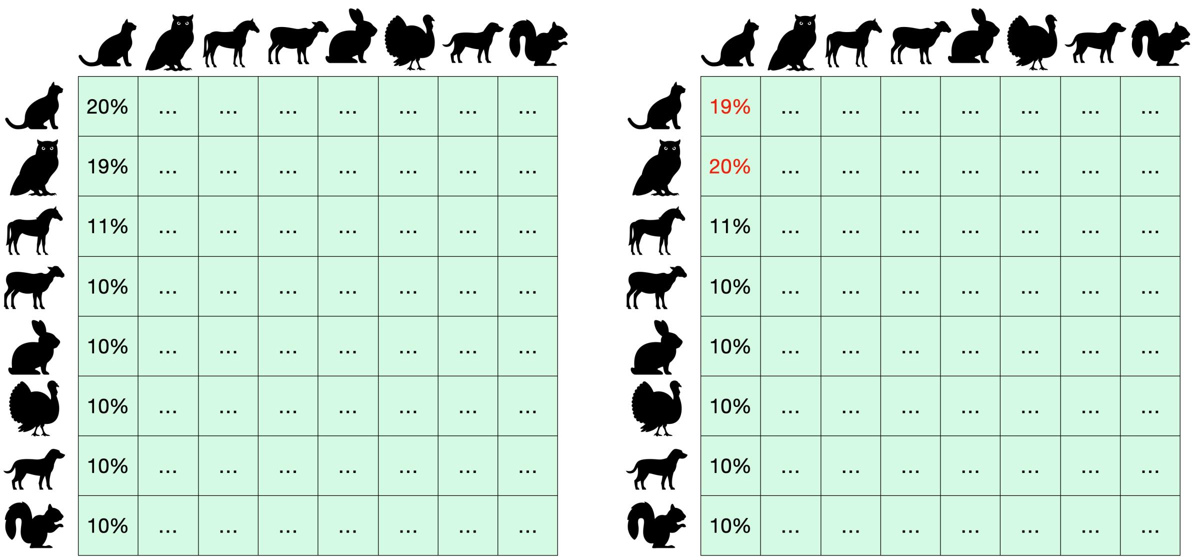
Asymmetric Loss Functions for Learning with Noisy Labels: Supplementary Materials

A. More Analysis about Clean Labels Domination Assumption

For robust training, we assume that samples in the training dataset have bigger probability of keeping their true semantic label than wrong class labels, which is referred to as *clean labels domination assumption*. In the following, we provide more intuitive analysis about this assumption to show its reasonability.



(a) Class "cat" dominates

(b) Class "cat" does not dominate

Figure 1. Illustration of label noise model under clean-labels-dominate and -non-dominate settings.

In Figure 1, label noise models under clean-labels-dominate and -non-dominate setting are shown, from which an intuitive understanding about the clean labels domination assumption can be derived. Figure 1(a) and 1(b) exhibit noise transmission matrices, which denote the probability of flipping the class of columns to the class of rows. In Figure 1(a), *cats* have a 20% probability of keeping the true label, while having smaller probability of wrongly flipping to labels of any other classes. For example, they have a 19% probability to be annotated as *owls*. In this case, we call *cats* clean-labels-dominant, since images with true cats labels dominate in the cats class, a classifier can be learned to correctly separate *cats* from other classes by classifying a sample to the dominant class . In Figure 1(b), the situation is reversed, where *cats* have bigger probability of flipping to *owls* than keeping the true label, which is denoted as the case of clean-labels-non-dominate. It means that *owls* account for the largest proportion in *cats* class, which sounds ridiculous. On the other hand, without the help of prior knowledge, even if there exists a learned classifier that works well in a clean-labels-non-dominant dataset, it would produce wrong results on a clean-labels-dominant dataset since it tends to classify a sample into a non-dominant class rather than the corresponding dominant class (*i.e.*, the true class).

B. Classification-calibration and Excess Risk Bound

In the binary classification problem with label set $\{0, 1\}$ which is different from $\{-1, 1\}$, we need to slightly modified the definition of classification-calibration in (Tong, 2003; Bartlett et al., 2006).

Let $f(\mathbf{x})$ denote the predictive result of $p(y=1|\mathbf{x})$, and $R_\ell(f)$ denote the risk of a classifier f based on a loss function ℓ or the ℓ-risk, i.e., $R_\ell(f) = \mathbb{E}\ell(f(\mathbf{x}), y)$. And the risk of a global minimizer is $R_\ell^* = \inf_f R_\ell(f)$.

For the zero-one loss ℓ_{0-1}, we have $R_{\ell_{0-1}}(f) = \mathbb{E}[\mathbb{I}(\text{sign}(f(\mathbf{x}) - 1/2) \neq \text{sign}(y - 1/2))]$. R^* denote the Bayes risk, i.e.,

$$R^*_{\ell_{0-1}}(f) = \inf_f R_{\ell_{0-1}}(f) \tag{1}$$

Given a loss function $\ell(t)$ (eg., exponential loss, cross entropy loss, or unhinged loss), where $t = yf(\mathbf{x}) + (1-y)(1-f(\mathbf{x}))$ is the predictive probability of data point (x, y), the *conditional* $\ell-$*risk* is defined as

$$C_{\eta_\mathbf{x}}(f(\mathbf{x}), \ell) = \mathbb{E}_{y|\mathbf{x}}[\ell(yf(\mathbf{x}) + (1-y)(1-f(\mathbf{x}))] = \eta_\mathbf{x}\ell(f(\mathbf{x})) + (1-\eta_\mathbf{x})\ell(1-f(\mathbf{x})) \tag{2}$$

where $\eta_\mathbf{x} = p(y=1|\mathbf{x})$. Similarly, we define the "optimal ℓ-risk" as

$$R_\ell^* = \inf_f R_\ell(f) = \inf_f \mathbb{E}[\eta_\mathbf{x}\ell(f(\mathbf{x})) + (1-\eta_\mathbf{x})\ell(1-f(\mathbf{x}))] \tag{3}$$

When ℓ is the *zero-one* loss, we obtain the Bayes-optimal classifier $\mathbb{I}(\eta_\mathbf{x} > \frac{1}{2})$.

The *excess risk* for a classifier f is given by $R_{\ell_{0-1}}(f) - R^*_{\ell_{0-1}}$, and the "excess ℓ-risk" is $R_\ell(f) - R_\ell^*$.

For a fixed value of $\mathbf{x}$, the minimum of the expectation is given by

$$H_\ell(\eta) = \inf_{\alpha\in[0,1]} (\eta\ell(\alpha) + (1-\eta)\ell(1-\alpha)), \tag{4}$$

so we write

$$R_\ell^* = \mathbb{E}[H_\ell(\eta_\mathbf{x})]. \tag{5}$$

For a good classifier, we want $\text{sign}(f(\mathbf{x}) - \frac{1}{2}) = \text{sign}(\alpha - \frac{1}{2}) = \text{sign}(\eta - \frac{1}{2})$, i.e., $(\alpha - \frac{1}{2})(\eta - \frac{1}{2}) \geq 0$. So we define a quantity similar to Eq. 4 but optimized only where α is not a good classifier:

$$H_\ell^-(\eta) = \inf_{\{\alpha:(\alpha-\frac{1}{2})(\eta-\frac{1}{2})\leq 0, \alpha\in[0,1]\}} (\eta\ell(\alpha) + (1-\eta)\ell(1-\alpha)). \tag{6}$$

We define a loss function ℓ to be "classification-calibrated" if $H_\ell^-(\eta) > H_\ell(\eta)$ for all $\eta \neq \frac{1}{2}$. Intuitively, this means that the loss function strictly penalizes a classifier f for not classifying in accordance with $\eta_\mathbf{x}$.

