Supplementary Material for Discriminative Complementary-Label Learning with Weighted Loss

A. The Proof of Lemma 1

Lemma 1. *Based on Eq.(5) and Assumption 1, it holds that*

$$\widehat{\mathfrak{R}}_n(\bar{\ell} \circ \mathcal{F}) \leq c^2 L_\ell \widehat{\mathfrak{R}}_n\left(\mathcal{F}_k\right)$$

Proof. Given

$$\widehat{\mathfrak{R}}_n(\bar{\ell} \circ \mathcal{F}) = \mathbb{E}_\sigma\left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \bar{\ell}(\boldsymbol{f}(\boldsymbol{x}_i), e^{\bar{y}_i})\right]$$

where $\sigma = [\sigma_1, \ldots, \sigma_n]$, which denotes n Rademacher varibles. Let us first assume $c = 2$ and use the max operator $\max(a, b) = \frac{1}{2}(a + b + |a - b|)$. Thus, we have

$$\begin{aligned}
\widehat{\mathfrak{R}}_n(\bar{\ell} \circ \mathcal{F}) &= \mathbb{E}_\sigma\left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \bar{\ell}(\boldsymbol{f}(\boldsymbol{x}_i), e^{\bar{y}_i})\right] \\
&= \mathbb{E}_\sigma\left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \sum_{k=1}^{c} \bar{\ell}(f_k(\boldsymbol{x}_i), e_k^{\bar{y}_i})\right] \\
&\leq \mathbb{E}_\sigma\left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} c\sigma_i \max_{k \in \{1, \ldots, c\}} \bar{\ell}(f_k(\boldsymbol{x}_i), e_k^{\bar{y}_i})\right] \\
&\leq \mathbb{E}_\sigma\left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{2n} \sum_{i=1}^{n} c\sigma_i \bar{\ell}(f_1(\boldsymbol{x}_i), e_1^{\bar{y}_i})\right] \\
&+ \mathbb{E}_\sigma\left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{2n} \sum_{i=1}^{n} c\sigma_i \bar{\ell}(f_2(\boldsymbol{x}_i), e_2^{\bar{y}_i})\right] \\
&+ \mathbb{E}_\sigma\left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{2n} \sum_{i=1}^{n} c\sigma_i |\bar{\ell}(f_1(\boldsymbol{x}_i), e_1^{\bar{y}_i}) - \bar{\ell}(f_2(\boldsymbol{x}_i), e_2^{\bar{y}_i})|\right] \\
&\leq 2\mathbb{E}_\sigma\left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} c\sigma_i \bar{\ell}(f_k(\boldsymbol{x}_i), e_k^{\bar{y}_i})\right]
\end{aligned}$$

When there are c classes, the general case can be derived from $\max\{z_1, \ldots, z_c\} = \max\{z_1, \max\{z_2, \ldots, z_c\}\}$, by recurrence, we will have

$$\widehat{\mathfrak{R}}_n(\bar{\ell} \circ \mathcal{F}) \leq c^2 \mathbb{E}_\sigma\left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \bar{\ell}(f_k(\boldsymbol{x}_i), e_k^{\bar{y}_i})\right]$$

By our Assumption 1 and the definition of $\bar{\ell}(\cdot)$, we further have

$$\begin{aligned}
&\mathbb{E}_\sigma\left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \bar{\ell}(f_k(\boldsymbol{x}_i), e_k^{\bar{y}_i})\right] \\
&\leq \mathbb{E}_\sigma\left[\sup_{\boldsymbol{f} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(1 - f_k(\boldsymbol{x}_i), 1 - e_k^{y_i})\right]
\end{aligned}$$

$$\leq \mathbb{E}_\sigma \left[\sup_{\boldsymbol{f}\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(f_k(\boldsymbol{x}_i), e_k^{y_i}) \right] = \widehat{\mathfrak{R}}_n(\ell \circ \mathcal{F}_k)$$

According to Talagrand's contraction lemma (Ledoux & Talagrand, 1991), we have $\widehat{\mathfrak{R}}_n(\bar{\ell} \circ \mathcal{F}) \leq c^2 L_\ell \widehat{\mathfrak{R}}_n(\mathcal{F}_k)$. □

B. The Proof of Lemma 2

Given the upper bound for $\widehat{\mathfrak{R}}_n(\bar{\ell} \circ \mathcal{F})$, we can prove Lemma 2 that defines the uniform deviation bound.

