Noisy Label Learning with Instance-Dependent Outliers: Identifiability via Crowd Wisdom

Tri Nguyen[1] Shahana Ibrahim[2] Xiao Fu[1]

[1]School of EECS
Oregon State University

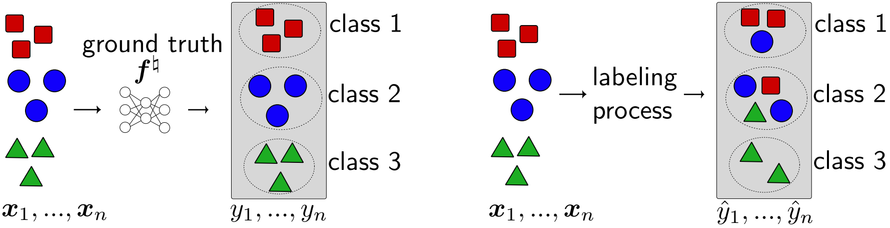
[2]Department of ECE
University of Central Florida

NeurIPS 2024

Oregon State University

UCF

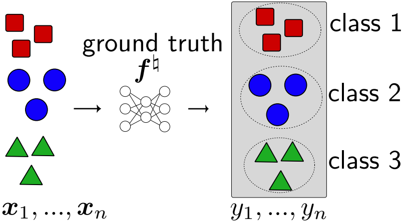
Noisy Label Learning Problem



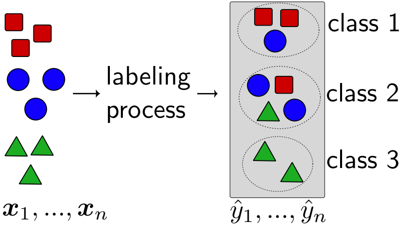
(a) Clean label (b) Noisy label

Figure: Source: internet/chatgpt.

Noisy Label Learning Problem



(a) Clean label



(b) Noisy label

Figure: Source: internet/chatgpt.

Goal: Recover the ground truth (GT) classifier $\boldsymbol{f}^\natural$ given $(\boldsymbol{x}_1, \widehat{y}_1), \ldots, (\boldsymbol{x}_N, \widehat{y}_N)$.

Noise Generation Modeling Approach

- Among approaches including noisy label filtering [1–4], robust noise design [5–9], the **noise generation modeling** [10–18] approach is the most popular.

Noise Generation Modeling Approach

- Among approaches including noisy label filtering [1–4], robust noise design [5–9], the **noise generation modeling** [10–18] approach is the most popular.
- Let $\boldsymbol{x}, y, \hat{y}$ be the sample's feature, its true label, and the obtained noisy label,

$$\underbrace{\begin{bmatrix} \Pr\left(\widehat{y}=1 \mid \boldsymbol{x}\right) \\ \dots \\ \Pr\left(\widehat{y}=K \mid \boldsymbol{x}\right) \end{bmatrix}}_{\boldsymbol{g}^{\natural}(\boldsymbol{x})} = \underbrace{\begin{bmatrix} \Pr\left(\widehat{y}=1 \mid y=1, \boldsymbol{x}\right) & \dots & \Pr\left(\widehat{y}=1 \mid y=K, \boldsymbol{x}\right) \\ \dots & \dots & \dots \\ \Pr\left(\widehat{y}=K \mid y=1, \boldsymbol{x}\right) & \dots & \Pr\left(\widehat{y}=K \mid y=K, \boldsymbol{x}\right) \end{bmatrix}}_{\boldsymbol{T}^{\natural}(\boldsymbol{x})} \underbrace{\begin{bmatrix} \Pr\left(y=1 \mid \boldsymbol{x}\right) \\ \dots \\ \Pr\left(y=K \mid \boldsymbol{x}\right) \end{bmatrix}}_{\boldsymbol{f}^{\natural}(\boldsymbol{x})}$$

