Asymmetric Loss Functions for Learning with Noisy Labels

Xiong Zhou[1 2], Xianming Liu[1 2], Junjun Jiang[1 2]
Xin Gao[2 3], Xiangyang Ji[4]

ICML International Conference On Machine Learning

ICML | 2021
Thirty-eighth International Conference on Machine Learning

[1] *Harbin Institute of Technology*

[2] *Peng Cheng Laboratory*

[3] *King Abdullah University of Science and Technology*

[4] *Tsinghua University*

July 18-25, 2021

Clean Labels Dominate

- The label noise model can be formulated as

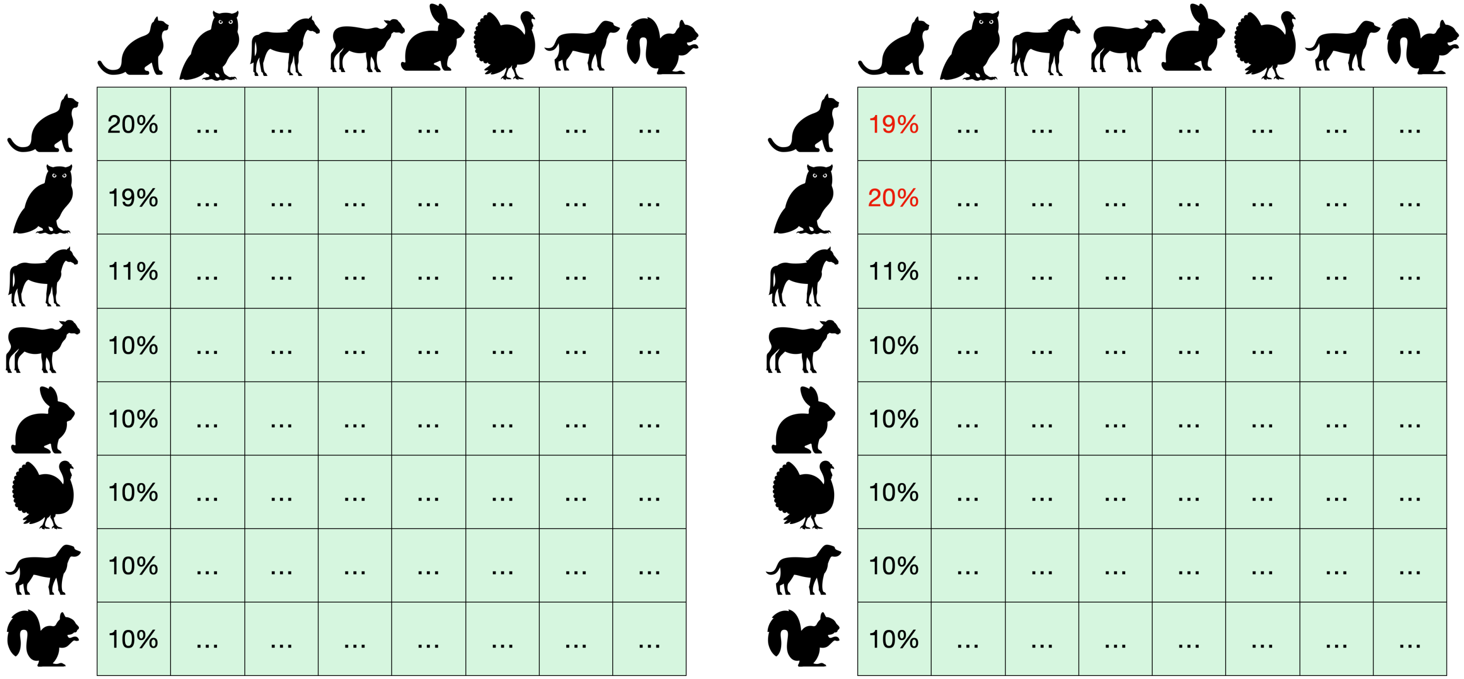
$$\tilde{y}_n = \begin{cases} y_n & \text{with probability } (1 - \eta_{\boldsymbol{x}_n}) \\ i, i \in [k], i \neq y_n & \text{with probability } \eta_{\boldsymbol{x}_n, i} \end{cases}.$$

Clean Labels Dominate

- The label noise model can be formulated as

$$\tilde{y}_n = \begin{cases} y_n & \text{with probability } (1 - \eta_{\boldsymbol{x}_n}) \\ i, i \in [k], i \neq y_n & \text{with probability } \eta_{\boldsymbol{x}_n, i} \end{cases}.$$

- **Assumption 1**. *The label noise model is clean-labels-dominant, i.e.,* $\forall \boldsymbol{x},\ 1 - \eta_x > \max_{j \neq y} \eta_{\boldsymbol{x},j}$.