B.1. Classification-calibration

Theorem 1. *Completely asymmetric loss functions are classification-calibrated.*

Proof. For any weights w_1, w_2 and $w_1 \neq w_2$, we define a completely asymmetric loss function ℓ as follows

$$\arg\min_{u\in[0,1]} w_1\ell(u) + w_2\ell(1-u) = \mathbb{I}[w_1 > w_2], \tag{7}$$

i.e., $w_1\ell(u) + w_2\ell(1-u) \geq \mathbb{I}[w_1 > w_2] \cdot [w_1\ell(1) + w_2\ell(0)] + \mathbb{I}[w_1 < w_2] \cdot [w_1\ell(0) + w_2\ell(1)]$, and the equality holds if and only if $u = \mathbb{I}(w_1 > w_2)$. In other words, the **conditional risk minimizer** of ℓ can be expressed as $\mathbb{I}(\eta_\mathbf{x} > 1 - \eta_\mathbf{x})$, which is equivalent to the Bayes-optimal classifier $\mathbb{I}(\eta_\mathbf{x} > \frac{1}{2})$.

Then if ℓ is asymmetric on $\eta, 1-\eta$, where $\eta \neq \frac{1}{2}$, we have

$$H_\ell(\eta) = \inf_{\alpha\in[0,1]} (\eta\ell(\alpha) + (1-\eta)\ell(1-\alpha) = \begin{cases} \eta\ell(1) + (1-\eta)\ell(0), & \eta > \frac{1}{2} \\ \eta\ell(0) + (1-\eta)\ell(1), & \eta < \frac{1}{2} \end{cases} \tag{8}$$

and

$$H_\ell^-(\eta) = \begin{cases} \inf_{0\leq\alpha\leq\frac{1}{2}} (\eta\ell(\alpha) + (1-\eta)\ell(1-\alpha)), & \eta > \frac{1}{2} \\ \inf_{\frac{1}{2}\leq\alpha\leq 1} (\eta\ell(\alpha) + (1-\eta)\ell(1-\alpha)), & \eta < \frac{1}{2} \end{cases}. \tag{9}$$

Because ℓ is asymmetric on $\eta, 1-\eta$, then for all $\eta > \frac{1}{2}$, we have

$$\inf_{0\leq\alpha\leq\frac{1}{2}} (\eta\ell(\alpha) + (1-\eta)\ell(1-\alpha)) > \eta\ell(1) + (1-\eta)\ell(0), \tag{10}$$

and for all $\eta < \frac{1}{2}$,

$$\inf_{\frac{1}{2}\leq\alpha\leq 1} (\eta\ell(\alpha) + (1-\eta)\ell(1-\alpha)) > \eta\ell(0) + (1-\eta)\ell(1). \tag{11}$$

so it follows that $H_\ell^-(\eta) > H_\ell(\eta)$ for all $\eta \neq \frac{1}{2}$, so asymmetric loss functions are classification-calibrated. $\square$

B.2. Excess Risk Bound

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

$$R_{\ell_{0-1}}(f) - R^*_{\ell_{0-1}} \leq \frac{2(R_\ell(f) - R^*_\ell)}{\ell(0) - \ell(1)}, \tag{12}$$

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

Proof. Consider a loss function ℓ, the transform $\tilde{\psi} : [-1, 1] \to R_+$ from (Bartlett et al., 2006) is defined as

$$\tilde{\psi}(\theta) = H_\ell^-\left(\frac{1+\theta}{2}\right) - H_\ell\left(\frac{1+\theta}{2}\right) \tag{13}$$

For $\theta \in (0, 1]$, we have

$$\begin{aligned}\tilde{\psi}(\theta) &= H_\ell^-\left(\frac{1+\theta}{2}\right) - H_\ell\left(\frac{1+\theta}{2}\right) \\ &= \inf_{0\leq\alpha\leq\frac{1}{2}} \left[\frac{1+\theta}{2}\ell(\alpha) + \frac{1-\theta}{2}\ell(1-\alpha)\right] - \left[\frac{1+\theta}{2}\ell(1) + \frac{1-\theta}{2}\ell(0)\right] \\ &= \frac{1}{2}[2\ell(1/2) - \ell(0) - \ell(1)] + \frac{\theta}{2}[\ell(0) - \ell(1)].\end{aligned} \tag{14}$$

where $\frac{1+\theta}{2}\ell(\alpha) + \frac{1-\theta}{2}\ell(1-\alpha) \geq \frac{1+\theta}{2}\ell(1/2) + \frac{1-\theta}{2}\ell(1/2)$, for $\alpha \in [0, 1/2]$, since ℓ is strictly asymmetric.

For $\theta \in [-1, 0)$, we have

$$\begin{aligned}\tilde{\psi}(\theta) &= H_\ell^-\left(\frac{1+\theta}{2}\right) - H_\ell\left(\frac{1+\theta}{2}\right) \\ &= \inf_{\frac{1}{2}\leq\alpha\leq 1} \left[\frac{1+\theta}{2}\ell(\alpha) + \frac{1-\theta}{2}\ell(1-\alpha)\right] - \left[\frac{1+\theta}{2}\ell(0) + \frac{1-\theta}{2}\ell(1)\right] \\ &= \frac{1}{2}[2\ell(1/2) - \ell(0) - \ell(1)] - \frac{\theta}{2}[\ell(0) - \ell(1)].\end{aligned} \tag{15}$$

where $\frac{1+\theta}{2}\ell(\alpha) + \frac{1-\theta}{2}\ell(1-\alpha) \geq \frac{1+\theta}{2}\ell(1/2) + \frac{1-\theta}{2}\ell(1/2)$, for $\alpha \in [1/2, 1]$, since ℓ is strictly asymmetric.

We can see that $\tilde{\psi}$ is symmetric about 0, i.e., $\tilde{\psi}(-t) = \tilde{\psi}(t)$, and $\tilde{\psi}(0) = \frac{1}{2}[2\ell(1/2) - \ell(0) - \ell(1)] \geq 0$. Therefore, $\tilde{\psi}(\theta)$ is convex. For simplicity, let $\sigma(t) = (t - 1/2)$. Then, according to Jensen's inequality, we have

