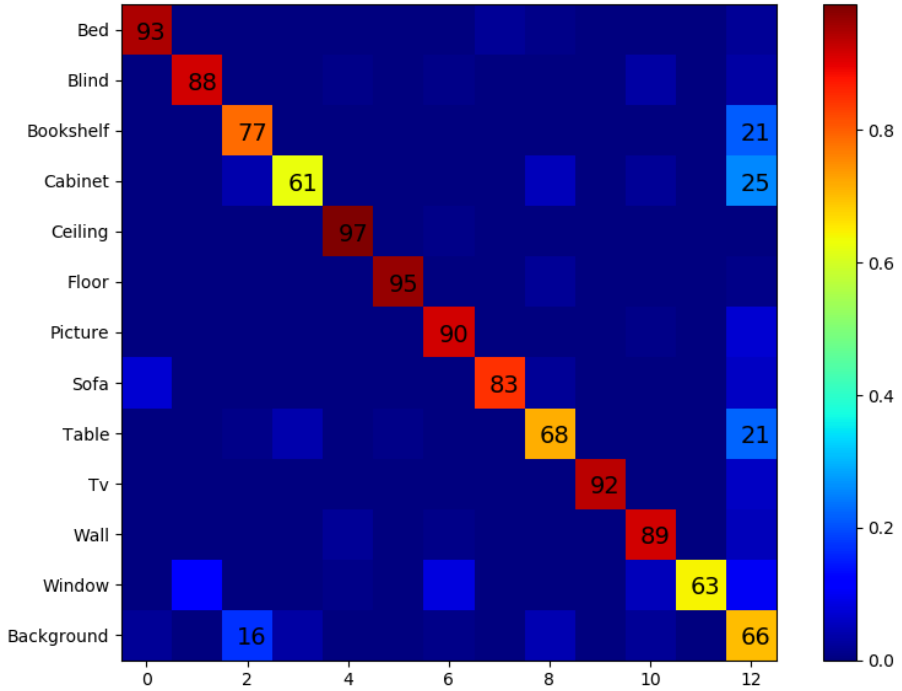
Beyond Class-Conditional Assumption: A Primary Attempt to Combat Instance-Dependent Label Noise

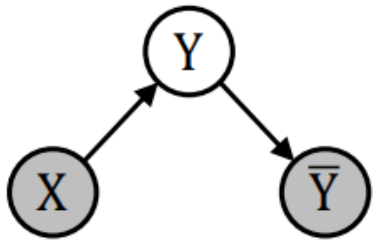
Pengfei Chen,[1] **Junjie Ye,**[2*] **Guangyong Chen,**[3*] **Jingwei Zhao,**[2] **Pheng-Ann Heng**[1,3]

[1] The Chinese University of Hong Kong
[2] VIVO AI Lab
[3] Shenzhen Key Laboratory of Virtual Reality and Human Interaction Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
{pfchen, pheng}@cse.cuhk.edu.hk, {junjie.ye, jingwei.zhao}@vivo.com, gy.chen@siat.ac.cn

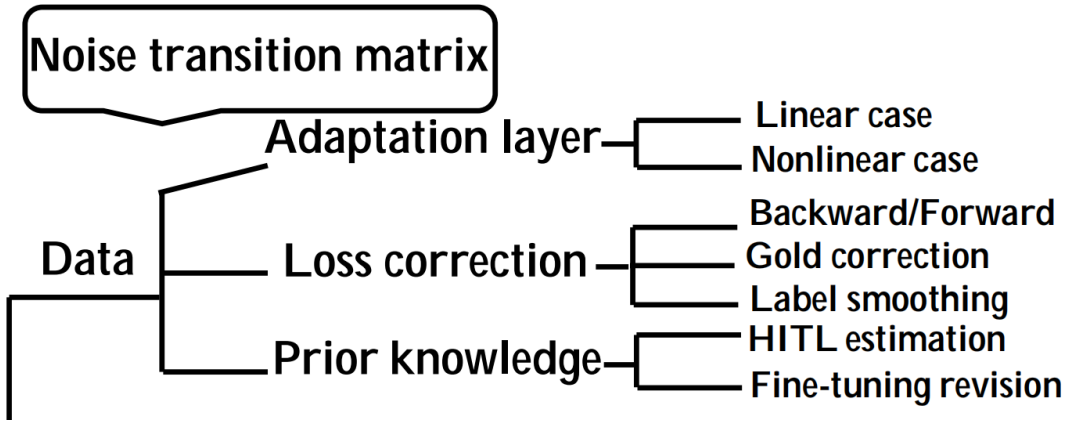
AAAI 2021

Class-Conditional Noise





(a) CCN



A perfect mapping from truth to noisy labels

Class-Conditional Noise

Figure 1: Examples of *8* in MNIST (first row) and *Airplane* in CIFAR-10 (second row). It is problematic to assume a same probability of mislabeling for diverse samples in each class.

