A. Proofs for Lemma 1, Lemma 2 and Lemma 3

Our proofs are inspired by (Ghosh et al., 2017).

Lemma 1. *In a multi-class classification problem, any normalized loss function $\mathcal{L}_{norm}$ is noise tolerant under symmetric (or uniform) label noise, if noise rate $\eta < \frac{K-1}{K}$.*

Proof. For symmetric label noise, the noise risk can be defined as:

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

where the last equality holds due to $\sum_{k=1}^{K}\mathcal{L}_{norm}(f(\boldsymbol{x}),k)=1$, following Eq. (1). Thus,

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

because $\eta < \frac{K-1}{K}$ and f^* is a global minimizer of $R(f)$. This proves f^* is also the global minimizer of risk $R^{\eta}(f)$, that is, $\mathcal{L}_{norm}$ is noise tolerant to symmetric label noise. □

Lemma 2. *In a multi-class classification problem, given $R(f^*)=0$ and $0 \leq \mathcal{L}_{norm}(f^*(\boldsymbol{x}),k) \leq \frac{1}{K-1}$, any normalized loss function $\mathcal{L}_{norm}$ is noise tolerant under asymmetric (or class-conditional) label noise, if noise rate $\eta_{jk} < 1-\eta_y$.*

Proof. For asymmetric or class-conditional 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 $\boldsymbol{x}$ 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,

$$\begin{aligned}
R^{\eta}(f) &= \mathbb{E}_{\boldsymbol{x},\hat{y}}\mathcal{L}_{norm}(f(\boldsymbol{x}),\hat{y}) = \mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{y|\boldsymbol{x}}\mathbb{E}_{\hat{y}|\boldsymbol{x},y}\mathcal{L}_{norm}(f(\boldsymbol{x}),\hat{y}) \\
&= \mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{y|\boldsymbol{x}}\Big[(1-\eta_y)\mathcal{L}_{norm}(f(\boldsymbol{x}),y) + \sum_{k\neq y}\eta_{yk}\mathcal{L}_{norm}(f(\boldsymbol{x}),k)\Big] \\
&= \mathbb{E}_{\boldsymbol{x},y}\Big[(1-\eta_y)\Big(\sum_{k=1}^{K}\mathcal{L}_{norm}(f(\boldsymbol{x}),k) - \sum_{k\neq y}\mathcal{L}_{norm}(f(\boldsymbol{x}),k)\Big)\Big] + \mathbb{E}_{\boldsymbol{x},y}\Big[\sum_{k\neq y}\eta_{yk}\mathcal{L}_{norm}(f(\boldsymbol{x}),k)\Big] \\
&= \mathbb{E}_{\boldsymbol{x},y}\Big[(1-\eta_y)\big(1-\sum_{k\neq y}\mathcal{L}_{norm}(f(\boldsymbol{x}),k)\big)\Big] + \mathbb{E}_{\boldsymbol{x},y}\Big[\sum_{k\neq y}\eta_{yk}\mathcal{L}_{norm}(f(\boldsymbol{x}),k)\Big] \\
&= \mathbb{E}_{\boldsymbol{x},y}(1-\eta_y) - \mathbb{E}_{\boldsymbol{x},y}\Big[\sum_{k\neq y}(1-\eta_y-\eta_{yk})\mathcal{L}_{norm}(f(\boldsymbol{x}),k)\Big].
\end{aligned} \tag{7}$$

As f^*_η is the minimizer of $R^{\eta}(f)$, $R^{\eta}(f^*_\eta) - R^{\eta}(f^*) \leq 0$. So, from 7 above, we have,

$$\mathbb{E}_{\boldsymbol{x},y}\Big[\sum_{k\neq y}(1-\eta_y-\eta_{yk})\big(\underbrace{\mathcal{L}_{norm}(f^*(\boldsymbol{x}),k)}_{\mathcal{L}^*_{norm}} - \underbrace{\mathcal{L}_{norm}(f^*_\eta(\boldsymbol{x}),k)}_{\mathcal{L}^{\eta*}_{norm}}\big)\Big] \leq 0. \tag{8}$$