$$
\begin{aligned}
&\tilde{\psi}(R_{\ell_{0-1}}(f) - R^*_{\ell_{0-1}}) \\
&= \tilde{\psi}\left(\mathbb{E}[\mathbb{I}(\sigma(f(\mathbf{x})) \neq \sigma(\eta_{\mathbf{x}})|2\eta_{\mathbf{x}} - 1|]\right) \\
&\leq \mathbb{E}[\tilde{\psi}\left(\mathbb{I}(\sigma(f(\mathbf{x})) \neq \sigma(\eta_{\mathbf{x}})|2\eta_{\mathbf{x}} - 1|\right)] \\
&= \mathbb{E}\left[\mathbb{I}(\sigma(f(\mathbf{x})) = \sigma(\eta_{\mathbf{x}})) \cdot \tilde{\psi}(0)\right] + \mathbb{E}\left[\mathbb{I}(\sigma(f(\mathbf{x})) \neq \sigma(\eta_{\mathbf{x}})) \cdot \tilde{\psi}(|2\eta_{\mathbf{x}} - 1|)\right] \\
&\leq \tilde{\psi}(0) + \mathbb{E}\left[\mathbb{I}(\sigma(f(\mathbf{x})) \neq \sigma(\eta_{\mathbf{x}})) \cdot \left(H^-_{\ell}(\eta_{\mathbf{x}}) - H_{\ell}(\eta_{\mathbf{x}})\right)\right] \\
&= \tilde{\psi}(0) + \mathbb{E}\left[\mathbb{I}(\sigma(f(\mathbf{x})) \neq \sigma(\eta_{\mathbf{x}})) \cdot \left(\inf_{\{\alpha:(\alpha-\frac{1}{2})(\eta_{\mathbf{x}}-\frac{1}{2})\leq 0, \alpha\in[0,1]\}} C_{\eta_{\mathbf{x}}}(\alpha, \ell) - H_{\ell}(\eta_{\mathbf{x}})\right)\right] \\
&\leq \tilde{\psi}(0) + \mathbb{E}\left[\mathbb{I}(\sigma(f(\mathbf{x})) \neq \sigma(\eta_{\mathbf{x}})) \cdot (C_{\eta_{\mathbf{x}}}(f(\mathbf{x}), \ell) - H_{\ell}(\eta_{\mathbf{x}}))\right] \\
&\leq \tilde{\psi}(0) + \mathbb{E}\left[\mathbb{I}(\sigma(f(\mathbf{x})) \neq \sigma(\eta_{\mathbf{x}})) \cdot (C_{\eta_{\mathbf{x}}}(f(\mathbf{x}), \ell) - H_{\ell}(\eta_{\mathbf{x}}))\right] + \mathbb{E}\left[\mathbb{I}(\sigma(f(\mathbf{x})) = \sigma(\eta_{\mathbf{x}})) \cdot (C_{\eta_{\mathbf{x}}}(f(\mathbf{x}), \ell) - H_{\ell}(\eta_{\mathbf{x}}))\right] \\
&= \tilde{\psi}(0) + \mathbb{E}\left[C_{\eta_{\mathbf{x}}}(f(\mathbf{x}), \ell) - H_{\ell}(\eta_{\mathbf{x}})\right] \\
&= \tilde{\psi}(0) + R_{\ell}(f) - R^*_{\ell},
\end{aligned}
$$

where we have used the fact that for any $\mathbf{x}$, and in particular when $\text{sign}(f(\mathbf{x}) - 1/2) = \text{sign}(\eta_{\mathbf{x}} - 1/2)$, $C_{\eta_{\mathbf{x}}}(f(\mathbf{x}), \ell) \geq H_{\ell}(\eta_{\mathbf{x}})$. On the other hand, since $\tilde{\psi}(\theta) = \tilde{\psi}(0) + \frac{|\theta|}{2}[\ell(0) - \ell(1)]$, we have

$$
\tilde{\psi}(0) + \frac{R_{\ell_{0-1}}(f) - R^*_{\ell_{0-1}}}{2}[\ell(0) - \ell(1)] = \tilde{\psi}(R_{\ell_{0-1}}(f) - R^*_{\ell_{0-1}}) \leq \tilde{\psi}(0) + R_{\ell}(f) - R^*_{\ell}, \tag{16}
$$

i.e., we obtain the excess risk bound as follows

$$
R_{\ell_{0-1}}(f) - R^*_{\ell_{0-1}} \leq \frac{2(R_{\ell}(f) - R^*_{\ell})}{\ell(0) - \ell(1)}. \tag{17}
$$

The result suggests that the excess risk bound of any completely asymmetric loss function is controlled only by the difference of $\ell(0) - \ell(1)$. Intuitively, the excess risk bound suggests that if the prediction function f minimizes the surrogate risk $R_{\ell}(f) = R^*_{\ell}$, then the prediction function f must also minimize the misclassification risk $R_{\ell_{0-1}}(f) = R^*_{\ell_{0-1}}$. □

C. Proof for Theorems and Corollaries

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

Proof. For any weights $w_1, ..., w_k$, $\exists t$, s.t., $w_t > \max_{i\neq t} w_i$, i.e., $w_i - w_t < 0$. Let L be a symmetric loss function, then

$$
\begin{aligned}
\sum_{i=1}^{k} w_i L(\mathbf{u}, i) &= w_t L(\mathbf{u}, t) + \sum_{i\neq t} w_i L(\mathbf{u}, i) \\
&= w_t C + \sum_{i\neq t}(w_i - w_t) L(\mathbf{u}, i) \\
&\geq w_t C + \min_{\mathbf{u}\in U'} \sum_{i\neq t}(w_i - w_t) L(\mathbf{u}, i)
\end{aligned} \tag{18}
$$

where $U' = \{\mathbf{u} : \sum_{i\neq t} L(\mathbf{u}, i) = C - \min_{\mathbf{u}} L(\mathbf{u}, t)\} = \{\arg\min_{\mathbf{u}} L(\mathbf{u}, t)\}$. Therefore, $\arg\min_{\mathbf{u}} \sum_{i=1}^{k} w_i L(\mathbf{u}, i) = \arg\min_{\mathbf{u}} L(\mathbf{u}, t)$, i.e., L is a completely asymmetric loss function. □

C.1. Proof for theorems

Theorem 4 (Noise-Tolerance)**.** *In a multi-classification problem, given an appropriate neural network class $\mathcal{H}$ which satisfies Assumption 7, then the loss function L is noise-tolerant if L is asymmetric on the label noise model.*

Proof. Let $f^* = \arg\min_{f\in\mathcal{H}} R_L^\eta(f)$, when we regard the conditional risk $L^\eta(\mathbf{x}, y)$ as a new loss function, then f^* minimizes $L^\eta(f(\mathbf{x}), y)$ for each $(\mathbf{x}, y)$. Because L is an asymmetric loss and $1-\eta_{\mathbf{x}}$ is bigger than $\eta_{\mathbf{x},i}$, f^* also minimizes $L(f(\mathbf{x}), y)$. Therefore, we have

$$R_L(f) = \mathbb{E}_{\mathbf{x},y} L(f(\mathbf{x}), y) \geq \mathbb{E}_{\mathbf{x},y} L(f^*(\mathbf{x}), y) = R_L(f^*),$$

so f^* minimizes $R_L(f)$. □

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

Proof. Given weights $w_1, ..., w_k$, $w_t > \max_{i\neq t} w_i$, because L_1 and L_2 are asymmetric, let $\mathbf{u}^* = \mathbf{e}_t = \arg\min_{\mathbf{u}} L_1(\mathbf{u}, t) = \arg\min_{\mathbf{u}} L_2(\mathbf{u}, t)$, i.e.,

$$\begin{aligned} &\sum_{i=1}^k w_i L_1(\mathbf{u}, i) \geq \sum_{i=1}^k w_i L_1(\mathbf{u}^*, i) \quad \text{and} \\ &\sum_{i=1}^k w_i L_2(\mathbf{u}, i) \geq \sum_{i=1}^k w_i L_2(\mathbf{u}^*, i) \end{aligned} \tag{19}$$

Then we have $\sum_{i=1}^k w_i[\alpha L_1(\mathbf{u}, i) + \beta L_2(\mathbf{u}, i)] \geq \sum_{i=1}^k w_i(\alpha L_1(\mathbf{u}^*, i) + \beta L_2(\mathbf{u}^*, i)]$, and the equality holds if and only if $\mathbf{u} = \mathbf{u}^*$, so $\alpha L_1 + \beta L_2$ is asymmetric. □

Lemma 1. *Consider a loss function $L(\mathbf{u}, i) = \ell(u_i)$, for any $w_1 > w_2 \geq 0$, $\mathbf{u} \in \mathcal{C}$, if ℓ satisfies $w_1\ell(u_1) + w_2\ell(u_2) \geq w_1\ell(u_1 + u_2) + w_2\ell(0)$, and the equality holds only if $u_2 = 0$, then L is completely asymmetric.*