Lemma 2. *For any* $\delta > 0$, *with probability at least* $1 - \delta$,

$$\underset{\boldsymbol{f}\in\mathcal{F}}{sup} \left| \bar{R}(\boldsymbol{f}) - \bar{R}_n(\boldsymbol{f}) \right| \leq 2c^2 L_\ell \widehat{\mathfrak{R}}_n(\mathcal{F}_k) + M\sqrt{\frac{log(2/\delta)}{2n}}$$

where $\bar{R}(\boldsymbol{f})$ and $\bar{R}_n(\boldsymbol{f})$ is defined by Eq.(6) and Eq.(7) respectively.

Proof. Consider the single direction $sup_{\boldsymbol{f}\in\mathcal{F}}(\bar{R}(\boldsymbol{f}) - \bar{R}_n(\boldsymbol{f}))$ with probability at least $1-\delta/2$. Because M is the upper bound of ℓ, the change of $sup_{\boldsymbol{f}\in\mathcal{F}}\left(\bar{R}(\boldsymbol{f}) - \bar{R}_n(\boldsymbol{f})\right)$ is no greater than M/n after some x are replaced. So using McDiarmid's inequality (McDiarmid, 2013) to $sup_{\boldsymbol{f}\in\mathcal{F}}\left(\bar{R}(\boldsymbol{f}) - \bar{R}_n(\boldsymbol{f})\right)$, we have

$$\underset{\boldsymbol{f}\in\mathcal{F}}{sup}\left(\bar{R}(\boldsymbol{f}) - \bar{R}_n(\boldsymbol{f})\right) \leq \mathbb{E}\left[\underset{\boldsymbol{f}\in\mathcal{F}}{sup}\left(\bar{R}(\boldsymbol{f}) - \bar{R}_n(\boldsymbol{f})\right)\right] + M\sqrt{\frac{log\,(2/\delta)}{2n}}$$

By symmetrization (Mohri et al., 2012), it is a routine work to show that

$$\mathbb{E}\left[\underset{\boldsymbol{f}\in\mathcal{F}}{sup}\left(\bar{R}(\boldsymbol{f}) - \bar{R}_n(\boldsymbol{f})\right)\right] \leq 2\widehat{\mathfrak{R}}_n(\bar{\ell} \circ \mathcal{F}) = 2c^2 L_\ell \widehat{\mathfrak{R}}_n(\mathcal{F}_k)$$

□

C. The Proof of Theorem 1

According to Lemma 2, we can establish the estimation error bound for the proposed CLL risk estimator. The estimation error bound is shown in Theorem 1.

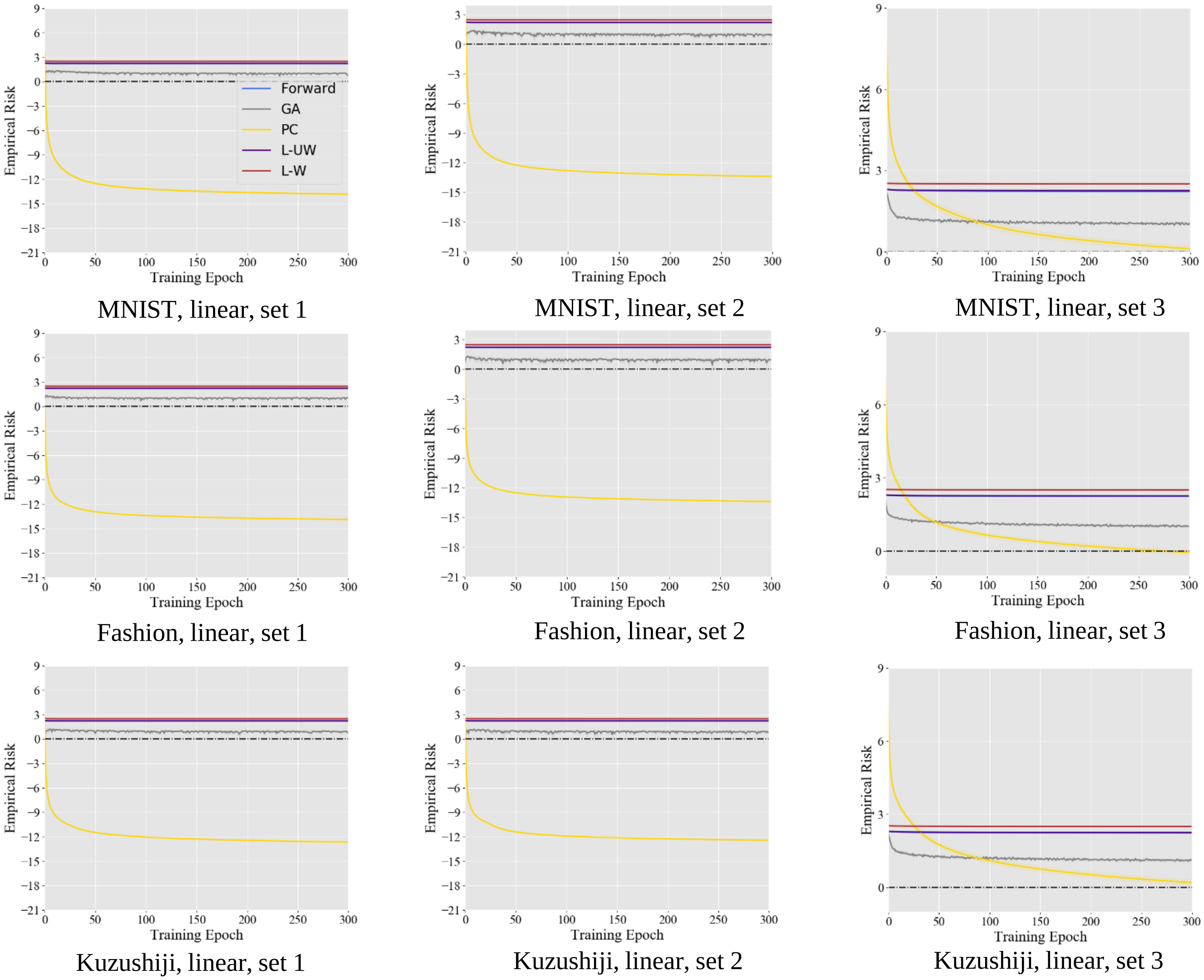
Theorem 1. *For any* $\delta > 0$, *with probability at least* $1 - \delta$,