Generalization error $er_{\bar{\mathcal{D}}}^{0-1}[f] \geq 1 - \sum_{p=1}^{c} w_p \max_{q \in \mathcal{Y}} M_{p,q}.$

Validation error $\hat{er}_{\bar{D}'}^{0-1}[f] = \sum_{i=1}^{m'} \frac{1}{m} \mathbb{1}(f(x_i) \neq \bar{y}_i)$

Hoeffding's inequality

$$\mathbb{P}(\overline{X} - \mathbb{E}[\overline{X}] \geq t) \leq \exp\left(-\frac{2t^2 n^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right),$$

Lower Bound of generalization error (0.3817)

0.2212

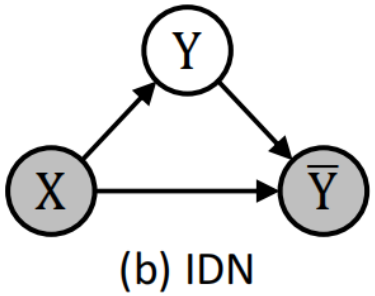
$$\Pr\left[1 - \sum_{p=1}^{c} w_p \max_{q \in \mathcal{Y}} M_{p,q} - \hat{er}_{\bar{D}'}^{0-1}[f] \geq \varepsilon\right]$$

$$\leq \Pr\left[er_{\bar{\mathcal{D}}}^{0-1}[f] - \hat{er}_{\bar{D}'}^{0-1}[f] \geq \varepsilon\right] \leq e^{-2m\varepsilon^2}.$$

m=500K

Validation error (0.1605)

Instance-dependent Noise



(b) IDN

Definition 2. *(IDN Model) Under the IDN model,* $M : \mathcal{X} \to [0,1]^{c\times c}$ *is a function of* X*. We observe samples* $(X, \bar{Y}) \sim \bar{\mathcal{D}} = \mathrm{IDN}(\mathcal{D}, M)$*, where first we draw* $(X, Y) \sim \mathcal{D}$ *as usual, then flip* Y *to produce* $\bar{Y}$ *according to the conditional probability defined by* $M(X)$*, i.e.,* $\Pr(\bar{Y} = q|Y = p) = M_{p,q}(X)$*, where* $p, q \in \mathcal{Y}$*.*

1. Assumption

Stronger noise for samples closer to the decision boundary of the Bayesian optimal classifier

2. Challenges:
   - No recognized formation of noise pattern (Part transition matrix)
   - Distribution of instances change $supp(P(X|\bar{Y} = Y, Y = p)) \stackrel{?}{=} supp(P(X|Y = p))$
   - DNN fit IDN instance easier
   - Memorization effect performs less significantly

Xia, X.; et. 2020. Parts-dependent label noise: Towards instance-dependent label noise. NIPS 2020.

A controllable IDN generator

Sym-flipping: $T = \begin{bmatrix} 1-\tau & \frac{\tau}{n-1} & \cdots & \frac{\tau}{n-1} \\ \frac{\tau}{n-1} & 1-\tau & & \frac{\tau}{n-1} \\ \vdots & & \ddots & \vdots \\ \frac{\tau}{n-1} & \frac{\tau}{n-1} & \cdots & 1-\tau \end{bmatrix}$

Pair-flipping: $T = \begin{bmatrix} 1-\tau & \tau & 0 & 0 \\ 0 & 1-\tau & \tau & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & & \tau \\ \tau & 0 & \cdots & 1-\tau \end{bmatrix}$

The score of mislabeling

$$S = \sum_{t=1}^{T} S^t / T \in \mathbb{R}^{n \times c},$$

$$N(x_i) = \max_{k \neq y_i} S_{i,k}, \quad \tilde{y}(x_i) = \arg\max_{k \neq y_i} S_{i,k},$$