Proof. Given any weights $w_1, ..., w_k$, $w_t > \max_{i\neq t} w_i$, the optimal solution is $\mathbf{u}^* = \mathbf{e}_t$, then

$$\begin{aligned} \sum_{i=1}^k w_i L(\mathbf{u}, i) &= w_t\ell(u_t) + \sum_{i\neq t} w_i \ell(u_i) \\ &\geq w_t \ell(u_t + \sum_{i\neq t} u_i) + \sum_{i\neq t} w_i \ell(0) \\ &= \sum_{i=1}^k w_i L(\mathbf{u}^*, i) \end{aligned} \tag{20}$$

The equality holds if and only if $u_i = 0$, for $i \neq t$, i.e., $\mathbf{u}^*$ is the only one minimizes $\sum_{i=1}^k w_i L(\mathbf{u}, i)$, so L is completely asymmetric. □

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

Proof. If $\frac{w_m}{w_n} \cdot r(\ell) \geq 1$, then for any $i \neq m$, we have

$$\frac{w_m}{w_i} \geq \frac{1}{r(\ell)} \geq \sup_{\substack{0\leq u_m, u_i \leq 1 \\ u_m + u_i \leq 1}} \frac{\ell(0) - \ell(u_i)}{\ell(u_m) - \ell(u_m + u_i)} \geq \frac{\ell(0) - \ell(u_i)}{\ell(u_m) - \ell(u_m + u_i)} \tag{21}$$

i.e., $w_m\ell(u_n) + w_i\ell(u_i) \geq w_m\ell(u_m + u_i) + w_i\ell(0)$, so L is asymmetric according to Theorem 1. □

Theorem 7. *In a binary classification problem, we assume that L is strictly asymmetric on the label noise model which keeps dominant, for any $\mathcal{H}$, let $f^* = \arg\min_{f\in\mathcal{H}} R_L^\eta(f)$. If $\forall \mathbf{x}$, $\frac{1-\eta_{\mathbf{x}}}{\eta_{\mathbf{x}}} \cdot r(L) > 1$ hold, then f^* also minimizes a positive weighted L-risk $R_{w,L}(h) = \mathbb{E}w(\mathbf{x}, y)L(f(\mathbf{x}), y)$.*

Proof. Without loss of generality, let the label set be $\{0,1\}$, and $f^* = \arg\min_{f\in\mathcal{H}} R_L^\eta(f)$, then we have

$$
\begin{aligned}
&R_L^\eta(f^*) - R_L^\eta(f)\\
=&\mathbb{E}_{\mathbf{x},y}\Big[(1-\eta_{\mathbf{x}})\big[L(f^*(\mathbf{x}),y) - L(f(\mathbf{x}),y)\big] + \eta_{\mathbf{x}}\big[L(f^*(\mathbf{x}),1-y) - L(f(\mathbf{x}),1-y)\big]\Big]\\
=&\mathbb{E}_{\mathbf{x},y}\mathbb{I}(f^*(\mathbf{x})_y < f(\mathbf{x})_y)\Big[(1-\eta_{\mathbf{x}})\big[L(f^*(\mathbf{x}),y) - L(f(\mathbf{x}),y)\big] + \eta_{\mathbf{x}}\big[L(f^*(\mathbf{x}),1-y) - L(f(\mathbf{x}),1-y)\big]\Big]+\\
&\mathbb{E}_{\mathbf{x},y}\mathbb{I}(f^*(\mathbf{x})_y > f(\mathbf{x})_y)\Big[(1-\eta_{\mathbf{x}})\big[L(f^*(\mathbf{x}),y) - L(f(\mathbf{x}),y)\big] + \eta_{\mathbf{x}}\big[L(f^*(\mathbf{x}),1-y) - L(f(\mathbf{x}),1-y)\big]\Big]\\
\geq&\mathbb{E}_{\mathbf{x},y}\mathbb{I}(f^*(\mathbf{x})_y < f(\mathbf{x})_y)\Big[(1-\eta_{\mathbf{x}})\big[L(f^*(\mathbf{x}),y) - L(f(\mathbf{x}),y)\big] - \frac{\eta_{\mathbf{x}}}{r(L)}\big[L(f^*(\mathbf{x}),y) - L(f(\mathbf{x}),y)\big]\Big]+\\
&\mathbb{E}_{\mathbf{x},y}\mathbb{I}(f^*(\mathbf{x})_y > f(\mathbf{x})_y)\Big[(1-\eta_{\mathbf{x}})\big[L(f^*(\mathbf{x}),y) - L(f(\mathbf{x}),y)\big] + \frac{\eta_{\mathbf{x}}}{r(L)}\big[L(f^*(\mathbf{x}),y) - L(f(\mathbf{x}),y)\big]\Big]\\
=&\mathbb{E}_{\mathbf{x},y}w(\mathbf{x},y)L(f^*(\mathbf{x}),y) - \mathbb{E}_{\mathbf{x},y}w(\mathbf{x},y)L(f(\mathbf{x}),y)
\end{aligned}
\tag{22}
$$

where we have

$$
L(f^*(\mathbf{x}),1-y) - L(f(\mathbf{x}),1-y) \geq \begin{cases} -\frac{1}{r(L)}\big[L(f^*(\mathbf{x}),y) - L(f(\mathbf{x}),y)\big], & f^*(\mathbf{x})_y < f(\mathbf{x})_y \\ \frac{1}{r(L)}\big[L(f^*(\mathbf{x}),y) - L(f(\mathbf{x}),y)\big], & f^*(\mathbf{x})_y > f(\mathbf{x})_y \end{cases}
\tag{23}
$$

and

$$
0 < w(\mathbf{x},y) = \begin{cases} (1-\eta_{\mathbf{x}} - \frac{\eta_{\mathbf{x}}}{r(L)}), & f^*(\mathbf{x})_y < f(\mathbf{x})_y \\ 1-\eta_{\mathbf{x}}, & f^*(\mathbf{x})_y = f(\mathbf{x})_y \\ (1-\eta_{\mathbf{x}} + \frac{\eta_{\mathbf{x}}}{r(L)}), & f^*(\mathbf{x})_y < f(\mathbf{x})_y \end{cases}
\tag{24}
$$

Otherwise, $R_L^\eta(f^*) - R_L^\eta(f) \leq 0$, so we obtain

$$
\mathbb{E}_{\mathbf{x},y}w(\mathbf{x},y)L(f^*(\mathbf{x}),y) \leq \mathbb{E}_{\mathbf{x},y}w(\mathbf{x},y)L(h(\mathbf{x}),y),
\tag{25}
$$

i.e., f^* also minimizes the positive weighted L-risk $\mathbb{E}_{\mathbf{x},y}w(\mathbf{x},y)L(f^*(\mathbf{x}),y)$.

