eta_i) + \Pr\left(\boldsymbol{y}_i^j = 1 \,\middle|\, s_i = 1, \tilde{\boldsymbol{y}}_i, \boldsymbol{x}_i\right) \cdot \eta_i. \tag{3}$$

The first term refers to the case that the sample $\boldsymbol{x}_i$ is not confusing, which euqals $\mathbb{I}\left(\tilde{\boldsymbol{y}}_i^j = \boldsymbol{y}_i^j\right) = \tilde{\boldsymbol{y}}_i^j$. The second term can be represented as the model's prediction $\hat{\boldsymbol{y}}_i^j$. Combine the above equations, and use the notation $\boldsymbol{v}_i^j$ to represent the ratio between $\Pr(\tilde{\boldsymbol{y}} \mid \boldsymbol{x}_i)$ and $\Pr\left(\tilde{\boldsymbol{y}}_i \,\middle|\, \boldsymbol{y}_i^j = 1, \boldsymbol{x}_i\right)$, we finally get (1).

Loss terms. The loss terms are consist of three parts:

- Classification loss with refined label:

$$\mathcal{L}_c = \frac{1}{N} \sum_{i \in [N]} \mathrm{CrossEntropy}(\boldsymbol{q}_i, \hat{\boldsymbol{y}}_i). \tag{4}$$

- Expectation-maximization(EM) for updating $\boldsymbol{v}_i$:

$$\mathcal{L}_v = -\frac{1}{N \cdot c} \sum_{i \in [N]} \sum_{j \in [c]} \boldsymbol{q}_i^j \log(\psi_i \cdot [\eta_i \cdot \hat{\boldsymbol{y}}_i + (1 - \eta_i) \cdot \tilde{\boldsymbol{y}}_i]). \tag{5}$$

- Regularization terms, e.g., ELR loss [2]:

$$\mathcal{L}_r = \frac{1}{N} \sum_{i \in [N]} \log(1 - \langle \hat{\boldsymbol{y}}_i, \boldsymbol{t}_i \rangle). \tag{6}$$

Highlights

- Estimating the confusing probability helps modeling IDN label noise; π-LR assigns a probability η_i to each instance, showing how IDN affects true labels.
- π-LR shows robustness against both realistic and synthetic label noise, while remaining efficient in time and space.

π-LR: Probabilistic Instance-dependent Label Refinement

Input: Training set $\{(\boldsymbol{x}_i, \tilde{\boldsymbol{y}}_i)\}_{i=1}^N$; training steps T; estimation step list $\mathcal{T}$.
Output: Optimized parameters θ
Initialize $\eta_i = 0, \forall i \in [N]$
Initialize $\boldsymbol{v}_i = \boldsymbol{1}, \forall i \in [N]$
For $t = 1$ to T **do**
  Estimate the true label as $\boldsymbol{q}_i = \boldsymbol{v}_i \cdot (\eta_i \cdot \hat{\boldsymbol{y}}_i + (1 - \eta_i) \cdot \tilde{\boldsymbol{y}}_i)$.
  Calculate the loss terms, ref (4) (5) and (6).
  Update θ.
  If $t \in \mathcal{T}$ **then**
    Estimate $\eta_i, \forall i \in [N]$
  End if
End for

Algorithm 1. Overall algorithm of π-LR.