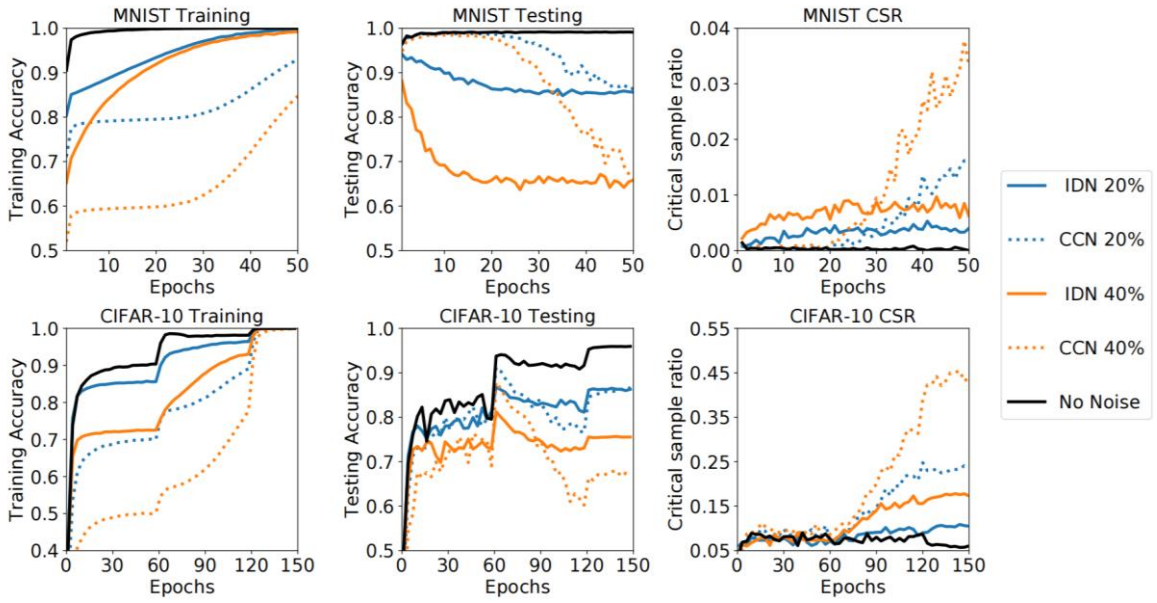
The instance with noisy label and high confidence

Algorithm 1 IDN Generation.

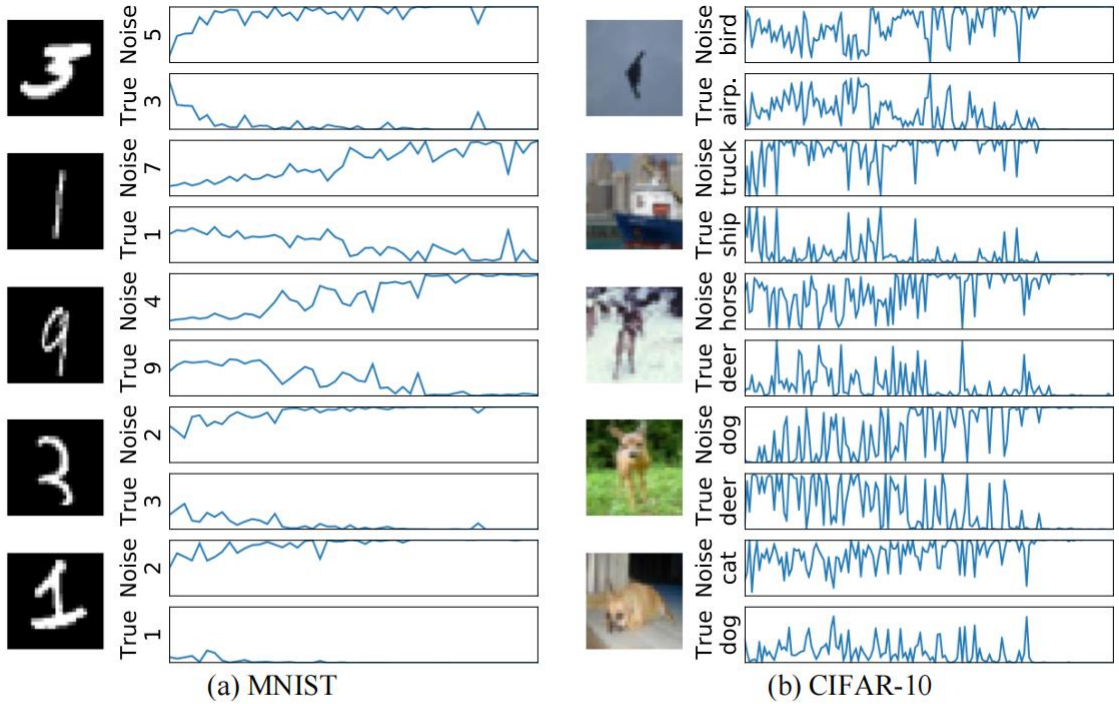
Input: Clean samples $D = \{(x_i, y_i)\}_{i=1}^n$, a targeted noise fraction p, epochs T.
Initialize a network f.
for $t = 1$ **to** T **do**
  for batches $\{(x_i, y_i)\}_{i \in \mathcal{B}}$ **do**
    Train f on $\{(x_i, y_i)\}_{i \in \mathcal{B}}$ using cross-entropy loss:
    $\mathcal{L}_{CE} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \log(f_{y_i}^t(x_i))$
  end for
  Record output $S^t = [f^t(x_i)]_{i=1}^n \in \mathbb{R}^{n \times c}$.
end for
Compute $N(x_i)$, $\tilde{y}(x_i)$ using $\{S^t\}_{t=1}^T$ (Eq. (3)).
Compute the index set $\mathcal{I} = \{p\% \arg\max_{1 \leq i \leq n} N(x_i)\}$.
Flip $\bar{y}_i = \tilde{y}_i$ if $i \in \mathcal{I}$ else keep $\bar{y}_i = y_i$.
Output: A dataset with IDN: $\bar{D} = \{(x_i, \bar{y}_i)\}_{i=1}^n$.

