1. Notations

$x \in \mathbb{R}^d$: input image which is assumed to be sampled from distribution $\mathcal{X}$

$y_x \in \mathbb{R}^+$: the label of x

$\hat{y_x} \in \mathbb{R}^+$: the noisy label of x

$p(i|x)$: the ground truth distribution

$q(i|x)$: the predicted distribution

$f(\cdot) : \mathbb{R}^d \to \mathbb{R}^+$: the trained network

$L(f(x), i)$: the loss function. In this report, it refers to the improved loss function

$R_L = \mathbb{E}_{x,y_x}[f(x), y_x]$: the empirical risk under clean data

$R_L^\eta = \mathbb{E}_{x,y_x}[f(x), \hat{y_x}]$: the empirical risk under noisy data

f^*: the global minimizer of R_L

2. Symmetric Loss Function

I'd like to first validate whether the improved loss function: $l = \alpha l_{ce} + \beta l_{rce}$ is a symmetric loss function.

We have the definition that if a loss function satisfies the following property:

$$\exists\ C \text{ is constant}, \forall f, x \in \mathcal{X}$$
$$s.t. \sum_{i=1}^{k} L(f(x), i) = C,$$

For an arbitrary i, we have:

$$L(f(x), i) = -\alpha log\, q(i|x) - \beta A(1 - q(i|x)) \tag{1}$$

Then take the sum over i, we can get:

$$\sum_{i=1}^{k} L(f(x), i) = -\beta A(k-1) - \alpha \sum_{i=1}^{k} log\, q(i|x) \tag{2}$$

Since the first term in (2) is a constant and the second term is a variable, it is not a constant. With α taking a very small value, it could be treated as a constant to some extent. The improved loss function is thus not a symmetric one. That explains why the loss function doesn't perform well in the case of symmetric noise $\eta = 0.8$. However, it stills performs very well in the case of symmetric noise with small values.

3. Still a global optimizer?

With theorem 1 proposed in the referenced paper, we could examine the value of $R_L^\eta(f^*) - R_L(f^*)$

Firstly, I'd like to get the form of $R_L^\eta(f)$ under the new loss function:

$$\begin{aligned}
R_L^\eta(f) &= \mathbb{E}_x \mathbb{E}_{y_x|x} \mathbb{E}_{\hat{y_x}|x,y_x} L(f(x), \hat{y_x}) \\
&= \mathbb{E}_x \mathbb{E}_{y_x|x} [(1-\eta) L(f(x), y_x) + \frac{\eta}{k-1}(\hat{C} - \alpha \sum_{i=1}^{k} log\, q(i|x) + \alpha L(f(x), y_x))] \\
&= (\frac{\eta\alpha}{k-1} + 1 - \eta) R_L(f) + \frac{\eta}{k-1}(\alpha \mathbb{E}_x [\sum_{i=1}^{k} log\, q(i|x)] + \hat{C}) \\
&= M R_L(f) + N(\alpha \mathbb{E}_x [\sum_{i=1}^{k} log\, q(i|x)] + \hat{C}) \\
&\qquad \text{where } \hat{C} = -\beta A(k-1), M = \frac{\eta\alpha}{k-1} + 1 - \eta, N = \frac{\eta}{k-1}
\end{aligned}$$

Expand $R_L(f)$:

$$\begin{gathered}
R_L(f) = \mathbb{E}_{x,y_x}[-\alpha log\, q(y_x|x) - \beta A(1 - q(y_x|x)], \\
\text{where } q \text{ is the output of } f
\end{gathered} \tag{3}$$

Then insert (3) in and evaluate $R_L^\eta(f^*) - R_L^\eta(f)$:

$$\begin{aligned}
R_L^\eta(f^*) - R_L^\eta(f) &= M \mathbb{E}_{x,y_x}[-\alpha log \frac{q^*(y_x|x)}{q(y_x|x)} - \beta A(-q^*(y_x|x) + q(y_x|x))] \\
&\quad + N(\mathbb{E}_{x,y_x}[\alpha log \frac{q^*(y_x|x)}{q(y_x|x)}] + \mathbb{E}_x[\alpha \sum_{i \neq y_x} log \frac{q^*(i|x)}{q(i|x)}]) \\
&= (N - M) \cdot \mathbb{E}_{x,y_x}[\alpha log \frac{q^*(y_x|x)}{q(y_x|x)}] \\
&\quad + N \cdot \mathbb{E}_{x,y_x}[\alpha \sum_{i \neq y_x} log \frac{q^*(i|x)}{q(i|x)}] \\
&\quad + M\beta A \cdot \mathbb{E}_{x,y_x}[q^*(y_x|x) - q(y_x|x)]
\end{aligned}$$

Let's look at the three terms of the result:

For the first term, since f^* is the global optimum, we could easily conclude that $\mathbb{E}_{x,y_x}[\alpha log \frac{q^*(y_x|x)}{q(y_x|x)}]$ should be positive. Then with some algebra, we can get $N - M = \frac{k-\alpha}{k-1}\eta - 1 < 0$ when η is small. Further, when η is increased, the value of $N - M$ comes to greater than 0, which means that $(N-M)\mathbb{E}_{x,y_x}[\alpha log \frac{q^*(y_x|x)}{q(y_x|x)}]$ is a negative value with η being small and its value gets to be positive with higher η

For the third term, similar to the first term, we could easily conclude that $\mathbb{E}_{x,y_x}[q^*(y_x|x) - q(y_x|x)]$ should be a positive one as f^* is the global optimum. $M\beta A$ should be a negative number since A is negative. Since the value of M approaches 0 when η increases, so the value of the third approaches 0 when η increases.

For the second term, we could expect $\mathbb{E}_x[\alpha \sum_{i \neq y_x} log\frac{q^*(i|x)}{q(i|x)}]$ to be negative since the possibility of classes except the true class predicted by the global optimum should be generally less than that of other predictors. So the value of the second term is a negative one. While the N increases when η increases, the decreasing speed should be slower than the increasing speed of the first term and the third term because of the A in the coefficient of the third term.

So we can easily imagine such as process: At first, when η is very small, $R_L^\eta(f^*) - R_L^\eta(f)$ should be negative so it is still a global optimizer under noisy data. When η increases, the first term gets larger and the third term approaches 0. At some point, the value of $R_L^\eta(f^*) - R_L^\eta(f)$ flips to a positive number and f^* is no longer a global optimum. Then that explains why in the case that $\eta = 0.8$, the model doesn't perform very well.