□

Theorem 8 (Necessity)**.** *On the given weights $w_1,..,w_k$, where $w_m > w_n$ and $w_n = \max_{i\neq m} w_i$, the loss function $L(\mathbf{u},i) = \ell(u_i)$ is asymmetric only if $\frac{w_m}{w_n}\cdot r_u(\ell) \geq 1$.*

Proof. If loss function $L_q(\mathbf{u},i) = \ell(u_i)$ is asymmetric, then for $w_m > w_n$, let $u_i = 0$, $i \neq m,n$, then $w_m\ell(u_m) + w_n\ell(u_n) \geq w_m\ell(1) + w_n\ell(0)$ always holds, i.e.,

$$
\frac{w_m}{w_n}\cdot \inf_{\substack{0\leq u_m,u_n\leq 1\\ u_m+u_n=1}} \frac{\ell(u_m) - \ell(u_m+u_n)}{\ell(0) - \ell(u_n)} \geq 1,
\tag{26}
$$

so $\frac{w_m}{w_n}\cdot r_u(\ell) \geq 1$. □

C.2. Proof for corollaries

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

Proof. $\Rightarrow$ If loss function $L_q(\mathbf{u},i) = \ell(u_i)$ is asymmetric, then for $w_m > w_n$, let $u_i = 0$, $i \neq m,n$, then $w_m\ell(u_m) + w_n\ell(u_n) \geq w_m\ell(1) + w_n\ell(0)$ always holds, i.e.,

$$
w_m[(a+1)^q - (a+u_m)^q] \geq w_n[(a+u_n)^q - a^q].
\tag{27}
$$

$$
\frac{(a+u_1+\Delta u)^q - (a+u_1)^q}{(a+u_2)^q - (a+u_2-\Delta u)^q}
\tag{28}
$$

so we have
$$\frac{w_m}{w_n} \geq \sup_{0 \leq u \leq 1} \frac{(a+1-u)^q - a^q}{(a+1)^q - (a+u)^q}.$$
RHS equals to $(\frac{a+1}{a})^{1-q}$ if $q \leq 1$, and equals to 1 when $q > 1$.

$\Leftarrow$ According to Theorem 1, L is asymmetric

$$\begin{aligned}
&\Leftarrow w_m\ell(u_m) + w_i\ell(u_i) \geq w_m\ell(u_m + u_i) + w_i\ell(0) \\
&\Leftrightarrow \frac{w_m}{w_i} \geq \sup_{\substack{u_i, u_m \geq 0 \\ u_i + u_m \leq 1}} \frac{\ell(0) - \ell(u_i)}{\ell(u_m) - \ell(u_m + u_i)} \\
&\Leftrightarrow \frac{w_m}{w_i} \geq \sup_{\substack{u_i, u_m \geq 0 \\ u_i + u_m \leq 1}} \frac{(a+u_i)^q - a^q}{(a+u_i+u_m)^q - (a+u_m)^q} \\
&\Leftrightarrow \frac{w_m}{w_i} \geq \mathbb{I}(q \leq 1) \cdot \sup_{0 \leq u_m \leq 1} \left(\frac{a+u_m}{a}\right)^{1-q} + \mathbb{I}(q > 1) \\
&\Leftrightarrow \frac{w_m}{w_i} \geq \left(\frac{a+1}{a}\right)^{1-q} \cdot \mathbb{I}(q \leq 1) + \mathbb{I}(q > 1).
\end{aligned}$$

On the other hand, if $\frac{w_m}{w_n} \geq (\frac{a+1}{a})^{1-q} \cdot \mathbb{I}(q \leq 1) + \mathbb{I}(q > 1)$. Then for any $i \neq m$, we have $\frac{w_m}{w_i} \geq (\frac{a+1}{a})^{1-q} \cdot \mathbb{I}(q \leq 1) + \mathbb{I}(q > 1)$. $\square$

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

Proof. $\Rightarrow$ If $L_p(\mathbf{u}, i) = \ell(u_i)$ is asymmetric, then for $w_m > w_n \geq 0$, let $u_i = 0$, $i \neq m, n$, then $w_m\ell(u_m) + w_n\ell(u_n) \geq w_m\ell(1) + w_n\ell(0)$ always holds, i.e.,
$$w_m[(a-u_m)^p - (a-1)^p)] \geq w_n[a^p - (a-u_n)^p],$$
so we have
$$\frac{w_m}{w_n} \geq \sup_{0 \leq u \leq 1} \frac{a^p - (a-1+u)^p}{(a-u)^p - (a-1)^p}.$$
RHS equals to $(\frac{a}{a-1})^{p-1}$ if $p > 1$, and equals to 1 when $p \leq 1$.

$\Leftarrow$ According to Theorem 1, L is asymmetric

$$\begin{aligned}
&\Leftarrow w_m\ell(u_m) + w_i\ell(u_i) \geq w_m\ell(u_m + u_i) + w_i\ell(0) \\
&\Leftrightarrow \frac{w_n}{w_i} \geq \sup_{\substack{u_i, u_m \geq 0 \\ u_i + u_m \leq 1}} \frac{\ell(0) - \ell(u_i)}{\ell(u_m) - \ell(u_n + u_i)} \\
&\Leftrightarrow \frac{w_m}{w_i} \geq \sup_{\substack{u_i, u_m \geq 0 \\ u_i + u_m \leq 1}} \frac{a^p - (a-u_i)^p}{(a-u_m)^p - (a-u_i-u_m)^p} \\
&\Leftrightarrow \frac{w_m}{w_i} \geq \mathbb{I}(p > 1) \cdot \sup_{0 \leq u_m \leq 1} \left(\frac{a}{a-u_m}\right)^{p-1} + \mathbb{I}(p \leq 1) \\
&\Leftrightarrow \frac{w_m}{w_i} \geq \left(\frac{a}{a-1}\right)^{p-1} \cdot \mathbb{I}(p > 1) + \mathbb{I}(p \leq 1).
\end{aligned}$$

On the other hand, if $\frac{w_m}{w_n} \geq (\frac{a+1}{a})^{1-q} \cdot \mathbb{I}(q \leq 1) + \mathbb{I}(q > 1)$. Then for any $i \neq m$, we have $\frac{w_m}{w_i} \geq (\frac{a}{a-1})^{p-1} \cdot \mathbb{I}(p > 1) + \mathbb{I}(p \leq 1)$. $\square$

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

Proof. $\Rightarrow$ If $L_a(\mathbf{u}, i) = \ell(u_i)$ is asymmetric, then for $w_m > w_n \geq 0$, let $u_i = 0$, $i \neq m, n$, then $w_m\ell(u_m) + w_n\ell(u_n) \geq w_m\ell(u_m + u_n) + w_n\ell(0)$ always holds, i.e.,

$$w_m[\exp(\frac{-u_m}{a}) - \exp(\frac{-u_m - u_n}{a})] \geq w_n[1 - \exp(\frac{-u_n}{a})],$$

so we have

$$\frac{w_m}{w_n} \geq \exp\left(\frac{u_m}{a}\right) \Rightarrow a \geq \frac{1}{\ln w_m - \ln w_n}.$$

$\Leftarrow$ According to Theorem 1, L_a is asymmetric

$$\Leftarrow w_m\ell(u_m) + w_i\ell(u_i) \geq w_m\ell(u_m + u_i) + w_i\ell(0)$$
$$\Leftrightarrow \frac{w_m}{w_i} \geq \exp\left(\frac{u_m}{a}\right).$$

On the other hand, when $a \geq \frac{1}{\ln w_m - \ln w_n}$, then for any $i \neq m$, we have $\frac{w_m}{w_i} \geq \exp(1/a)$. $\square$

D. Experiments

D.1. Evaluation on Benchmark Datasets

Noise generation. The noisy labels are generated following standard approaches in previous works (Ma et al., 2020; Patrini et al., 2017). For symmetric noise, we corrupt the training labels by flipping labels in each class randomly to incorrect labels to other classes with flip probability $\eta \in \{0.2, 0.3, 0.6, 0.8\}$. For asymmetric noise, we flip the labels within a specific set of classes. For MNIST, flipping 7 $\to$ 1, 2 $\to$ 7, 5 $\leftrightarrow$ 6, 3 $\to$ 8. For CIFAR-10, flipping TRUCK $\to$ AUTOMOBILE, BIRD $\to$ AIRPLANE, DEER $\to$ HORSE, CAR $\leftrightarrow$ DOG. For CIFAR-100, the 100 classes are grouped into 20 super-classes with each has 5 sub-classes, and each class are flipped within the same super-class into the next in a circular fashion.

