A. Proof for Theorem 1

Theorem 1. *In a multi-class classification problem, ℓ_{rce} is noise tolerant under symmetric or uniform label noise if noise rate $\eta < 1 - \frac{1}{K}$. And, if $R(f^*) = 0$, ℓ_{rce} is also noise tolerant under asymmetric or class-dependent label noise when noise rate $\eta_{yk} < 1 - \eta_y$ with $\sum_{k\neq y} \eta_{yk} = \eta_y$.*

Proof. For symmetric noise:

$$\begin{aligned}
R^\eta(f) &= \mathbb{E}_{\mathbf{x},\hat{y}}\ell_{rce}(f(\mathbf{x}),\hat{y}) = \mathbb{E}_{\mathbf{x}}\mathbb{E}_{y|\mathbf{x}}\mathbb{E}_{\hat{y}|\mathbf{x},y}\ell_{rce}(f(\mathbf{x}),\hat{y}) \\
&= \mathbb{E}_{\mathbf{x}}\mathbb{E}_{y|\mathbf{x}}\Big[(1-\eta)\ell_{rce}(f(\mathbf{x}),y) + \frac{\eta}{K-1}\sum_{k\neq y}\ell_{rce}(f(\mathbf{x}),k)\Big] \\
&= (1-\eta)R(f) + \frac{\eta}{K-1}\Big(\mathbb{E}_{\mathbf{x},y}\Big[\sum_{k=1}^{K}\ell_{rce}(f(\mathbf{x}),k)\Big] - R(f)\Big) \\
&= R(f)\left(1 - \frac{\eta K}{K-1}\right) - A\eta,
\end{aligned}$$

where the last equality holds due to $\sum_{k=1}^{K}\ell_{rce}(f(\mathbf{x}),k) = -(K-1)A$ following Eq. (5) and the definition of $\log 0 = A$ (a negative constant). Thus,

$$R^\eta(f^*) - R^\eta(f) = (1 - \frac{\eta K}{K-1})(R(f^*) - R(f)) \leq 0,$$

because $\eta < 1 - \frac{1}{K}$ and f^* is a global minimizer of $R(f)$. This proves f^* is also the global minimizer of risk $R^\eta(f)$, that is, ℓ_{rce} is noise tolerant.

For asymmetric or class-dependent noise, $1 - \eta_y$ is the probability of a label being correct (*i.e.,* $k = y$), and the noise condition $\eta_{yk} < 1 - \eta_y$ generally states that a sample $\mathbf{x}$ still has the highest probability of being in the correct class y, though it has probability of η_{yk} being in an arbitrary noisy (incorrect) class $k \neq y$. Considering the noise transition matrix between classes $[\eta_{ij}], \forall i,j \in \{1,2,\cdots,K\}$, this condition only requires that the matrix is diagonal dominated by η_{ii} (*i.e.,* the correct class probability $1-\eta_y$). Following the symmetric case, here we have,

$$\begin{aligned}
R^\eta(f) &= \mathbb{E}_{\mathbf{x},\hat{y}}\ell_{rce}(f(\mathbf{x}),\hat{y}) = \mathbb{E}_{\mathbf{x}}\mathbb{E}_{y|\mathbf{x}}\mathbb{E}_{\hat{y}|\mathbf{x},y}\ell_{rce}(f(\mathbf{x}),\hat{y}) \\
&= \mathbb{E}_{\mathbf{x}}\mathbb{E}_{y|\mathbf{x}}\Big[(1-\eta_y)\ell_{rce}(f(\mathbf{x}),y) + \sum_{k\neq y}\eta_{yk}\ell_{rce}(f(\mathbf{x}),k)\Big] \\
&= \mathbb{E}_{\mathbf{x},y}\Big[(1-\eta_y)\big(\sum_{k=1}^{K}\ell_{rce}(f(\mathbf{x}),k) - \sum_{k\neq y}\ell_{rce}(f(\mathbf{x}),k)\big)\Big] + \mathbb{E}_{\mathbf{x},y}\Big[\sum_{k\neq y}\eta_{yk}\ell_{rce}(f(\mathbf{x}),k)\Big] \\
&= \mathbb{E}_{\mathbf{x},y}\Big[(1-\eta_y)\big(-(K-1)A - \sum_{k\neq y}\ell_{rce}(f(\mathbf{x}),k)\big)\Big] + \mathbb{E}_{\mathbf{x},y}\Big[\sum_{k\neq y}\eta_{yk}\ell_{rce}(f(\mathbf{x}),k)\Big] \\
&= -(K-1)A\mathbb{E}_{\mathbf{x},y}(1-\eta_y) - \mathbb{E}_{\mathbf{x},y}\Big[\sum_{k\neq y}(1-\eta_y-\eta_{yk})\ell_{rce}(f(\mathbf{x}),k)\Big].
\end{aligned} \tag{12}$$

As f^*_η is the minimizer of $R^\eta(f)$, $R^\eta(f^*_\eta) - R^\eta(f^*) \leq 0$. So, from Eq.(12), we have,

$$\mathbb{E}_{\mathbf{x},y}\Big[\sum_{k\neq y}(1-\eta_y-\eta_{yk})\big(\underbrace{\ell_{rce}(f^*(\mathbf{x}),k)}_{\ell^*_{rce}} - \underbrace{\ell_{rce}(f^*_\eta(\mathbf{x}),k)}_{\ell^{\eta*}_{rce}}\big)\Big] \leq 0. \tag{13}$$