Noise Generation Modeling Approach

- Among approaches including noisy label filtering [1–4], robust noise design [5–9], the **noise generation modeling** [10–18] approach is the most popular.
- Let $\boldsymbol{x}, y, \hat{y}$ be the sample's feature, its true label, and the obtained noisy label,

$$\underbrace{\begin{bmatrix} \Pr(\widehat{y}=1 \mid \boldsymbol{x}) \\ \dots \\ \Pr(\widehat{y}=K \mid \boldsymbol{x}) \end{bmatrix}}_{\boldsymbol{g}^{\natural}(\boldsymbol{x})} = \underbrace{\begin{bmatrix} \Pr(\widehat{y}=1 \mid y=1, \boldsymbol{x}) & \dots & \Pr(\widehat{y}=1 \mid y=K, \boldsymbol{x}) \\ \dots & \dots & \dots \\ \Pr(\widehat{y}=K \mid y=1, \boldsymbol{x}) & \dots & \Pr(\widehat{y}=K \mid y=K, \boldsymbol{x}) \end{bmatrix}}_{\boldsymbol{T}^{\natural}(\boldsymbol{x})} \underbrace{\begin{bmatrix} \Pr(y=1 \mid \boldsymbol{x}) \\ \dots \\ \Pr(y=K \mid \boldsymbol{x}) \end{bmatrix}}_{\boldsymbol{f}^{\natural}(\boldsymbol{x})}$$

$$\underbrace{\boldsymbol{g}^{\natural}(\boldsymbol{x})}_{\substack{\text{noisy label} \\ \text{probability vector}}} = \underbrace{\boldsymbol{T}^{\natural}(\boldsymbol{x})}_{\text{confusion matrix}} \underbrace{\boldsymbol{f}^{\natural}(\boldsymbol{x})}_{\substack{\text{true label} \\ \text{probability vector}}}$$

Instance-Dependent Noise Model

Noise generation model:

$$\boldsymbol{g}^{\natural}(\boldsymbol{x}) = \boldsymbol{T}^{\natural}(\boldsymbol{x})\boldsymbol{f}^{\natural}(\boldsymbol{x}).$$

- Learning under instance-dependent confusion matrices is an ill-posed problem.

Instance-Dependent Noise Model

Noise generation model:

$$\boldsymbol{g}^{\natural}(\boldsymbol{x}) = \boldsymbol{T}^{\natural}(\boldsymbol{x})\boldsymbol{f}^{\natural}(\boldsymbol{x}).$$

- Learning under instance-dependent confusion matrices is an ill-posed problem.
- Most existing works [12, 14, 15, 17, 19, 20] resort to a simplification: $\boldsymbol{T}^{\natural}(\boldsymbol{x}) = \boldsymbol{A}^{\natural}, \ \forall \boldsymbol{x}$.

Instance-Dependent Noise Model

Noise generation model:

$$\boldsymbol{g}^{\natural}(\boldsymbol{x}) = \boldsymbol{T}^{\natural}(\boldsymbol{x})\boldsymbol{f}^{\natural}(\boldsymbol{x}).$$

- Learning under instance-dependent confusion matrices is an ill-posed problem.
- Most existing works [12, 14, 15, 17, 19, 20] resort to a simplification: $\boldsymbol{T}^{\natural}(\boldsymbol{x}) = \boldsymbol{A}^{\natural}, \forall \boldsymbol{x}$.
- However, real data exhibits a more complex confusion matrix.

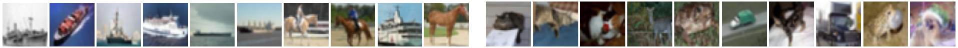


Figure: Nominal images (left) exhibits similar labeling difficulty, whereas special/outlier images (right) display a wide range of labeling challenges.