Networks and training. We follow the experimental settings in (Ma et al., 2020): 4-layer CNN for MNIST, an 8-layer CNN for CIFAR-10 and a ResNet-34 (He et al., 2016) for CIFAR-100. The networks are trained for 50, 120, 200 epochs for MNIST, CIFAR-10, CIFAR-100, respectively. For all the training, we use SGD optimizer with momentum 0.9 and cosine learning rate annealing. Weight decay is set to 1×10^{-3}, 1×10^{-4} and 1×10^{-5} for MNIST, CIFAR-10 and CIFAR-100, respectively. The initial learning rate is set to 0.01 for MNIST/CIFAR-10 and 0.1 for CIFAR-100. Batch size is set to 128. Typical data augmentations including random width/height shift and horizontal flip are applied.

Parameter settings.We set the parameter settings which match their original papers for all baseline methods. The details can be seen in Table 1.

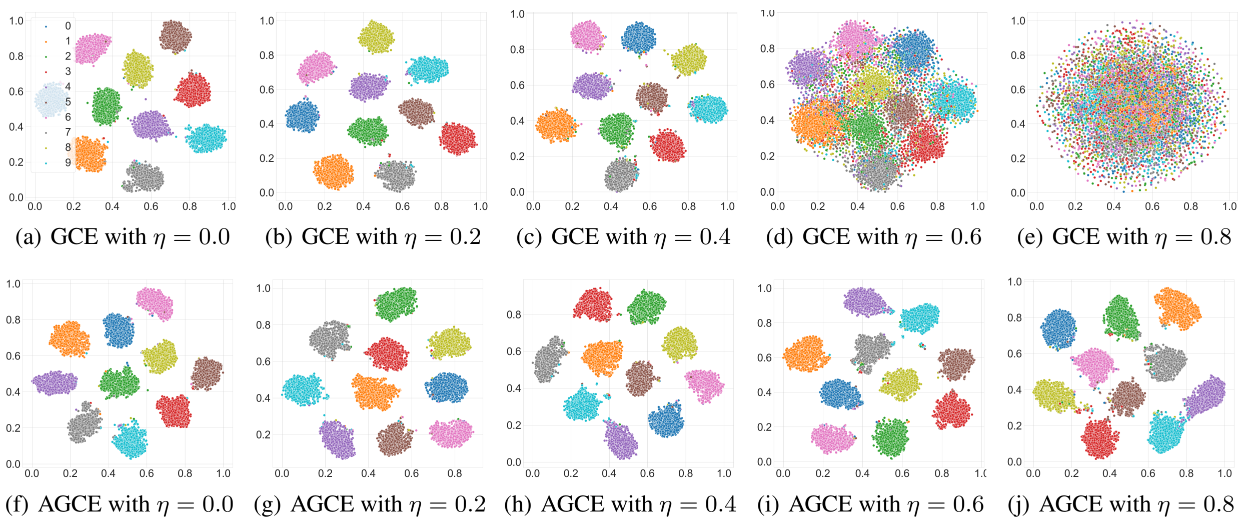
Table 1. Parameters settings for different methods.

Method	MNIST	CIFAR-10	CIFAR-100	WebVision
GCE& NGCE (q)	(0.7)	(0.7)	(0.7)	(0.7)
SCE (A, α, β)	(-4, 0.01, 1.0)	(-4, 0.1, 1.0)	(-4, 6.0, 1.0)	(-4, 10.0, 1.0)
FL& NFL (γ)	(0.5)	(0.5)	(0.5)	-
AGCE (a, q)	(4, 0.2)	(0.6, 0.6)	-	(1e-5,0.5)
AUL (a, p)	(3, 0.1)	(5.5, 3)	-	-
AEL (a)	(3.5)	(2.5)	-	-
NFL+RCE (A, α, β)	(-4, 1.0, 100.0)	(-4, 1.0, 1.0)	(-4, 10.0, 1.0)	-
NCE+MAE (α, β)	(1.0, 100.0)	(1.0, 1.0)	(10.0, 1.0)	-
NCE+RCE (α, β)	(1.0, 100.0)	(1.0, 1.0)	(10.0, 1.0)	(50.0, 0.1)
NCE+AGCE (a, q, α, β)	(4, 0.2, 0, 1)	(6, 1.5, 1, 4)	(1.8, 3, 10, 0.1)	(2.5, 3, 50, 0.1)
NCE+AUL (a, p, α, β)	(3, 0.1, 0, 1)	(6.3, 1.5, 1, 4)	(6, 3, 10, 0.015)	-
NCE+AEL (a, α, β)	(3.5, 0, 1)	(5, 1, 4)	(1.5, 10, 0.1)	-

Results. The experimental results of symmetric and asymmetric label noise are shown in Table 3 and Table 4, respectively. And we also visualize the learned features by the AGCE loss function and the GCE loss function. Figure 2 validates AGCE's ability of separating samples and robustness to label noise with any noise rate $\eta \in \{0.0, 0.2, 0.4, 0.6, 0.8\}$. Figure 3 validates different loss functions of separating samples and robustness to symmetric label noise with noise rates 0.0 and 0.4.

D.2. Evaluation on Real-world Noisy Labels

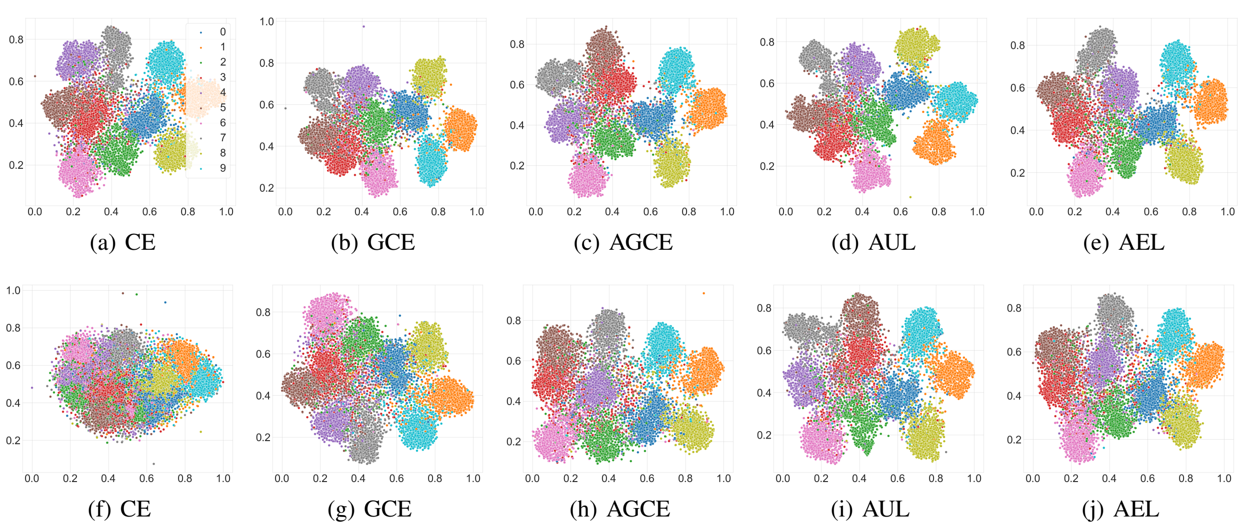
Here, we evaluate our asymmetric loss functions on large-scale real-world noisy dataset WebVision 1.0 (Li et al., 2017). It contains 2.4 million images of real-world noisy labels, crawled from the web using the 1,000 concepts in ImageNet ILSVRC12. Since the dataset is very big, for quick experiments, we follow the training setting in (Jiang et al., 2018; Ma et al., 2020) that only takes the first 50 classes of the Google resized image subset. We evaluate the trained networks on the same 50 classes of WebVision 1.0 validation set, which can be considered as a clean validation. ResNet-50 (He et al., 2016) is the model to be learnt. We compare our NCE+AGCE with GCE, SCE and NCE+RCE. The training details follow (Ma et al., 2020), where for each loss, we train a ResNet-50 (He et al., 2016) using SGD for 250 epochs with initial learning rate 0.4, nesterov momentum 0.9 and weight decay 3×10^{-5} and batch size 512. The learning rate is multiplied by 0.97 after every epoch of training. All the images are resized to 224×224. Typical data augmentations including random width/height shift, color jittering and random horizontal flip are applied. Experiments can be reported in Table 2.