$$\bar{R}(\bar{\boldsymbol{f}}_n^*) - \bar{R}(\bar{\boldsymbol{f}}^*) \leq 4c^2 L_\ell \widehat{\mathfrak{R}}_n(\mathcal{F}_k) + M\sqrt{\frac{2log(2/\delta)}{n}}$$

Proof.

$$\begin{aligned}
\bar{R}(\bar{\boldsymbol{f}}_n^*) - \bar{R}(\bar{\boldsymbol{f}}^*) &= (\bar{R}_n(\bar{\boldsymbol{f}}_n^*) - \bar{R}_n(\bar{\boldsymbol{f}}^*)) + (\bar{R}(\bar{\boldsymbol{f}}_n^*) - \bar{R}_n(\bar{\boldsymbol{f}}_n^*)) + (\bar{R}_n(\bar{\boldsymbol{f}}^*) - \bar{R}(\bar{\boldsymbol{f}}^*)) \\
&\leq \bar{R}(\bar{\boldsymbol{f}}_n^*) - \bar{R}_n(\bar{\boldsymbol{f}}_n^*) + \bar{R}_n(\bar{\boldsymbol{f}}^*) - \bar{R}(\bar{\boldsymbol{f}}^*) \\
&\leq 2\underset{\boldsymbol{f}\in\mathcal{F}}{sup}\left|\bar{R}_n(\bar{\boldsymbol{f}}) - \bar{R}(\bar{\boldsymbol{f}})\right| \\
&\leq 4c^2 L_\ell \widehat{\mathfrak{R}}_n(\mathcal{F}_k) + M\sqrt{\frac{2\log(2/\delta)}{n}}
\end{aligned}$$

Since $\bar{R}_n(\bar{\boldsymbol{f}}_n^*) - \bar{R}_n(\bar{\boldsymbol{f}}^*) \leq 0$, the second step in the above equation naturally follows from the first step. The proof is complete. □



MNIST, linear, set 1

MNIST, linear, set 2

MNIST, linear, set 3

Fashion, linear, set 1

Fashion, linear, set 2

Fashion, linear, set 3

Kuzushiji, linear, set 1

Kuzushiji, linear, set 2

Kuzushiji, linear, set 3

Figure 1. The experimental results on various biased settings on the linear model for 300 epochs. The dark color is the mean accuracy and the light color corresponds to the std.

D. Empirical Risk for Biased Settings

Figure 1 and Figure 2 are corresponding empirical risks for the linear model and MLP model on various datasets and biased settings.

Results From Figure 1, the empirical risk of PC on three datasets goes non-negative when the generation setting of complementary labels gradually becomes uniform. Furthermore, as shown in Figure 1, all approaches work normally with linear base model on MNIST, Fashion-MNIST and Kuzushiji-MNIST, while empirical risk of URE-based methods, such as PC and GA, goes zero or even negative when the more complex models are applied (shown in 2). Specifically, under the case of using MLP model, the performance of PC becomes the worst. This is due to the property that URE-based methods are easy to suffer from over-fitting problem when using complex models (Chou et al., 2020).