Instance-Dependent Noise Model

Noise generation model:

$$\boldsymbol{g}^{\natural}(\boldsymbol{x}) = \boldsymbol{T}^{\natural}(\boldsymbol{x})\boldsymbol{f}^{\natural}(\boldsymbol{x}).$$

- Learning under instance-dependent confusion matrices is an ill-posed problem.
- Most existing works [12, 14, 15, 17, 19, 20] resort to a simplification: $\boldsymbol{T}^{\natural}(\boldsymbol{x}) = \boldsymbol{A}^{\natural}, \ \forall \boldsymbol{x}$.
- However, real data exhibits a more complex confusion matrix.

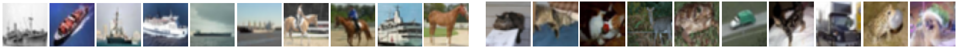


Figure: Nominal images (left) exhibits similar labeling difficulty, whereas special/outlier images (right) display a wide range of labeling challenges.

- We consider an instance-dependent noise model

$$\begin{aligned} &\text{Nominal samples:} \quad \boldsymbol{T}^{\natural}(\boldsymbol{x}_n) = \boldsymbol{A}^{\natural} \quad , \text{if } n \in \mathcal{O} \subseteq [N] \\ &\text{Outlier samples:} \quad \boldsymbol{T}^{\natural}(\boldsymbol{x}_n) = \boldsymbol{A}^{\natural}(\boldsymbol{x}_n) \quad \text{for some } \boldsymbol{A}^{\natural}(\cdot), \quad \text{otherwise.} \end{aligned}$$

Identifiability Guarantee

Proposed criterion:

$$\underset{\{\boldsymbol{A}_m\in\mathcal{A}\},\{\boldsymbol{e}_n^{(m)}\in\mathcal{E}\},\boldsymbol{f}\in\mathcal{F}}{\text{minimize}} \;\; \mathsf{L}_{\mathsf{ce}} \triangleq -\frac{1}{S}\sum_{(m,n)\in\mathcal{S}}\sum_{k=1}^{K}\mathbb{1}[\widehat{y}_n^{(m)}=k]\log\left[\boldsymbol{A}_m\boldsymbol{f}(\boldsymbol{x}_n)+\boldsymbol{e}_n^{(m)}\right]_k, \tag{1a}$$

$$\text{subject to} \;\; \sum_{n=1}^{N}\mathbb{1}\left\{\sum_{m=1}^{M}\|\boldsymbol{e}_n^{(m)}\|_2>0\right\}\leq C, \tag{1b}$$

Identifiability Guarantee

Proposed criterion:

$$\underset{\{\boldsymbol{A}_m \in \mathcal{A}\}, \{\boldsymbol{e}_n^{(m)} \in \mathcal{E}\}, \boldsymbol{f} \in \mathcal{F}}{\text{minimize}} \quad \mathsf{L}_{\mathsf{ce}} \triangleq -\frac{1}{S} \sum_{(m,n) \in \mathcal{S}} \sum_{k=1}^{K} \mathbb{1}[\widehat{y}_n^{(m)} = k] \log \left[\boldsymbol{A}_m \boldsymbol{f}(\boldsymbol{x}_n) + \boldsymbol{e}_n^{(m)} \right]_k, \tag{1a}$$

$$\text{subject to} \quad \sum_{n=1}^{N} \mathbb{1} \left\{ \sum_{m=1}^{M} \|\boldsymbol{e}_n^{(m)}\|_2 > 0 \right\} \leq C, \tag{1b}$$

Theorem (Identifiability and Generalization)