	20%	…	…	…	…	…	…	…
	19%	…	…	…	…	…	…	…
	11%	…	…	…	…	…	…	…
	10%	…	…	…	…	…	…	…
	10%	…	…	…	…	…	…	…
	10%	…	…	…	…	…	…	…
	10%	…	…	…	…	…	…	…
	10%	…	…	…	…	…	…	…

	19%	…	…	…	…	…	…	…
	20%	…	…	…	…	…	…	…
	11%	…	…	…	…	…	…	…
	10%	…	…	…	…	…	…	…
	10%	…	…	…	…	…	…	…
	10%	…	…	…	…	…	…	…
	10%	…	…	…	…	…	…	…
	10%	…	…	…	…	…	…	…

Figure 1. Illustrations of clean-labels-dominant and clean-labels-non-dominant cases.

Asymmetric Loss Functions

- **Definition 1.** *On the given weights* $w_1, \dots, w_k \geq 0$*, where* $\exists t \in [k]$*, s.t.,* $w_t > \max_{i \neq t} w_i$*, a loss function* $L(\boldsymbol{u}, i)$ *is called* ***asymmetric*** *if* L *satisfies*

$$\arg\min_{\boldsymbol{u}} \sum_{i=1}^{k} w_i L(\boldsymbol{u}, i) = \arg\min_{u} L(\boldsymbol{u}, t),$$

where we always have $\arg\min_{u} L(\boldsymbol{u}, t) = \boldsymbol{e}_t$.

Asymmetric Loss Functions

- **Definition 1.** *On the given weights* $w_1, \dots, w_k \geq 0$*, where* $\exists t \in [k]$*, s.t.,* $w_t > \max_{i \neq t} w_i$*, a loss function* $L(\boldsymbol{u}, i)$ *is called* ***asymmetric*** *if* L *satisfies*

$$\arg\min_{\boldsymbol{u}} \sum_{i=1}^{k} w_i L(\boldsymbol{u}, i) = \arg\min_{u} L(\boldsymbol{u}, t),$$

where we always have $\arg\min_{u} L(\boldsymbol{u}, t) = \boldsymbol{e}_t$.

- Recall to the conditional L-risk $L^{\eta}(f(x), y) = (1 - \eta_{\boldsymbol{x}})L(f(\boldsymbol{x}), y) + \sum_{i \neq y} \eta_{x,i} L(f(\boldsymbol{x}), i)$. We have $(1 - \eta_{\mathbf{x}}) > \max_{i \neq y} \eta_{\mathbf{x},i}$ according to Assumption 1.

Asymmetric Loss Functions

- **Definition 1.** *On the given weights* $w_1, \ldots, w_k \geq 0$*, where* $\exists t \in [k]$*, s.t.,* $w_t > \max_{i \neq t} w_i$*, a loss function* $L(\boldsymbol{u}, i)$ *is called* ***asymmetric*** *if* L *satisfies*

$$\arg\min_{\boldsymbol{u}} \sum_{i=1}^{k} w_i L(\boldsymbol{u}, i) = \arg\min_{u} L(\boldsymbol{u}, t),$$

where we always have $\arg\min_{u} L(\boldsymbol{u}, t) = \boldsymbol{e}_t$.

- Recall to the conditional L-risk $L^{\eta}(f(x), y) = (1 - \eta_{\boldsymbol{x}}) L(f(\boldsymbol{x}), y) + \sum_{i \neq y} \eta_{\boldsymbol{x},i} L(f(\boldsymbol{x}), i)$. We have $(1 - \eta_{\mathbf{x}}) > \max_{i \neq y} \eta_{\mathbf{x},i}$ according to Assumption 1.

- We define that L is *asymmetric* on the label noise model satisfying Assumption 1, if $\forall(\boldsymbol{x}, y)$, L is asymmetric on $\{1 - \eta_{\boldsymbol{x}}\} \cup \{\eta_{\boldsymbol{x},i}\}_{i \neq y}$.