(a) GCE with $\eta = 0.0$ (b) GCE with $\eta = 0.2$ (c) GCE with $\eta = 0.4$ (d) GCE with $\eta = 0.6$ (e) GCE with $\eta = 0.8$

(f) AGCE with $\eta = 0.0$ (g) AGCE with $\eta = 0.2$ (h) AGCE with $\eta = 0.4$ (i) AGCE with $\eta = 0.6$ (j) AGCE with $\eta = 0.8$

Figure 2. Visualization for GCE (top) and AGCE (bottom) on MNIST with different symmetric noise ($\eta \in [0.0, 0.2, 0.4, 0.6, 0.8]$) by t-SNE (Van der Maaten & Hinton, 2008) 2D embeddings of deep features.



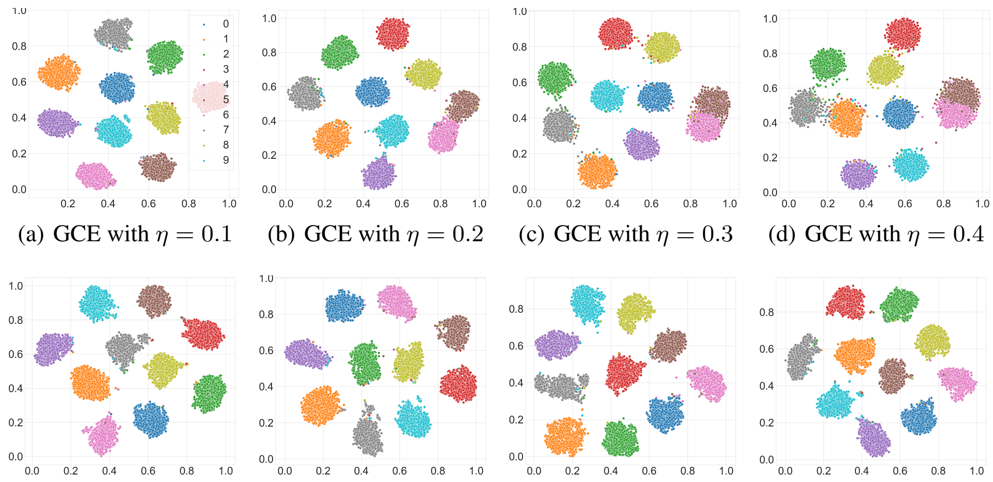
(a) CE (b) GCE (c) AGCE (d) AUL (e) AEL

(f) CE (g) GCE (h) AGCE (i) AUL (j) AEL

Figure 3. Visualization for CE, GCE, AGCE, AUL, and AEL on CIFAR10 with different symmetric noise (0.0 for top, 0.4 for bottom) by t-SNE (Van der Maaten & Hinton, 2008) 2D embeddings of deep features.

Table 2. Top-1 validation accuracies (%) on WebVision validatoin set of ResNet-50 models trained on WebVision using different loss functions, under the Mini setting in (Jiang et al., 2018; Ma et al., 2020).

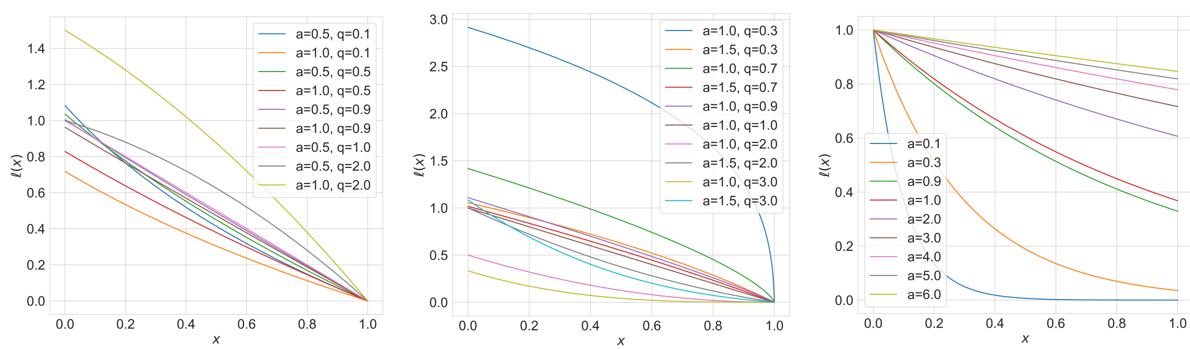
Loss	CE	GCE	SCE	NCE+RCE	**NCE+AGCE**	**AGCE**
Acc	66.96	61.76	66.92	66.32	**67.12**	**69.40**



(a) GCE with $\eta = 0.1$ (b) GCE with $\eta = 0.2$ (c) GCE with $\eta = 0.3$ (d) GCE with $\eta = 0.4$

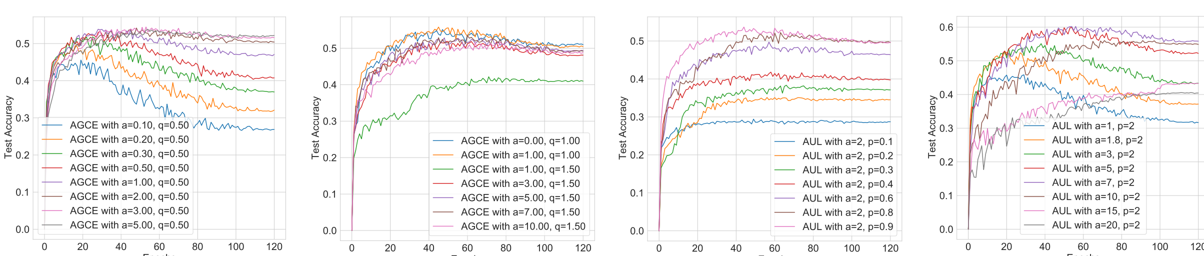
(e) AGCE with $\eta = 0.1$ (f) AGCE with $\eta = 0.2$ (g) AGCE with $\eta = 0.3$ (h) AGCE with $\eta = 0.4$

Figure 4. Visualization for GCE (top) and AGCE (bottom) on MNIST with different asymmetric noise ($\eta \in [0.1, 0.2, 0.3, 0.4]$) by t-SNE (Van der Maaten & Hinton, 2008) 2D embeddings of deep features.



(a) AGCE with different parameters (b) AUL with different parameters (c) AEL with different parameters

Figure 5. Illustration of asymmetric loss functions.