Let $(\{\widehat{\boldsymbol{A}}_m\}, \{\widehat{\boldsymbol{e}}_n^{(m)}\}, \widehat{\boldsymbol{f}})$ be any optimal solution of (1)*. The following result holds with probability greater than $1 - 2/S - K/T^\alpha$:*

$$\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_{\boldsymbol{x}}} \left[\min_{\boldsymbol{\Pi}} \|\widehat{\boldsymbol{f}}(\boldsymbol{x}) - \boldsymbol{\Pi}^\top \boldsymbol{f}^\natural(\boldsymbol{x})\|_2^2 \right] \leq K(\eta + \xi_1 + \xi_2),$$

$$\min_{\boldsymbol{\Pi}} \|\widehat{\boldsymbol{A}}_m - \boldsymbol{A}_m^\natural \boldsymbol{\Pi}\|_{\mathrm{F}}^2 = K\sigma^2(\eta + \xi_1 + \xi_2), \ \forall m,$$

where $\eta^2 = \mathcal{O}\left(\beta M T^\alpha \sqrt{S} \left(\sqrt{M \log S} + (\|\boldsymbol{X}\| \, R_{\mathcal{F}})^{0.25}\right)\right)$, $\boldsymbol{\Pi}$ a permutation matrix, and $T = N - |\mathcal{I}|$. In addition, we have exact outlier detection, i.e., $\widehat{\mathcal{I}} = \mathcal{I}$.

Identifiability Guarantee

Proposed criterion:

$$\underset{\{\boldsymbol{A}_m\in\mathcal{A}\},\{\boldsymbol{e}_n^{(m)}\in\mathcal{E}\},\boldsymbol{f}\in\mathcal{F}}{\text{minimize}} \quad \mathsf{L}_{\mathsf{ce}} \triangleq -\frac{1}{S}\sum_{(m,n)\in\mathcal{S}}\sum_{k=1}^{K}\mathbb{1}[\widehat{y}_n^{(m)}=k]\log\left[\boldsymbol{A}_m\boldsymbol{f}(\boldsymbol{x}_n)+\boldsymbol{e}_n^{(m)}\right]_k, \tag{1a}$$

$$\text{subject to} \quad \sum_{n=1}^{N}\mathbb{1}\left\{\sum_{m=1}^{M}\|\boldsymbol{e}_n^{(m)}\|_2>0\right\}\le C, \tag{1b}$$

Theorem (Identifiability and Generalization)

Let $(\{\widehat{\boldsymbol{A}}_m\},\{\widehat{\boldsymbol{e}}_n^{(m)}\},\widehat{\boldsymbol{f}})$ be any optimal solution of (1). The following result holds with probability greater than $1-2/S-K/T^\alpha$:

$$\underset{\boldsymbol{x}\sim\mathcal{D}_{\boldsymbol{x}}}{\mathbb{E}}\left[\min_{\boldsymbol{\Pi}}\|\widehat{\boldsymbol{f}}(\boldsymbol{x})-\boldsymbol{\Pi}^\top\boldsymbol{f}^\natural(\boldsymbol{x})\|_2^2\right]\le K(\eta+\xi_1+\xi_2),$$

$$\min_{\boldsymbol{\Pi}}\|\widehat{\boldsymbol{A}}_m-\boldsymbol{A}_m^\natural\boldsymbol{\Pi}\|_{\mathrm{F}}^2=K\sigma^2(\eta+\xi_1+\xi_2),\ \forall m,$$

where $\eta^2=\mathcal{O}\left(\beta MT^\alpha\sqrt{S}\left(\sqrt{M\log S}+(\|\boldsymbol{X}\|\,R_{\mathcal{F}})^{0.25}\right)\right)$, $\boldsymbol{\Pi}$ a permutation matrix, and $T=N-|\mathcal{I}|$. In addition, we have exact outlier detection, i.e., $\widehat{\mathcal{I}}=\mathcal{I}$.

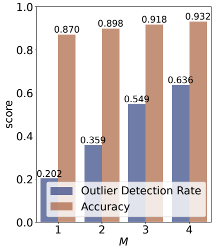


Figure: Performance of the proposal on CIFAR-10 with synthetic labels against different number of annotators.

Experiments using Real Datasets

Data.