Asymmetric Loss Functions

- **Definition 1.** *On the given weights* $w_1, \dots, w_k \geq 0$*, where* $\exists t \in [k]$*, s.t.,* $w_t > \max_{i \neq t} w_i$*, a loss function* $L(\boldsymbol{u}, i)$ *is called* ***asymmetric*** *if* L *satisfies*

$$\arg\min_{\boldsymbol{u}} \sum_{i=1}^{k} w_i L(\boldsymbol{u}, i) = \arg\min_{u} L(\boldsymbol{u}, t),$$

where we always have $\arg\min_{u} L(\boldsymbol{u}, t) = \boldsymbol{e}_t$.

- Recall to the conditional L-risk $L^\eta(f(x), y) = (1 - \eta_{\boldsymbol{x}}) L(f(\boldsymbol{x}), y) + \sum_{i \neq y} \eta_{x,i} L(f(\boldsymbol{x}), i)$. We have $(1 - \eta_{\mathbf{x}}) > \max_{i \neq y} \eta_{\mathbf{x},\mathrm{i}}$ according to Assumption 1.

- We define that L is *asymmetric* on the label noise model satisfying Assumption 1, if $\forall (\boldsymbol{x}, y)$, L is asymmetric on $\{1 - \eta_{\boldsymbol{x}}\} \cup \{\eta_{\boldsymbol{x},i}\}_{i \neq y}$. L is called *completely asymmetric* on any weights that contain a unique maximum.

Asymmetric Loss Functions

- **Definition 1.** *On the given weights* $w_1, \dots, w_k \geq 0$, *where* $\exists t \in [k]$, *s.t.*, $w_t > \max_{i \neq t} w_i$, *a loss function* $L(\boldsymbol{u}, i)$ *is called* ***asymmetric*** *if* L *satisfies*

$$\arg\min_{\boldsymbol{u}} \sum_{i=1}^{k} w_i L(\boldsymbol{u}, i) = \arg\min_{u} L(\boldsymbol{u}, t)\,,$$

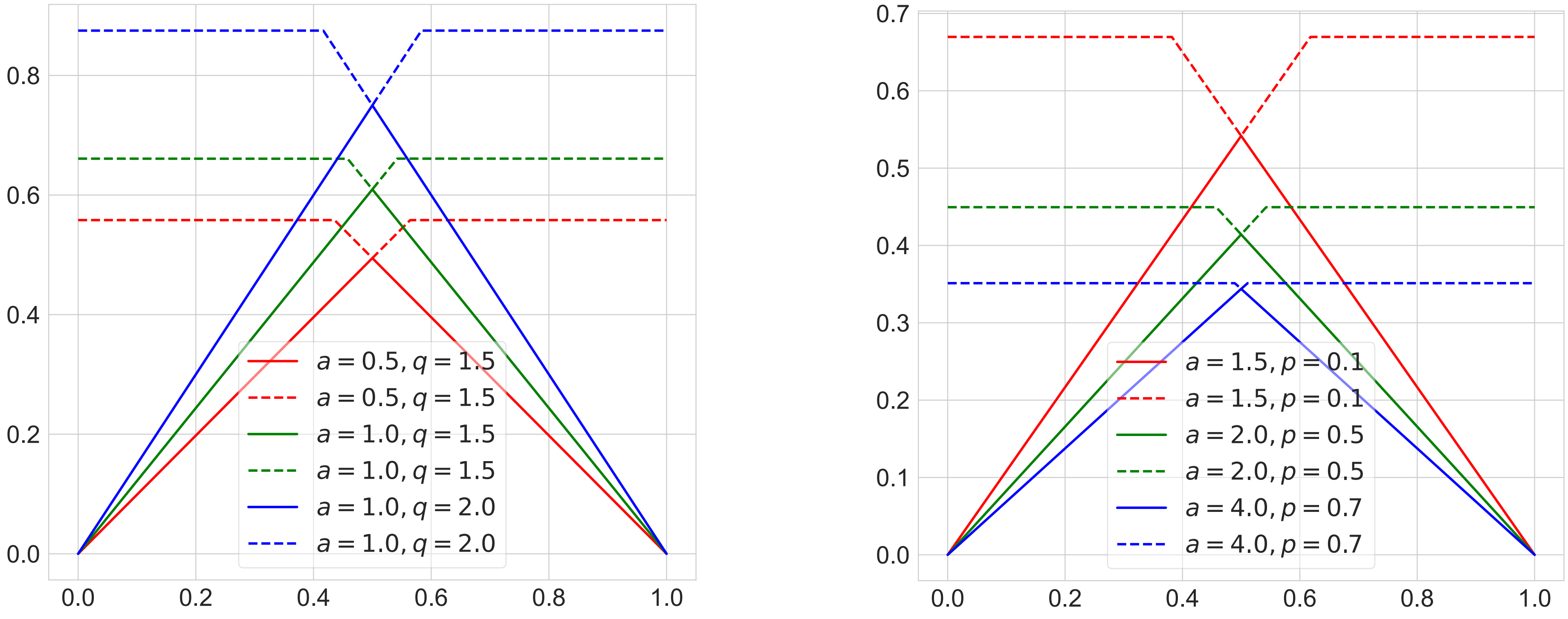
where we always have $\arg\min_{u} L(\boldsymbol{u}, t) = \boldsymbol{e}_t$.