Some effects on IDN

It is easier for DNNs to fit IDN



The memorization effect still exists under IDN, but it is less significant compared to CCN.



(a) MNIST (b) CIFAR-10

Figure 3: Examples of softmax outputs on the noisy label and latent true label. The x-axis is training epoch and the y-axis is DNN's output probability. The airp. is airplane for short.

A simple method : SEAL

Algorithm 2 An iteration of SEAL.

Input: Noisy samples $\bar{D} = \{(x_i, \bar{y}_i)\}_{i=1}^{n}$, epochs T, soft labels from the last iteration $\bar{S}$ (optional).
Initialize a network f.
if $\bar{S}$ is not available **then**
  # The initial iteration, using one-hot noisy labels $\bar{S} = [e_{\bar{y}_i}]_{i=1}^{n} \in \mathbb{R}^{n \times c}$ *where* $e_{\bar{y}_i}$ *is the one-hot label.*
end if
for $t = 1$ **to** T **do**
  for batches $\{(x_i, \bar{S}_i)\}_{i \in \mathcal{B}}$ **do**
    Train f on $\{(x_i, \bar{S}_i)\}_{i \in \mathcal{B}}$ using the loss:
      $\mathcal{L}_{SEAL} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \sum_{k=1}^{c} \bar{S}_{i,k} \log(f_k^t(x_i))$
  end for
  Record output $\bar{S}^t = [f^t(x_i)]_{i=1}^{n} \in \mathbb{R}^{n \times c}$.
end for
Update $\bar{S} = \sum_{t=1}^{T} \bar{S}^t / T \in \mathbb{R}^{n \times c}$.
Output: Trained f, $\bar{S}$ (can be used in next iteration).

Experiments

Table 1: Classification accuracies (%) on MNIST under instance-dependent label noise with different noise fractions.

Method	10%	20%	30%	40%
CE	94.07 ±0.29	85.62 ±0.56	75.75 ±0.09	65.83 ±0.56
Forward	93.93 ±0.14	85.39 ±0.92	76.29 ±0.81	68.30 ±0.42
Co-teaching	95.77 ±0.03	91.07 ±0.19	86.20 ±0.35	79.30 ±0.84
GCE	94.56 ±0.31	86.71 ±0.47	78.32 ±0.43	69.78 ±0.58
DAC	94.13 ±0.02	85.63 ±0.56	75.82 ±0.58	65.69 ±0.78
DMI	94.21 ±0.12	87.02 ±0.42	76.19 ±0.64	67.65 ±0.73
SEAL	**96.75** ±0.08	**93.63** ±0.33	**88.52** ±0.15	**80.73** ±0.41

Table 2: Classification accuracies (%) on CIFAR-10 under instance-dependent label noise with different noise fractions.

Method	10%	20%	30%	40%
CE	91.25 ±0.27	86.34 ±0.11	80.87 ±0.05	75.68 ±0.29
Forward	91.06 ±0.02	86.35 ±0.11	78.87 ±2.66	71.12 ±0.47
Co-teaching	91.22 ±0.25	87.28 ±0.20	84.33 ±0.17	78.72 ±0.47
GCE	90.97 ±0.21	86.44 ±0.23	81.54 ±0.15	76.71 ±0.39
DAC	90.94 ±0.09	86.16 ±0.13	80.88 ±0.46	74.80 ±0.32
DMI	91.26 ±0.06	86.57 ±0.16	81.98 ±0.57	77.81 ±0.85
SEAL	**91.32** ±0.14	**87.79** ±0.09	**85.30** ±0.01	**82.98** ±0.05

Table 3: Testing accuracy (%) on Clothing1M. The $*$ marks published results.