- CIFAR10-N [21]. $N = 60000, K = 10, M = 3$. The error rates of annotators are 17.23%, 18.12%, and 17.64%.
- LabelMe [22, 23]. $N = 2688, K = 8, M = 59$. The average error rate is 25.95%.
- ImageNet-15N: we acquire noisy annotations by asking AMT workers to annotate some images from ImageNet. $K = 15$, $N = 2,514$, $M = 100$. The average error rate of the annotators is 42.68%.

Table: Average classification accuracy on CIFAR-10N, LabelMe, and ImageNet-15N datasets, labeled by human annotators. **Bold black** represents the best and blue represents the second best.

Method/Dataset	CIFAR-10N	LabelMe	ImageNet-15N
PTD	89.52 $\pm$ 0.24	84.18 $\pm$ 1.36	65.53 $\pm$ 0.18
BLTM	75.68 $\pm$ 0.47	82.10 $\pm$ 0.56	66.57 $\pm$ 0.76
VolMinNet	86.58 $\pm$ 0.21	79.97 $\pm$ 0.16	63.11 $\pm$ 1.08
Reweight	89.56 $\pm$ 0.30	84.51 $\pm$ 0.50	65.85 $\pm$ 2.93
GCE	78.01 $\pm$ 7.23	83.41 $\pm$ 0.59	64.71 $\pm$ 1.38
MEIDTM	68.69 $\pm$ 0.31	83.53 $\pm$ 0.21	72.66 $\pm$ 0.58
CrowdLayer	87.38 $\pm$ 0.43	82.80 $\pm$ 0.90	61.36 $\pm$ 2.73
TraceReg	86.57 $\pm$ 0.24	82.83 $\pm$ 0.23	68.43 $\pm$ 0.12
MaxMIG	90.11 $\pm$ 0.09	83.73 $\pm$ 0.84	81.13 $\pm$ 1.42
GeoCrowdNet(F)	87.19 $\pm$ 0.37	87.21 $\pm$ 0.39	80.45 $\pm$ 1.77
GeoCrowdNet(W)	86.43 $\pm$ 0.44	82.83 $\pm$ 0.75	68.79 $\pm$ 0.27
COINNet (Ours)	**92.09** $\pm$ **0.47**	**87.60** $\pm$ **0.54**	**93.71** $\pm$ **3.26**

Qualitative Results

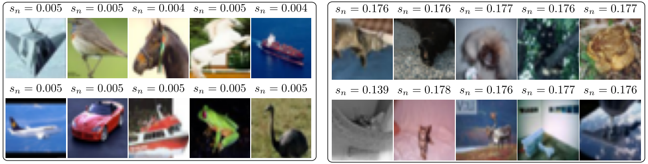


Figure: Examples from CIFAR-10N with low (left) and high (right) $s_n = \sum_{m=1}^{M} \|\widehat{\boldsymbol{e}}_n^{(m)}\|^2$.

Qualitative Results

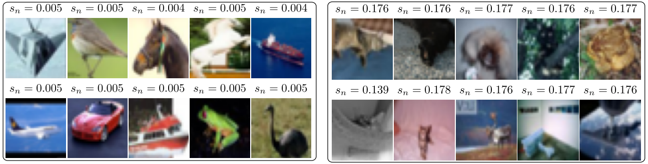


Figure: Examples from CIFAR-10N with low (left) and high (right) $s_n = \sum_{m=1}^{M} \|\widehat{\boldsymbol{e}}_n^{(m)}\|^2$.

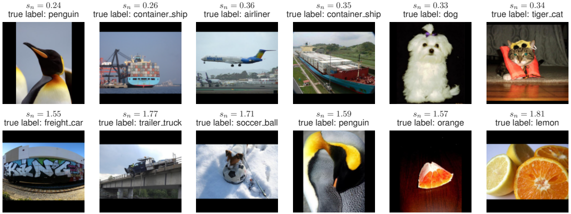


Figure: Examples from ImageNet-15N with low (top) and high (bottom) $s_n = \sum_{m=1}^{M} \|\widehat{\boldsymbol{e}}_n^{(m)}\|^2$.