- Recall to the conditional L-risk $L^\eta(f(x), y) = (1 - \eta_{\boldsymbol{x}}) L(f(\boldsymbol{x}), y) + \sum_{i \neq y} \eta_{x,i} L(f(\boldsymbol{x}), i)$. We have $(1 - \eta_{\mathbf{x}}) > \max_{i \neq y} \eta_{\mathbf{x},i}$ according to Assumption 1.

- We define that L is *asymmetric* on the label noise model satisfying Assumption 1, if $\forall(\boldsymbol{x}, y)$, L is asymmetric on $\{1 - \eta_{\boldsymbol{x}}\} \cup \{\eta_{\boldsymbol{x},i}\}_{i \neq y}$. L is called *completely asymmetric* on any weights that contain a unique maximum. L is called *strictly asymmetric*, if $\sum_i^k w_i L(\boldsymbol{u}, i) < \sum_i^k w_i L(\boldsymbol{u}', i)$, $\forall \boldsymbol{u}, \boldsymbol{u}' \in \mathcal{C}$, $u_t > u_t{}'$.

Properties of Asymmetric Loss Functions

Theorem 1 (Classification calibration). *Completely asymmetric loss functions are classification-calibrated.*



(a) AGCE (b) AUL

Figure 2. Verification of classification calibration. Solid and dashed lines denote the curve of $H_\ell(\eta)$ and $H_\ell^-(\eta)$

Properties of Asymmetric Loss Functions

Theorem 2 (Excess risk bound). *An excess risk bound of a strictly and completely asymmetric loss function* $L(\boldsymbol{u}, i) = \ell(u_i)$ *can be expressed as.*

$$R_{\ell_{0-1}}(f) - R^*_{\ell_{0-1}} \leq \frac{2(R_\ell(f) - R^*_\ell)}{\ell(0) - \ell(1)},$$

where $R^*_{\ell_{0-1}} = \inf_f R_{\ell_{0-1}}(f)$*, and* $R^*_\ell = \inf_f R_\ell(f)$.

Properties of Asymmetric Loss Functions

Theorem 2 (Excess risk bound). *An excess risk bound of a strictly and completely asymmetric loss function* $L(\boldsymbol{u}, i) = \ell(u_i)$ *can be expressed as.*

$$R_{\ell_{0-1}}(f) - R^*_{\ell_{0-1}} \leq \frac{2(R_\ell(f) - R^*_\ell)}{\ell(0) - \ell(1)},$$

where $R^*_{\ell_{0-1}} = \inf_f R_{\ell_{0-1}}(f)$*, and* $R^*_\ell = \inf_f R_\ell(f)$.

Theorem 3. *Symmetric loss functions are completely asymmetric.*

Properties of Asymmetric Loss Functions

Theorem 2 (Excess risk bound). *An excess risk bound of a strictly and completely asymmetric loss function* $L(\boldsymbol{u}, i) = \ell(u_i)$ *can be expressed as.*

$$R_{\ell_{0-1}}(f) - R^*_{\ell_{0-1}} \leq \frac{2(R_\ell(f) - R^*_\ell)}{\ell(0) - \ell(1)},$$

where $R^*_{\ell_{0-1}} = \inf_f R_{\ell_{0-1}}(f)$*, and* $R^*_\ell = \inf_f R_\ell(f)$*.*

Theorem 3. *Symmetric loss functions are completely asymmetric.*

Theorem 4 **(Noise tolerance)**. *In a multi-classification problem, assuming that there exists a hypothesis* $f \in \mathcal{H}, \forall(\boldsymbol{x}, y), f$ *minimizes* $L(f(\boldsymbol{x}), y)$*, then* L *is noise-tolerant if* L *is asymmetric on the label noise model.*

Properties of Asymmetric Loss Functions

Theorem 2 (Excess risk bound). *An excess risk bound of a strictly and completely asymmetric loss function* $L(\boldsymbol{u}, i) = \ell(u_i)$ *can be expressed as.*

$$R_{\ell_{0-1}}(f) - R^*_{\ell_{0-1}} \leq \frac{2(R_\ell(f) - R^*_\ell)}{\ell(0) - \ell(1)},$$

where $R^*_{\ell_{0-1}} = \inf_f R_{\ell_{0-1}}(f)$, *and* $R^*_\ell = \inf_f R_\ell(f)$.