Next, we prove, $f^*_\eta = f^*$ holds following Eq. (13). First, $(1-\eta_y-\eta_{yk}) > 0$ as per the assumption that $\eta_{yk} < 1-\eta_y$. Since we are given $R(f^*) = 0$, we have $\ell_{rce}(f^*(\mathbf{x}),k) = -A$ for all $k \neq y$. Also, by the definition of $\ell^{\eta*}_{rce}$, we have $\ell_{rce}(f^*_\eta(\mathbf{x}),k) = -A(1-p_k) \leq -A$, $\forall k \neq y$. Thus, for Eq. (13) to hold (*e.g.* $\ell_{rec}(f^*_\eta(\mathbf{x}),k) \geq \ell_{rec}(f^*(\mathbf{x}),k)$), it must be the case that $p_k = 0$, $\forall k \neq y$, that is, $\ell_{rec}(f^*_\eta(\mathbf{x}),k) = \ell_{rec}(f^*(\mathbf{x}),k)$ for all $k \in \{1,2,\cdots,K\}$, thus $f^*_\eta = f^*$ which completes the proof. ∎

B. Gradient Derivation of SL

The complete derivartion of the simplified SL ($\alpha, \beta = 1$) with respect to the logits is as follows:

$$\frac{\partial \ell_{sl}}{\partial z_j} = -\sum_{k=1}^{K} q_k \frac{1}{p_k}\frac{\partial p_k}{\partial z_j} - \sum_{k=1}^{K}\frac{\partial p_k}{\partial z_j}\log q_k, \tag{14}$$

where

$$\frac{\partial p_k}{\partial z_j} = \frac{\partial \left(\frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}} \right)}{\partial z_j} = \frac{\frac{\partial e^{z_k}}{\partial z_j} (\sum_{j=1}^{K} e^{z_j}) - e^{z_k} \frac{\partial \left(\sum_{j=1}^{K} e^{z_j} \right)}{\partial z_j}}{(\sum_{j=1}^{K} e^{z_j})^2}. \tag{15}$$

In the case of $k = j$:

$$\begin{aligned} \frac{\partial p_k}{\partial z_j} = \frac{\partial p_k}{\partial z_k} &= \frac{e^{z_k} \left(\sum_{k=1}^{K} e^{z_k} \right) - (e^{z_k})^2}{(\sum_{k=1}^{K} e^{z_k})^2} \\ &= \frac{e^{z_k}}{\sum_{k=1}^{K} e^{z_k}} - \left(\frac{e^{z_k}}{\sum_{k=1}^{K} e^{z_k}} \right)^2 \\ &= p_k - p_k^2 = p_k(1 - p_k); \end{aligned} \tag{16}$$

In the case of $k \neq j$:

$$\begin{aligned} \frac{\partial p_k}{\partial z_j} &= \frac{0 \cdot (\sum_{j=1}^{K} e^{z_j}) - e^{z_k} e^{z_j}}{(\sum_{j=1}^{K} e^{z_j})(\sum_{j=1}^{K} e^{z_j})} \\ &= -\frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}} \frac{e^{z_j}}{\sum_{j=1}^{K} e^{z_j}} \\ &= -p_k p_j. \end{aligned} \tag{17}$$

Combining Eq. (16) and (17) into Eq. (14), we can obtain:

$$\begin{aligned} \frac{\partial \ell_{sl}}{\partial z_j} &= -\sum_{k=1}^{K} q_k \frac{1}{p_k} \frac{\partial p_k}{\partial z_j} - \sum_{k=1}^{K} \frac{\partial p_k}{\partial z_j} \log q_k \\ &= -\sum_{k \neq j}^{K} \frac{q_k}{p_k} (-p_j p_k) - \frac{q_j}{p_j} (p_j(1 - p_j)) - \sum_{k \neq j}^{K} (-p_j p_k) \log q_k - p_j(1 - p_j) \log q_j \\ &= p_j - q_j + p_j (\sum_{k=1}^{K} p_k \log q_k - \log q_j). \end{aligned} \tag{18}$$

If $q_j = q_y = 1$, then the gradient of SL is:

$$\begin{aligned} \frac{\partial \ell_{sl}}{\partial z_j} &= p_j - q_j + p_j (\sum_{k=1}^{K} p_k \log q_k - \log q_j) \\ &= (p_j - 1) + p_j((1 - p_j)A - 0) \\ &= \frac{\partial \ell_{ce}}{\partial z_j} - (Ap_j^2 - Ap_j). \end{aligned} \tag{19}$$

Else if $q_j = 0$, then

$$\begin{aligned} \frac{\partial \ell_{sl}}{\partial z_j} &= p_j - q_j + p_j (\sum_{k=1}^{K} p_k \log q_k - \log q_j) \\ &= p_j + p_j((1 - p_y)A - A) \\ &= \frac{\partial \ell_{ce}}{\partial z_j} - Ap_j p_y. \end{aligned} \tag{20}$$