Method	Accuracy
CE$*$	68.94
Forward$*$	69.84
Co-teaching	70.15
GCE$*$	69.09
Joint Optimization$*$	72.16
DMI$*$	72.46
CE	69.07
SEAL	**70.63**
DMI	72.27
SEAL (DMI)	**73.40**

Experiments

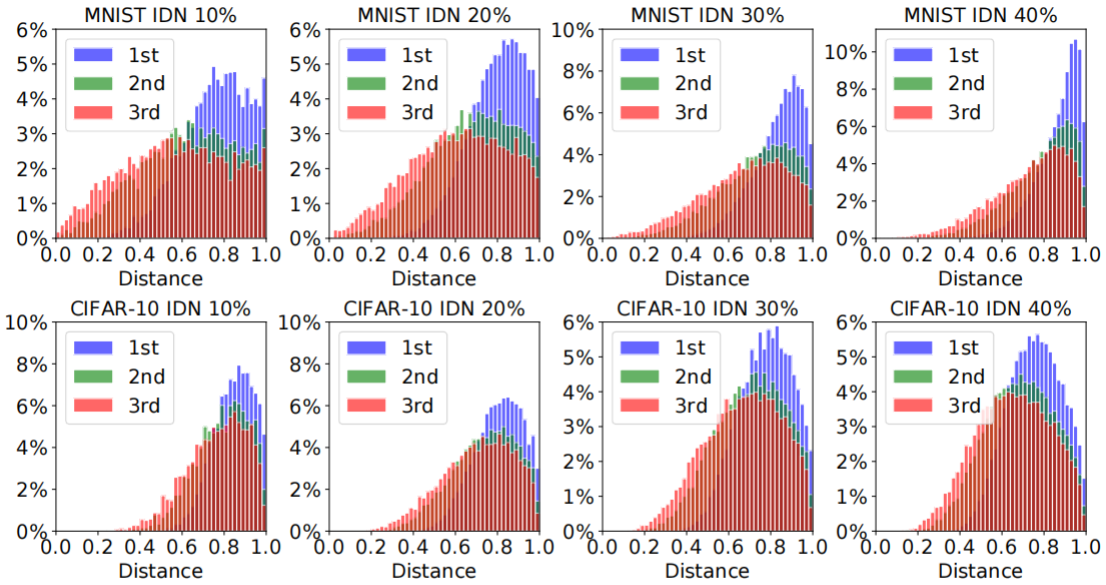


Figure 5: Histograms of distance distribution, where distance is evaluated between the true label and the soft label obtained by SEAL in 1-3 iteration.

Figure 6: The noisy label and label correction (in parentheses) obtained from SEAL. The airp. and auto. are airplane and automobile for short.

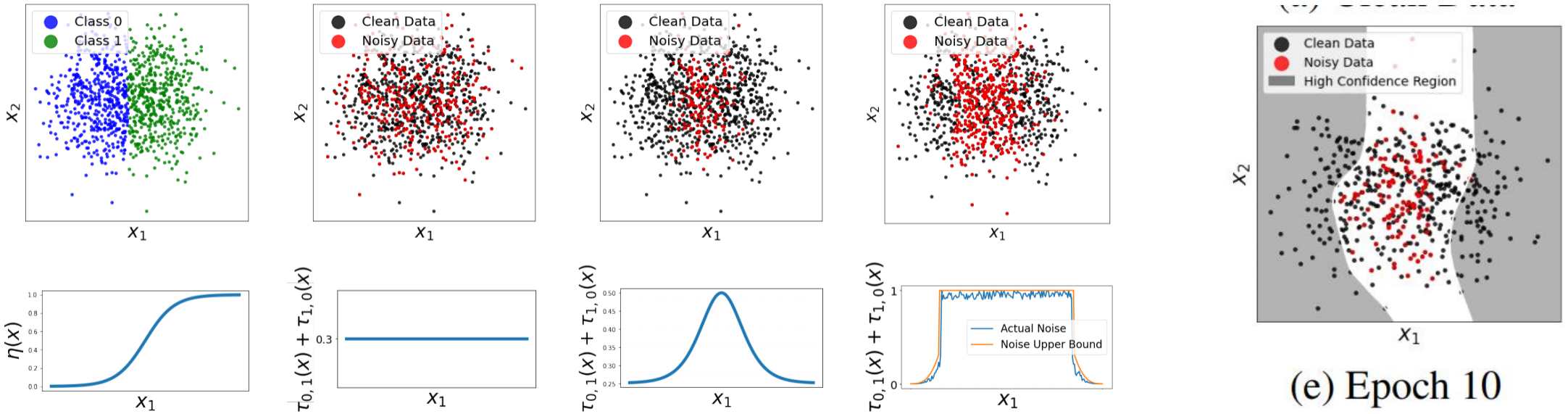
LEARNING WITH FEATURE-DEPENDENT LABEL NOISE: A PROGRESSIVE APPROACH

Yikai Zhang[1*], **Songzhu Zheng**[2*], **Pengxiang Wu**[1*], **Mayank Goswami**[3], **Chao Chen**[2]
[1]Rutgers University, {yz422,pw241}@cs.rutgers.edu
[2]Stony Brook University, {zheng.songzhu,chao.chen.1}@stonybrook.edu
[3]City University of New York, mayank.isi@gmail.com