Theorem 3. *Symmetric loss functions are completely asymmetric.*

Theorem 4 **(Noise tolerance)**. *In a multi-classification problem, assuming that there exists a hypothesis* $f \in \mathcal{H}, \forall(\boldsymbol{x}, y), f$ *minimizes* $L(f(\boldsymbol{x}), y)$, *then* L *is noise-tolerant if* L *is asymmetric on the label noise model.*

Theorem 5. $\forall \alpha, \beta > 0$, *if* L_1 *and* L_2 *are asymmetric, then* $\alpha L_1 + \beta L_2$ *is asymmetric.*

Asymmetry Ratio

Definition 2. *Consider a loss function $L(\boldsymbol{u}, i) = \ell(u_i)$, the* ***asymmetry ratio*** *is defined as*

$$r(\ell) = \inf_{\substack{0 \le u_1, u_2 \le 1 \\ u_1 + u_2 = 1 \\ 0 \le \Delta u \le u_2}} \frac{\ell(u_1) - \ell(u_1 + \Delta u)}{\ell(u_2 - \Delta u) - \ell(u_2)} \le \inf_{\substack{0 \le u_1, u_2 \le 1 \\ u_1 + u_2 = 1}} \frac{\ell(u_1) - \ell(1)}{\ell(0) - \ell(u_2)} = r^u(\ell)..$$

Asymmetry Ratio

Definition 2. *Consider a loss function* $L(\boldsymbol{u}, i) = \ell(u_i)$, *the* ***asymmetry ratio*** *is defined as*

$$r(\ell) = \inf_{\substack{0 \le u_1, u_2 \le 1 \\ u_1 + u_2 = 1 \\ 0 \le \Delta u \le u_2}} \frac{\ell(u_1) - \ell(u_1 + \Delta u)}{\ell(u_2 - \Delta u) - \ell(u_2)} \le \inf_{\substack{0 \le u_1, u_2 \le 1 \\ u_1 + u_2 = 1}} \frac{\ell(u_1) - \ell(1)}{\ell(0) - \ell(u_2)} = r^u(\ell)..$$

Theorem 6 (Sufficiency). *On the given weights* $w_1, \dots, w_k \ge 0$, *where* $w_m > w_n$ *and* $w_n = \max_{i \neq m} w_i$, *the loss function* $L(\boldsymbol{u}, i) = \ell(u_i)$ *is asymmetric if* $\frac{w_m}{w_n} \cdot r(\ell) \ge 1$.

Theorem 7 (Necessity). *On the given weights* $w_1, \dots, w_k \ge 0$, *where* $w_m > w_n$ *and* $w_n = \max_{i \neq m} w_i$, *the loss function* $L(\boldsymbol{u}, i) = \ell(u_i)$ *is asymmetric only if* $\frac{w_m}{w_n} \cdot r^u(\ell) \ge 1$.

Asymmetry Ratio

Definition 2. *Consider a loss function $L(\boldsymbol{u}, i) = \ell(u_i)$, the* ***asymmetry ratio*** *is defined as*

$$r(\ell) = \inf_{\substack{0\le u_1,u_2\le 1\\ u_1+u_2=1\\ 0\le\Delta u\le u_2}} \frac{\ell(u_1) - \ell(u_1+\Delta u)}{\ell(u_2-\Delta u) - \ell(u_2)} \le \inf_{\substack{0\le u_1,u_2\le 1\\ u_1+u_2=1}} \frac{\ell(u_1) - \ell(1)}{\ell(0) - \ell(u_2)} = r^u(\ell)..$$

Theorem 6 (Sufficiency). *On the given weights $w_1, \dots, w_k \ge 0$, where $w_m > w_n$ and $w_n = \max_{i\neq m} w_i$, the loss function $L(\boldsymbol{u}, i) = \ell(u_i)$ is asymmetric if $\frac{w_m}{w_n} \cdot r(\ell) \ge 1$.*

Theorem 7 (Necessity). *On the given weights $w_1, \dots, w_k \ge 0$, where $w_m > w_n$ and $w_n = \max_{i\neq m} w_i$, the loss function $L(\boldsymbol{u}, i) = \ell(u_i)$ is asymmetric only if $\frac{w_m}{w_n} \cdot r^u(\ell) \ge 1$.*

According to Theorem 6 and Theorem 8, when $r(\ell) = r^u(\ell)$, $\frac{w_m}{w_n} \cdot r(\ell) \ge 1$ will become the necessary and sufficient condition for $L(\boldsymbol{u}, i) = \ell(u_i)$ to be asymmetric.