Estimation of Confusing Probabilities

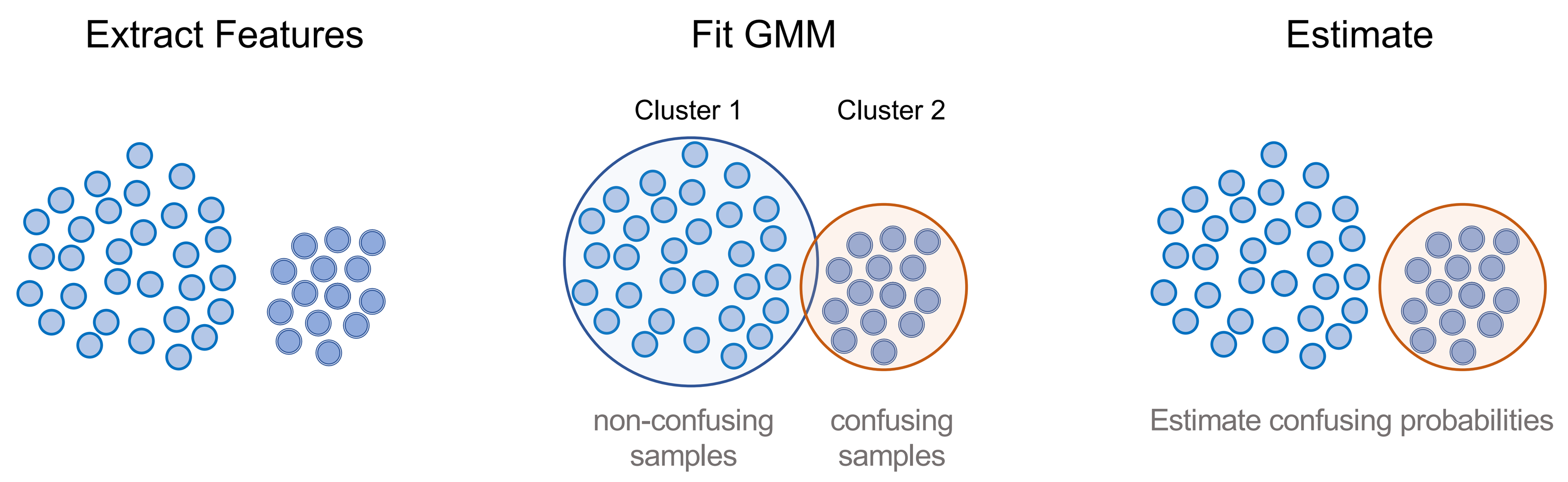

Figure 3. Estimation of confusing probabilities. Non-confusing samples are typically closer to the class center. Thus, we can use Gaussian mixture models for estimation.

Empirical Studies

Methods	Random 1	Random 2	Random 3	Aggregate	Worst	Noisy
CE	85.02±0.65	86.46±1.79	85.16±0.61	87.77±0.38	77.69±1.55	55.50±0.66
Forward	86.88±0.50	86.14±0.24	87.04±0.35	88.24±0.22	79.79±0.46	57.01±1.03
Backward	87.14±0.34	86.28±0.80	86.86±0.41	88.13±0.29	77.61±1.05	57.14±0.92
GCE	87.61±0.28	87.70±0.56	87.58±0.29	87.85±0.70	80.66±0.35	56.73±0.30
Peer Loss	89.06±0.11	88.76±0.19	88.57±0.09	90.75±0.25	82.53±0.52	57.59±0.61
VolMinNet	88.30±0.12	88.27±0.09	88.19±0.41	89.70±0.21	80.53±0.20	57.80±0.31
F-div	89.70±0.40	89.79±0.12	89.55±0.49	91.64±0.34	82.53±0.52	57.10±0.65
ELR	91.46±0.38	91.61±0.16	91.41±0.44	92.38±0.64	83.58±1.13	58.94±0.92
PM	85.12±2.90	87.55±0.13	84.83±3.18	88.78±0.95	84.83±1.13	29.87±0.08
CAL	90.93±0.31	90.75±0.30	90.74±0.24	91.97±0.32	85.36±0.16	61.73±0.42
CORES	89.66±0.32	89.91±0.45	89.79±0.50	91.23±0.11	83.60±0.53	61.15±0.73
π-LR (Ours)	92.02±0.32	91.96±0.28	92.09±0.12	92.99±0.24	86.76±0.42	62.73±0.46

Table 1. Experiments on CIFAR-N, comparision with SOTAs.