ICLR 2021

A more generalized noise pattern - PMD Noise

Polynomial Margin Diminishing (PMD) label noise



(a) Clean labels (b) Uniform noise (c) BCN noise (d) PMD noise

(e) Epoch 10

Definition 1 (PMD noise). *A pair of noise functions $\tau_{0,1}(\boldsymbol{x})$ and $\tau_{1,0}(\boldsymbol{x})$ are polynomial-margin diminishing (PMD), if there exist constants $t_0 \in (0, \frac{1}{2})$, and $c_1, c_2 > 0$ such that:*

$$\tau_{1,0}(\boldsymbol{x}) \leq c_1[1-\eta(\boldsymbol{x})]^{1+c_2}; \forall \eta(\boldsymbol{x}) \geq \frac{1}{2} + t_0, \text{ and}$$
$$\tau_{0,1}(\boldsymbol{x}) \leq c_1 \eta(\boldsymbol{x})^{1+c_2}; \forall \eta(\boldsymbol{x}) \leq \frac{1}{2} - t_0. \quad (1)$$

$\eta(\boldsymbol{x}) = \mathbb{P}[y=1 \mid \boldsymbol{x}]$

Margin

Jun Du and Zhihua Cai. Modelling class noise with symmetric and asymmetric distributions. AAAI 2015.

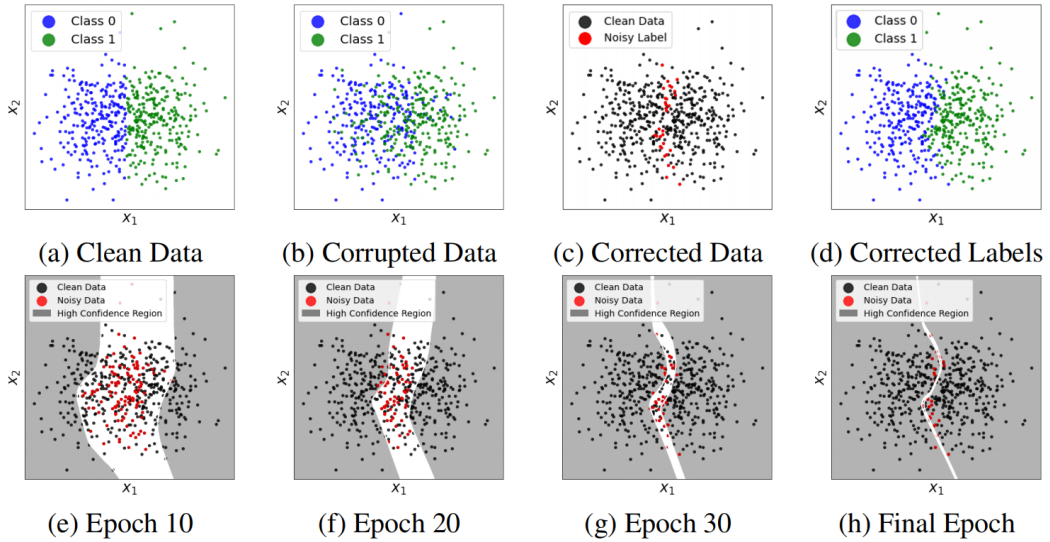
Progressive Label Correction

Algorithm 1 Progressive Label Correction

Input: Dataset $\tilde{S} = \{(\boldsymbol{x}_1, \widetilde{y}_1^0), \cdots, (\boldsymbol{x}_n, \widetilde{y}_n^0)\}$, initial NN $f(\boldsymbol{x})$, step size β, initial and end thresholds (T_0, T_{end}), warm-up m, total round N

Output: $f_{final}(\cdot)$

$T \leftarrow T_0$
$\theta \leftarrow 1/2 - T_0$
for $t \leftarrow 1, \cdots, N$ **do**
  Train $f(\boldsymbol{x})$ on $\tilde{S}$
  for all $(\boldsymbol{x}_i, \widetilde{y}_i^{t-1}) \in \tilde{S}$ **and** $|f(\boldsymbol{x}_i) - \frac{1}{2}| \geq \theta$ **do**
    $\widetilde{y}_i^t \leftarrow \mathbb{I}_{\{f(\boldsymbol{x}_i) \geq \frac{1}{2}\}}$
  end for
  if $t \geq m$ **then**
    $\theta \leftarrow 1/2 - T$
    if $\forall i \in [1, \cdots, n], \widetilde{y}_i^t = \widetilde{y}_i^{t-1}$ **then**
      $T \leftarrow \min(T(1+\beta), T_{end})$
    end if
  end if
  $\tilde{S} \leftarrow \{(\boldsymbol{x}_1, \widetilde{y}_1^t), \cdots, (\boldsymbol{x}_n, \widetilde{y}_n^t)\}$
end for