AGCE, AUL, AEL

Corollary 1 (AGCE). *On the given weights* $w_1, \ldots, w_k \geq 0$*, where* $w_m > w_n$ *and* $w_n = \max_{i \neq m} w_i$*, the loss function* $L_q(\boldsymbol{u}, i) = [(a+1)^q - (a+u_i)^q]/q$ *(where* $q > 0, a > 0$*) is asymmetric if and only if* $\frac{\boldsymbol{w_m}}{\boldsymbol{w_n}} \geq \left(\frac{\boldsymbol{a+1}}{\boldsymbol{a}}\right)^{\boldsymbol{1-q}} \cdot \mathbb{I}(\boldsymbol{q} \leq \boldsymbol{1}) + \mathbb{I}(\boldsymbol{q} > \boldsymbol{1})$.

Corollary 2 (AUL). *On the given weights* $w_1, \ldots, w_k \geq 0$*, where* $w_m > w_n$ *and* $w_n = \max_{i \neq m} w_i$*, the loss function* $L_p(\boldsymbol{u}, i) = [(a-u_i)^p - (a-1)^p]/p$ *(where* $p > 0, a > 1$*) is asymmetric if and only if* $\frac{\boldsymbol{w_m}}{\boldsymbol{w_n}} \geq \left(\frac{\boldsymbol{a}}{\boldsymbol{a-1}}\right)^{\boldsymbol{p-1}} \cdot \mathbb{I}(\boldsymbol{p} \geq \boldsymbol{1}) + \mathbb{I}(\boldsymbol{p} < \boldsymbol{1})$.

Corollary 3 (AEL). *On the given weights* $w_1, \ldots, w_k \geq 0$*, where* $w_m > w_n$ *and* $w_n = \max_{i \neq m} w_i$*, the exponential loss function* $L_p(\boldsymbol{u}, i) = \exp(-u_i/a)$ *(where* $a > 0$*) is asymmetric if and only if* $\frac{\boldsymbol{w_m}}{\boldsymbol{w_n}} \geq \mathbf{exp}\left(\frac{\boldsymbol{1}}{\boldsymbol{a}}\right)$.

Experimental Results

Table 1. Test accuracies (%) of different methods on benchmark datasets with clean or symmetric label noise ($\eta \in [0.2, 0.4, 0.6, 0.8]$). The results (mean±std) are reported over 3 random runs and the top 3 best results are **boldfaced**.