Next, we prove, $f^*_\eta = f^*$ holds following Eq. (8). First, $(1-\eta_y-\eta_{yk}) > 0$ as per the assumption that $\eta_{yk} < 1-\eta_y$. Thus, $\mathcal{L}^*_{norm} - \mathcal{L}^{\eta*}_{norm} \leq 0$ for Eq. (8) to hold. Since we are given $R(f^*)=0$, we have $\mathcal{L}(f^*(\boldsymbol{x}),y)=0$. Thus, following the definition of $\mathcal{L}_{norm}$ in Eq. (1) and assumption $\mathcal{L}_{\text{norm}}(f^*(\boldsymbol{x}),k) \leq \frac{1}{K-1}$, we have $\mathcal{L}_{norm}(f^*(\boldsymbol{x}),k) = \frac{\mathcal{L}(f^*(\boldsymbol{x})=0,k)}{\sum_j^K \mathcal{L}(f^*(\boldsymbol{x}),j)} = \frac{1}{K-1}$, for all $k \neq y$. Also, we have $\mathcal{L}_{norm}(f^*_\eta(\boldsymbol{x}),k) = \frac{\mathcal{L}(f^*_\eta(\boldsymbol{x}),k)}{\sum_j^K \mathcal{L}(f^*_\eta(\boldsymbol{x}),j)} \leq \frac{1}{K-1}, \forall k \neq y$. Thus, for Eq. (8) to hold (*e.g.* $\mathcal{L}_{norm}(f^*_\eta(\boldsymbol{x}),k) \geq \mathcal{L}_{norm}(f^*(\boldsymbol{x}),k)$), it must be the case that $p_k = 0, \forall k \neq y$, that is, $\mathcal{L}_{norm}(f^*_\eta(\boldsymbol{x}),k) = \mathcal{L}_{norm}(f^*(\boldsymbol{x}),k)$ for all $k \in \{1,2,\cdots,K\}$, thus $f^*_\eta = f^*$ which completes the proof. □

Lemma 3. *$\forall\alpha, \forall\beta$, if $\mathcal{L}_{Active}$ and $\mathcal{L}_{Passive}$ are noise tolerant, then $\mathcal{L}_{APL} = \alpha \cdot \mathcal{L}_{Active} + \beta \cdot \mathcal{L}_{Passive}$ is noise tolerant.*

Proof. Let $\alpha, \beta \in \mathbb{R}$, then $\sum_j^K \mathcal{L}_{\text{APL}}(f(\boldsymbol{x}), j) = \alpha \cdot \sum_j^K \mathcal{L}_{\text{Active}}(f(\boldsymbol{x}), j) + \beta \cdot \sum_j^K \mathcal{L}_{\text{Passive}}(f(\boldsymbol{x}), j) = \alpha \cdot C_{\text{Active}} + \beta \cdot C_{\text{Passive}} = C$. Following our proof for Lemma 1, for symmetric noise, we have,

$$R^\eta(f) = R(f)\left(1 - \frac{\eta K}{K-1}\right) + \frac{(\alpha \cdot C_{\text{Active}} + \beta \cdot C_{\text{Passive}})\eta}{K-1}.$$

Thus, $R^\eta(f^*) - R^\eta(f) = (1 - \frac{\eta K}{K-1})(R(f^*) - R(f)) \leq 0$. Given $\eta < \frac{K-1}{K}$ and f^* is a global minimizer of $R(f)$, $R(f^*) - R(f)$, that is, f^* is also the global minimizer of risk $R^\eta(f)$. Thus, $\mathcal{L}_{\text{APL}}$ is noise tolerant to symmetric label noise.

Following our proof for Lemma 2, for asymmetric noise, we have,

$$R^\eta(f) = (\alpha \cdot C_{\text{Active}} + \beta \cdot C_{\text{Passive}})\mathbb{E}_{\boldsymbol{x},y}(1 - \eta_y) - \mathbb{E}_{\boldsymbol{x},y}\Big[\sum_{k \neq y}(1 - \eta_y - \eta_{yk})\mathcal{L}_{norm}(f(\boldsymbol{x}), k)\Big]. \tag{9}$$

As f_η^* is the minimizer of $R^\eta(f)$, $R^\eta(f_\eta^*) - R^\eta(f^*) \leq 0$. So, from 9 above, we can derive the same equation as Eq. (8),

$$\mathbb{E}_{\boldsymbol{x},y}\Big[\sum_{k \neq y}(1 - \eta_y - \eta_{yk})\big(\underbrace{\mathcal{L}_{\text{APL}}(f^*(\boldsymbol{x}), k)}_{\mathcal{L}^*_{\text{APL}}} - \underbrace{\mathcal{L}_{\text{APL}}(f^*_\eta(\boldsymbol{x}), k)}_{\mathcal{L}^{\eta*}_{\text{APL}}}\big)\Big] \leq 0. \tag{10}$$

Thus, we can follow the same proof from Eq. (8), to $f_\eta^* = f^*$, that is, $\mathcal{L}_{\text{APL}}$ is also noise tolerant to asymmetric noise. □