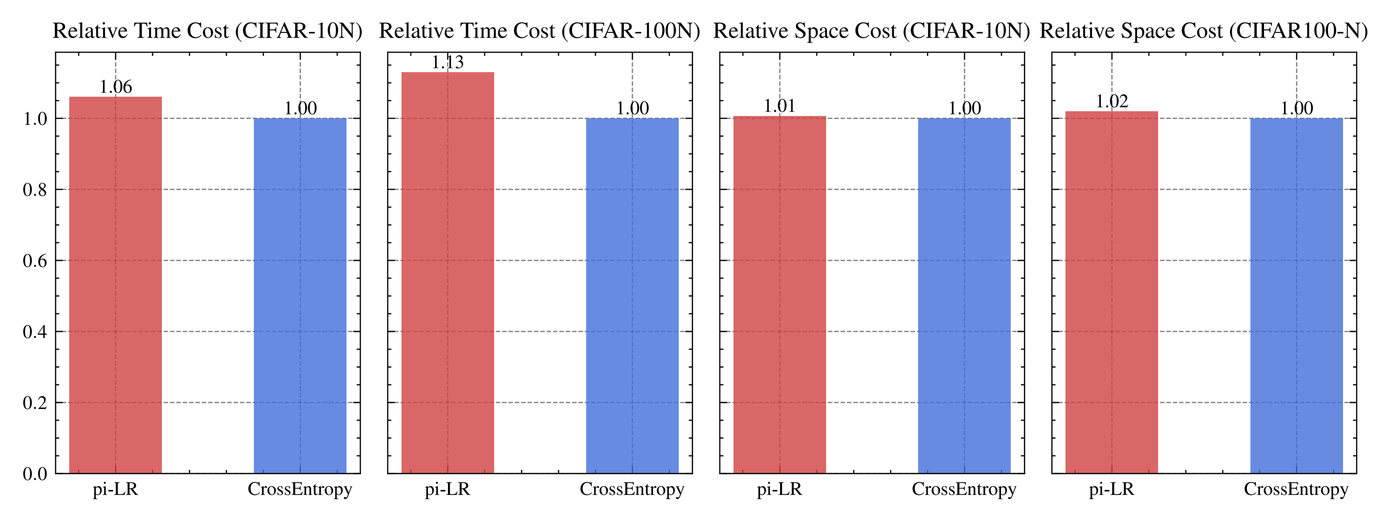

Figure 4. Efficiency analysis: π-LR has low time and space complexity.

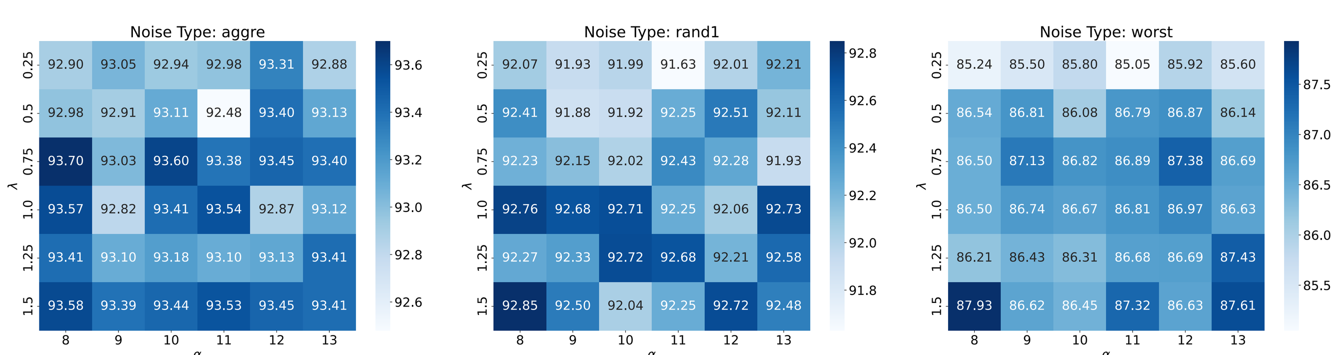

Figure 5. Sensitivity analysis: π-LR maintains good performance under various settings.

[1] G. Patrini, A. Rozza, A. K. Menon, R. Nock, and L. Qu, "Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2233–2241.

[2] S. Liu, J. Niles-Weed, N. Razavian, and C. Fernandez-Granda, "Early-Learning Regularization Prevents Memorization of Noisy Labels," in *Advances in Neural Information Processing Systems 33*, 2020, pp. 20331–20342.

Contact: hehy@lamda.nju.edu.cn