Datasets	Methods	Clean ($\eta = 0.0$)	Symmetric Noise Rate (η) 0.2	0.4	0.6	0.8
MNIST	CE	99.15 ± 0.05	91.62 ± 0.39	73.98 ± 0.27	49.36 ± 0.43	22.66 ± 0.61
	FL	99.13 ± 0.09	91.68 ± 0.14	74.54 ± 0.06	50.39 ± 0.28	22.65 ± 0.26
	GCE	99.27 ± 0.05	98.86 ± 0.07	97.16 ± 0.03	81.53 ± 0.58	33.95 ± 0.82
	NLNL	98.61 ± 0.13	98.02 ± 0.14	97.17 ± 0.09	95.42 ± 0.30	86.34 ± 1.43
	SCE	99.23 ± 0.10	98.92 ± 0.12	97.38 ± 0.15	88.83 ± 0.55	48.75 ± 1.54
	NCE	98.60 ± 0.06	98.57 ± 0.01	98.29 ± 0.05	97.65 ± 0.08	93.78 ± 0.41
	NCE+RCE	99.36 ± 0.05	**99.14 ± 0.03**	98.51 ± 0.06	95.60 ± 0.21	74.00 ± 1.68
	AUL	99.14 ± 0.05	**99.05 ± 0.09**	**98.90 ± 0.09**	**98.67 ± 0.04**	**96.73 ± 0.20**
	AGCE	99.05 ± 0.11	**98.96 ± 0.10**	**98.83 ± 0.06**	**98.57 ± 0.12**	**96.59 ± 0.12**
	AEL	99.03 ± 0.05	98.93 ± 0.06	**98.78 ± 0.13**	**98.51 ± 0.06**	**96.40 ± 0.11**
CIFAR10	CE	90.48 ± 0.11	74.68 ± 0.25	58.26 ± 0.21	38.70 ± 0.53	19.55 ± 0.49
	FL	89.82 ± 0.20	73.72 ± 0.08	57.90 ± 0.45	38.86 ± 0.07	19.13 ± 0.28
	GCE	89.59 ± 0.26	87.03 ± 0.35	82.66 ± 0.17	67.70 ± 0.45	26.67 ± 0.59
	SCE	91.61 ± 0.19	87.10 ± 0.25	79.67 ± 0.37	61.35 ± 0.56	28.66 ± 0.27
	NLNL	90.73 ± 0.20	73.70 ± 0.05	63.90 ± 0.44	50.68 ± 0.47	29.53 ± 1.55
	NCE	75.65 ± 0.26	72.89 ± 0.25	69.49 ± 0.39	62.64 ± 0.18	41.49 ± 0.66
	NCE+RCE	90.87 ± 0.37	**89.25 ± 0.42**	**85.81 ± 0.08**	**79.72 ± 0.20**	**55.74 ± 0.95**
	AUL	91.27 ± 0.12	89.21 ± 0.09	85.64 ± 0.19	78.86 ± 0.66	**52.92 ± 1.20**
	AGCE	88.95 ± 0.22	86.98 ± 0.12	83.39 ± 0.17	76.49 ± 0.53	44.42 ± 0.74
	AEL	86.38 ± 0.19	84.27 ± 0.12	81.12 ± 0.20	74.86 ± 0.22	51.41 ± 0.32
	NCE+AUL	91.10 ± 0.13	**89.31 ± 0.20**	**86.23 ± 0.18**	**79.70 ± 0.08**	**59.44 ± 1.14**
	NCE+AGCE	90.94 ± 0.12	**89.21 ± 0.08**	**86.19 ± 0.15**	**80.13 ± 0.18**	50.82 ± 1.46
	NCE+AEL	90.71 ± 0.04	88.57 ± 0.14	85.01 ± 0.38	77.33 ± 0.18	47.90 ± 1.21
CIFAR100	CE	71.33 ± 0.43	56.51 ± 0.39	39.92 ± 0.10	21.39 ± 1.17	7.59 ± 0.20
	FL	70.06 ± 0.70	55.78 ± 1.55	39.83 ± 0.43	21.91 ± 0.89	7.51 ± 0.09
	GCE	63.09 ± 1.39	61.57 ± 1.06	56.11 ± 1.35	45.28 ± 0.61	17.42 ± 0.06
	SCE	69.62 ± 0.42	52.25 ± 0.14	36.00 ± 0.69	20.14 ± 0.60	7.67 ± 0.63
	NLNL	68.72 ± 0.60	46.99 ± 0.91	30.29 ± 1.64	16.60 ± 0.90	11.01 ± 2.48
	NCE	29.96 ± 0.73	25.27 ± 0.32	19.54 ± 0.52	13.51 ± 0.65	8.55 ± 0.37
	NCE+RCE	68.65 ± 0.40	64.97 ± 0.49	58.54 ± 0.13	45.80 ± 1.02	**25.41 ± 0.98**
	NCE+AUL	68.96 ± 0.16	**65.36 ± 0.20**	**59.25 ± 0.23**	**46.34 ± 0.21**	23.03 ± 0.64
	NCE+AGCE	69.03 ± 0.37	**65.66 ± 0.46**	**59.47 ± 0.36**	**48.02 ± 0.58**	**24.72 ± 0.60**
	NCE+AEL	68.70 ± 0.20	**65.36 ± 0.14**	**59.51 ± 0.03**	**46.94 ± 0.07**	**24.48 ± 0.24**

Table 2. Test accuracies (%) of different methods on benchmark datasets with asymmetric label noise ($\eta \in [0.1, 0.2, 0.3, 0.4]$). The sults (mean±std) are reported over 3 random runs and the top 3 best results are **boldfaced**.