(a) AGCE with $q = 0.5$ (b) AGCE with $q = 1.5$ (c) AUL with $p < 1$ (d) AUL with $p = 2$

Figure 6. Test accuracies of AGCE and AUL with different parameters on CIFAR-10 under 0.8 symmetric noise.

Table 3. 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	99.17 ± 0.11	98.98 ± 0.07	98.54 ± 0.10	94.10 ± 0.70	46.60 ± 0.91
	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
	NFL	98.51 ± 0.03	98.35 ± 0.07	98.14 ± 0.06	97.48 ± 0.09	93.28 ± 0.40
	NGCE	98.72 ± 0.05	98.65 ± 0.04	98.42 ± 0.03	97.67 ± 0.12	94.76 ± 0.31
	NFL+RCE	99.41 ± 0.06	$\mathbf{99.13 \pm 0.07}$	98.46 ± 0.07	95.53 ± 0.36	73.52 ± 1.39
	NCE+MAE	99.34 ± 0.02	$\mathbf{99.14 \pm 0.05}$	98.42 ± 0.09	95.65 ± 0.13	72.97 ± 0.34
	NCE+RCE	99.36 ± 0.05	$\mathbf{99.14 \pm 0.03}$	98.51 ± 0.06	95.60 ± 0.21	74.00 ± 1.68
	AUL	99.14 ± 0.05	99.05 ± 0.09	$\mathbf{98.90 \pm 0.09}$	$\mathbf{98.67 \pm 0.04}$	$\mathbf{96.73 \pm 0.20}$
	AGCE	99.05 ± 0.11	98.96 ± 0.10	$\mathbf{98.83 \pm 0.06}$	$\mathbf{98.57 \pm 0.12}$	$\mathbf{96.59 \pm 0.12}$
	AEL	99.03 ± 0.05	98.93 ± 0.06	$\mathbf{98.78 \pm 0.13}$	$\mathbf{98.51 \pm 0.06}$	$\mathbf{96.40 \pm 0.11}$
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	80.73 ± 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
	NGCE	80.92 ± 0.16	78.82 ± 0.09	75.52 ± 0.37	69.79 ± 0.27	52.03 ± 0.88
	NFL	73.42 ± 0.35	70.93 ± 0.38	67.28 ± 0.24	60.30 ± 0.75	39.07 ± 0.40
	NFL+RCE	90.97 ± 0.19	88.89 ± 0.14	$\mathbf{86.03 \pm 0.33}$	79.65 ± 0.41	$\mathbf{54.33 \pm 0.80}$
	NCE+MAE	89.17 ± 0.09	86.98 ± 0.07	83.74 ± 0.10	76.02 ± 0.16	46.69 ± 0.31
	NCE+RCE	90.87 ± 0.37	$\mathbf{89.25 \pm 0.42}$	85.81 ± 0.08	$\mathbf{79.72 \pm 0.20}$	$\mathbf{55.74 \pm 0.95}$
	AUL	91.27 ± 0.12	89.21 ± 0.09	85.64 ± 0.19	$78,86 \pm 0.66$	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	$\mathbf{89.31 \pm 0.20}$	$\mathbf{86.23 \pm 0.18}$	$\mathbf{79.70 \pm 0.08}$	$\mathbf{59.44 \pm 1.14}$
	NCE+AGCE	90.94 ± 0.12	$\mathbf{89.21 \pm 0.08}$	$\mathbf{86.19 \pm 0.15}$	$\mathbf{80.13 \pm 0.18}$	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
	NGCE	22.83 ± 0.30	18.96 ± 1.41	15.09 ± 0.64	11.07 ± 0.77	6.14 ± 0.50
	NFL	28.73 ± 0.08	23.85 ± 0.24	18.96 ± 0.58	13.30 ± 0.80	8.20 ± 0.16
	NFL+RCE	67.90 ± 0.40	64.53 ± 0.69	57.85 ± 0.54	44.79 ± 1.00	$\mathbf{24.71 \pm 0.93}$
	NCE+MAE	67.60 ± 0.51	52.30 ± 0.11	36.09 ± 0.55	18.63 ± 0.60	7.48 ± 1.35
	NCE+RCE	68.65 ± 0.40	64.97 ± 0.49	58.54 ± 0.13	45.80 ± 1.02	$\mathbf{25.41 \pm 0.98}$
	NCE+AUL	68.96 ± 0.16	$\mathbf{65.36 \pm 0.20}$	$\mathbf{59.25 \pm 0.23}$	$\mathbf{46.34 \pm 0.21}$	23.03 ± 0.64
	NCE+AGCE	69.03 ± 0.37	$\mathbf{65.66 \pm 0.46}$	$\mathbf{59.47 \pm 0.36}$	$\mathbf{48.02 \pm 0.58}$	$\mathbf{24.72 \pm 0.60}$
	NCE+AEL	68.70 ± 0.20	$\mathbf{65.36 \pm 0.14}$	$\mathbf{59.51 \pm 0.03}$	$\mathbf{46.94 \pm 0.07}$	24.48 ± 0.24

Table 4. Test accuracies (%) of different methods on benchmark datasets with clean or asymmetric label noise ($\eta \in [0.1, 0.2, 0.3, 0.4]$). The results (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	99.13 ± 0.05	98.99 ± 0.01	97.27 ± 0.22	88.70 ± 0.49
	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
	NFL	98.35 ± 0.07	97.86 ± 0.16	96.33 ± 0.21	92.08 ± 0.28
	NGCE	98.73 ± 0.04	98.67 ± 0.05	98.32 ± 0.11	97.27 ± 0.08
	NFL+RCE	**99.38 ± 0.02**	98.98 ± 0.10	97.18 ± 0.14	89.58 ± 4.81
	NCE+MAE	**99.32 ± 0.09**	98.89 ± 0.04	96.93 ± 0.17	91.45 ± 0.40
	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	79.316 ± 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
	NGCE	80.18 ± 0.27	79.21 ± 0.08	76.76 ± 0.07	70.10 ± 1.82
	NFL	72.28 ± 0.15	70.78 ± 0.13	68.27 ± 0.43	65.09 ± 0.40
	NFL+RCE	89.91 ± 0.17	88.24 ± 0.16	**85.81 ± 0.23**	79.25 ± 0.25
	NCE+MAE	88.31 ± 0.20	86.50 ± 0.31	83.34 ± 0.39	77.14 ± 0.33
	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
	NGCE	20.89 ± 0.52	19.28 ± 0.23	17.77 ± 2.32	13.15 ± 2.90
	NFL	26.46 ± 0.31	25.39 ± 0.87	23.18 ± 0.80	20.10 ± 0.21
	NFL+RCE	65.97 ± 0.18	62.77 ± 0.31	**55.60 ± 0.25**	**41.66 ± 0.20**
	NCE+MAE	60.22 ± 0.37	52.20 ± 0.41	44.50 ± 0.46	35.82 ± 0.27
	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

References

Bartlett, P. L., Jordan, M. I., and Mcauliffe, J. D. Convexity, classi cation, and risk bounds. *Journal of the American Statistical Association*, 101, 2006.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pp. 2304–2313. PMLR, 2018.

Li, W., Wang, L., Li, W., Agustsson, E., and Van Gool, L. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.

Ma, X., Huang, H., Wang, Y., Romano, S., Erfani, S., and Bailey, J. Normalized loss functions for deep learning with noisy labels. In *ICML*, 2020.

Patrini, G., Rozza, A., Menon, A. K., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2233–2241, 2017. doi: 10.1109/CVPR.2017.240.

Tong, Z. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–134, 2003.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.