(a) Clean Data (b) Corrupted Data (c) Corrected Data (d) Corrected Labels

(e) Epoch 10 (f) Epoch 20 (g) Epoch 30 (h) Final Epoch

Experiments

$$\text{Type-I}: \tau_{u_{\boldsymbol{x}},s_{\boldsymbol{x}}} = -\frac{1}{2}\left[\eta_{u_{\boldsymbol{x}}}(\boldsymbol{x}) - \eta_{s_{\boldsymbol{x}}}(\boldsymbol{x})\right]^2 + \frac{1}{2}, \quad \text{Type-II}: \tau_{u_{\boldsymbol{x}},s_{\boldsymbol{x}}} = 1 - \left[\eta_{u_{\boldsymbol{x}}}(\boldsymbol{x}) - \eta_{s_{\boldsymbol{x}}}(\boldsymbol{x})\right]^3,$$

$$\text{Type-III}: \tau_{u_{\boldsymbol{x}},s_{\boldsymbol{x}}} = 1 - \frac{1}{3}\left[\left[\eta_{u_{\boldsymbol{x}}}(\boldsymbol{x}) - \eta_{s_{\boldsymbol{x}}}(\boldsymbol{x})\right]^3 + \left[\eta_{u_{\boldsymbol{x}}}(\boldsymbol{x}) - \eta_{s_{\boldsymbol{x}}}(\boldsymbol{x})\right]^2 + \left[\eta_{u_{\boldsymbol{x}}}(\boldsymbol{x}) - \eta_{s_{\boldsymbol{x}}}(\boldsymbol{x})\right]\right].$$

Table 1: Test accuracy (%) on CIFAR-10 and CIFAR-100 under different feature-dependent noise types and levels. The average accuracy and standard deviation over 3 trials are reported.

Dataset	Noise	Standard	Co-teaching+	GCE	SL	LRT	**PLC (ours)**
CIFAR-10	Type-I (35%)	78.11 ± 0.74	79.97 ± 0.15	80.65 ± 0.39	79.76 ± 0.72	80.98 ± 0.80	$\mathbf{82.80 \pm 0.27}$
	Type-I (70%)	41.98 ± 1.96	40.69 ± 1.99	36.52 ± 1.62	36.29 ± 0.66	41.52 ± 4.53	$\mathbf{42.74 \pm 2.14}$
	Type-II (35%)	76.65 ± 0.57	77.34 ± 0.44	77.60 ± 0.88	77.92 ± 0.89	80.74 ± 0.25	$\mathbf{81.54 \pm 0.47}$
	Type-II (70%)	45.57 ± 1.12	45.44 ± 0.64	40.30 ± 1.46	41.11 ± 1.92	44.67 ± 3.89	$\mathbf{46.04 \pm 2.20}$
	Type-III (35%)	76.89 ± 0.79	78.38 ± 0.67	79.18 ± 0.61	78.81 ± 0.29	81.08 ± 0.35	$\mathbf{81.50 \pm 0.50}$
	Type-III (70%)	43.32 ± 1.00	41.90 ± 0.86	37.10 ± 0.59	38.49 ± 1.46	44.47 ± 1.23	$\mathbf{45.05 \pm 1.13}$
CIFAR-100	Type-I (35%)	57.68 ± 0.29	56.70 ± 0.71	58.37 ± 0.18	55.20 ± 0.33	56.74 ± 0.34	$\mathbf{60.01 \pm 0.43}$
	Type-I (70%)	39.32 ± 0.43	39.53 ± 0.28	40.01 ± 0.71	40.02 ± 0.85	45.29 ± 0.43	$\mathbf{45.92 \pm 0.61}$
	Type-II (35%)	57.83 ± 0.25	56.57 ± 0.52	58.11 ± 1.05	56.10 ± 0.73	57.25 ± 0.68	$\mathbf{63.68 \pm 0.29}$
	Type-II (70%)	39.30 ± 0.32	36.84 ± 0.39	37.75 ± 0.46	38.45 ± 0.45	43.71 ± 0.51	$\mathbf{45.03 \pm 0.50}$
	Type-III (35%)	56.07 ± 0.79	55.77 ± 0.98	57.51 ± 1.16	56.04 ± 0.74	56.57 ± 0.30	$\mathbf{63.68 \pm 0.29}$
	Type-III (70%)	40.01 ± 0.18	35.37 ± 2.65	40.53 ± 0.60	39.94 ± 0.84	44.41 ± 0.19	$\mathbf{44.45 \pm 0.62}$

Experiments

Table 2: Test accuracy (%) on CIFAR-10 and CIFAR-100 under different hybrid noise types and levels. The average accuracy and standard deviation over 3 trials are reported.