Datasets	Methods	Asymmetric Noise Rate (η) 0.1	0.2	0.3	0.4
MNIST	CE	97.57 ± 0.22	94.56 ± 0.22	88.81 ± 0.10	82.27 ± 0.40
	FL	97.58 ± 0.09	94.25 ± 0.15	89.09 ± 0.25	82.13 ± 0.49
	GCE	99.01 ± 0.04	96.69 ± 0.12	89.12 ± 0.24	81.51 ± 0.19
	NLNL	98.63 ± 0.06	98.35 ± 0.01	97.51 ± 0.15	95.84 ± 0.26
	SCE	**99.14 ± 0.04**	98.03 ± 0.05	93.68 ± 0.43	85.36 ± 0.17
	NCE	98.49 ± 0.06	98.18 ± 0.12	96.99 ± 0.17	94.16 ± 0.19
	NCE+RCE	**99.35 ± 0.03**	98.99 ± 0.22	97.23 ± 0.20	90.49 ± 4.04
	AUL	**99.15 ± 0.09**	**99.15 ± 0.02**	**98.98 ± 0.05**	**98.62 ± 0.09**
	AGCE	99.10 ± 0.02	**99.07 ± 0.09**	**98.95 ± 0.03**	**98.44 ± 0.11**
	AEL	98.99 ± 0.05	**99.06 ± 0.07**	**98.90 ± 0.15**	**98.34 ± 0.08**
CIFAR10	CE	87.55 ± 0.14	83.32 ± 0.12	79.32 ± 0.59	74.67 ± 0.38
	FL	86.43 ± 0.30	83.37 ± 0.07	79.33 ± 0.08	74.28 ± 0.44
	GCE	88.33 ± 0.05	85.93 ± 0.23	80.88 ± 0.38	74.29 ± 0.43
	SCE	89.77 ± 0.11	86.20 ± 0.37	81.38 ± 0.35	75.16 ± 0.39
	NLNL	88.54 ± 0.25	84.74 ± 0.08	81.26± 0.43	76.97 ± 0.52
	NCE	74.06 ± 0.27	72.46 ± 0.32	69.86 ± 0.51	65.66 ± 0.42
	NCE+RCE	**90.06 ± 0.13**	**88.45 ± 0.16**	**85.42 ± 0.09**	**79.33 ± 0.15**
	AUL	**90.19 ± 0.16**	88.17 ± 0.11	84.87 ± 0.04	56.33 ± 0.07
	AGCE	88.08 ± 0.06	86.67 ± 0.14	83.59 ± 0.15	60.91 ± 0.20
	AEL	85.22 ± 0.15	83.82 ± 0.15	82.43 ± 0.16	58.81 ± 3.62
	NCE+AUL	90.05 ± 0.20	**88.72 ± 0.26**	**85.48 ± 0.18**	**79.26 ± 0.05**
	NCE+AGCE	**90.35 ± 0.15**	**88.48 ± 0.16**	**85.96 ± 0.24**	**80.00 ± 0.44**
	NCE+AEL	89.95 ± 0.04	87.93 ± 0.06	84.81 ± 0.26	77.27 ± 0.11
CIFAR100	CE	64.85 ± 0.37	58.11 ± 0.32	50.68 ± 0.55	40.17 ± 1.31
	FL	64.78 ± 0.50	58.05 ± 0.42	51.15 ± 0.84	**41.18 ± 0.68**
	GCE	63.01 ± 1.01	59.35 ± 1.10	53.83 ± 0.64	40.91 ± 0.57
	SCE	61.63 ± 0.84	53.81 ± 0.42	45.63 ± 0.07	36.43 ± 0.20
	NLNL	59.55 ± 1.22	50.19 ± 0.56	42.81 ± 1.13	35.10 ± 0.20
	NCE	27.59 ± 0.54	25.75 ± 0.50	24.28 ± 0.80	20.64 ± 0.40
	NCE+RCE	66.38 ± 0.16	**62.97 ± 0.24**	**55.38 ± 0.49**	**41.68 ± 0.56**
	NCE+AUL	**66.62 ± 0.09**	**63.86 ± 0.18**	50.38 ± 0.32	38.59 ± 0.48
	NCE+AGCE	**67.22 ± 0.12**	**63.69 ± 0.19**	**55.93 ± 0.38**	**43.76 ± 0.70**
	NCE+AEL	**66.92 ± 0.22**	62.50 ± 0.23	**52.42 ± 0.98**	39.99 ± 0.12

Table 3. Top-1 validation accuracies (%) on WebVision validation set using different loss functions.

Loss	CE	GCE	SCE	NCE+RCE	**NCE+AGCE**	**AGCE**
Acc	66.96	61.76	66.92	66.32	**67.12**	**69.40**

Thanks for your attention!

Any question? Please contact us!

Xiong Zhou: cszx@hit.edu.cn
Xianming Liu: csxm@hit.edu.cn