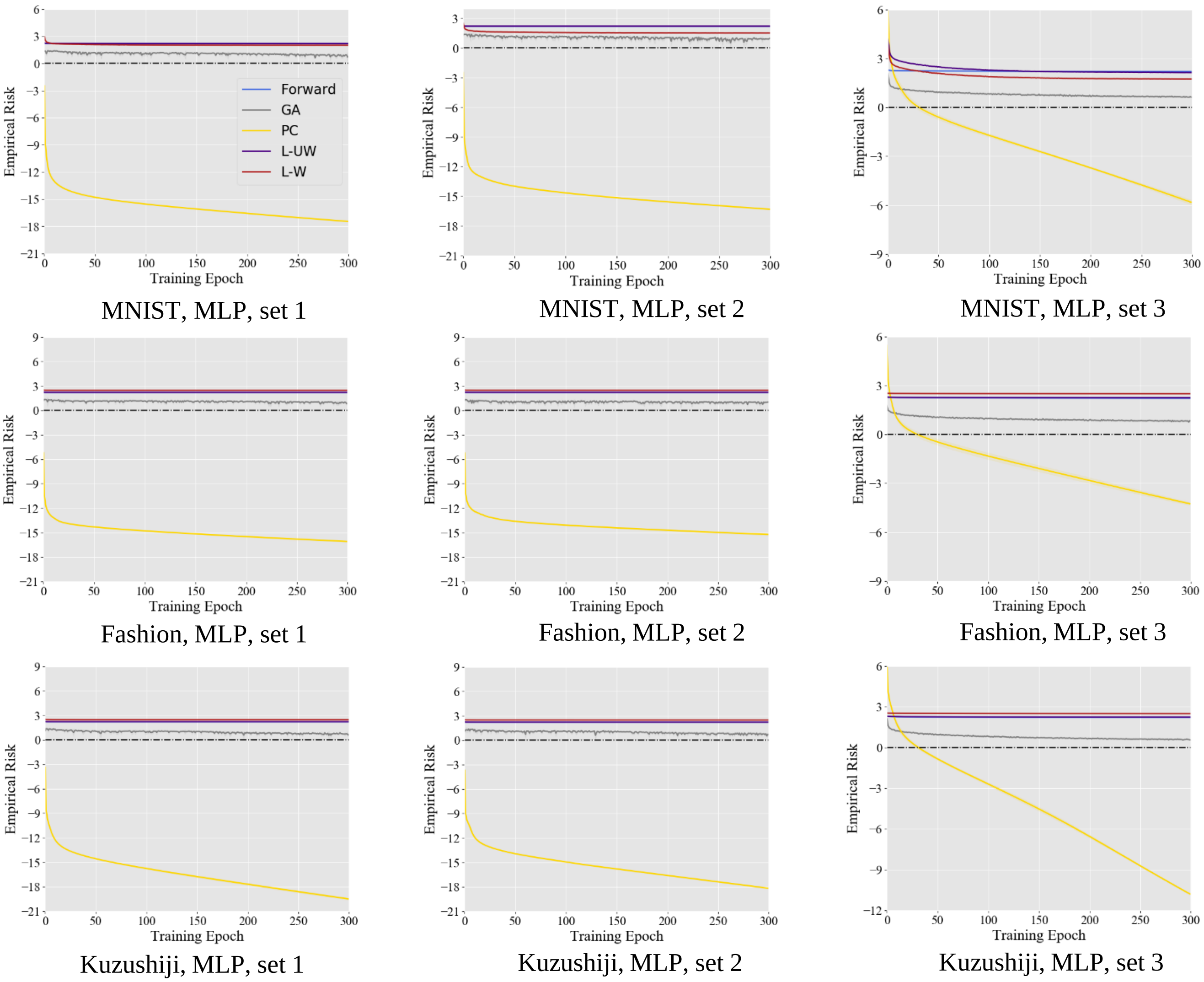

MNIST, MLP, set 1
MNIST, MLP, set 2
MNIST, MLP, set 3
Fashion, MLP, set 1
Fashion, MLP, set 2
Fashion, MLP, set 3
Kuzushiji, MLP, set 1
Kuzushiji, MLP, set 2
Kuzushiji, MLP, set 3

Figure 2. The experimental results on various biased settings on the MLP model for 300 epochs. The dark color is the mean accuracy and the light color corresponds to the std.