Dataset	Noise	Standard	Co-teaching+	GCE	SL	LRT	**PLC (ours)**
CIFAR-10	Type-I + 30% Uniform	75.26 ± 0.32	78.72 ± 0.53	78.08 ± 0.66	77.79 ± 0.46	75.97 ± 0.27	$\mathbf{79.04 \pm 0.50}$
	Type-I + 60% Uniform	64.25 ± 0.78	55.49 ± 2.11	67.43 ± 1.43	67.63 ± 1.36	59.22 ± 0.74	$\mathbf{72.21 \pm 2.92}$
	Type-I + 30% Asymmetric	75.21 ± 0.64	75.43 ± 2.96	76.91 ± 0.56	77.14 ± 0.70	76.96 ± 0.45	$\mathbf{78.31 \pm 0.41}$
	Type-II + 30% Uniform	74.92 ± 0.63	75.19 ± 0.54	75.69 ± 0.21	75.08 ± 0.47	75.94 ± 0.58	$\mathbf{80.08 \pm 0.37}$
	Type-II + 60% Uniform	64.02 ± 0.66	59.89 ± 0.63	66.39 ± 0.29	66.76 ± 1.60	58.99 ± 1.43	$\mathbf{71.21 \pm 1.46}$
	Type-II + 30% Asymmetric	74.28 ± 0.39	73.37 ± 0.83	75.30 ± 0.81	75.43 ± 0.42	77.03 ± 0.62	$\mathbf{77.63 \pm 0.30}$
	Type-III + 30% Uniform	74.00 ± 0.38	77.31 ± 0.11	77.00 ± 0.12	76.22 ± 0.12	75.66 ± 0.57	$\mathbf{80.06 \pm 0.47}$
	Type-III + 60% Uniform	63.96 ± 0.69	56.78 ± 1.56	67.53 ± 0.51	67.79 ± 0.54	59.36 ± 0.93	$\mathbf{73.48 \pm 1.84}$
	Type-III + 30% Asymmetric	75.31 ± 0.34	74.62 ± 1.71	75.70 ± 0.91	76.09 ± 0.10	77.19 ± 0.74	$\mathbf{77.54 \pm 0.70}$
CIFAR-100	Type-I + 30% Uniform	48.86 ± 0.56	52.33 ± 0.64	52.90 ± 0.53	51.34 ± 0.64	45.66 ± 1.60	$\mathbf{60.09 \pm 0.15}$
	Type-I + 60% Uniform	35.97 ± 1.12	27.17 ± 1.66	38.62 ± 1.65	37.57 ± 0.43	23.37 ± 0.72	$\mathbf{51.68 \pm 0.10}$
	Type-I + 30% Asymmetric	45.85 ± 0.93	51.21 ± 0.31	52.69 ± 1.14	50.18 ± 0.97	52.04 ± 0.15	$\mathbf{56.40 \pm 0.34}$
	Type-II + 30% Uniform	49.32 ± 0.36	51.99 ± 0.75	53.61 ± 0.46	50.58 ± 0.25	43.86 ± 1.31	$\mathbf{60.01 \pm 0.63}$
	Type-II + 60% Uniform	35.16 ± 0.05	25.91 ± 0.64	39.58 ± 3.13	37.93 ± 0.22	23.05 ± 0.99	$\mathbf{49.35 \pm 1.53}$
	Type-II + 30% Asymmetric	46.50 ± 0.95	51.07 ± 1.44	51.98 ± 0.37	49.46 ± 0.23	52.11 ± 0.46	$\mathbf{61.43 \pm 0.33}$
	Type-III + 30% Uniform	48.94 ± 0.61	49.94 ± 0.44	52.07 ± 0.35	50.18 ± 0.54	42.79 ± 1.78	$\mathbf{60.14 \pm 0.97}$
	Type-III + 60% Uniform	34.67 ± 0.16	22.89 ± 0.75	36.82 ± 0.49	37.65 ± 1.42	22.81 ± 0.72	$\mathbf{50.73 \pm 2.16}$
	Type-III + 30% Asymmetric	45.70 ± 0.12	49.38 ± 0.86	50.87 ± 1.12	48.15 ± 0.90	50.31 ± 0.39	$\mathbf{54.56 \pm 1.11}$

Experiments

Table 5: Test accuracy (%) on Clothing1M.

Method	Standard	Forward	D2L	JO	PENCIL	DY	GCE	SL	MLNT	LRT	**PLC (ours)**
Accuracy	68.94	69.84	69.47	72.23	73.49	71.00	69.75	71.02	73.47	71.74	**74.02**

Table 6: Test accuracy (%) on Food-101N.

Method	Accuracy
Standard	81.67
CleanNet (Lee et al., 2018)	83.95
PLC (ours)	**85.28 ± 0.04**

Table 7: Test accuracy (%) on ANIMAL-10N.

Method	Accuracy
Standard	79.4 ± 0.14
SELFIE (Song et al., 2019)	81.8 ± 0.09
PLC (ours)	**83.4 ± 0.43**