

# Supplementary Materials: Learning from Concealed Labels

Anonymous Authors

This is the supplemental material for the paper "Learning from Concealed Labels"

Hence, from Eq.(4), we have

## 1 PROOF OF LEMMA 2

**Lemma 2.** Under the concealed labels assumption, we can express conditional distribution  $P(y = i \neq cl | x)$  and  $P(y = cl | x)$  in terms of  $P(s = i | x)$ ,  $P(s = s_{none} | x)$  as

$$P(y = i \neq cl | x) = \frac{K}{L} P(s = i | x)$$

$$P(y = cl | x) = \frac{K}{L} P(s = s_{none} | x) - \frac{K - L}{L}$$

*Proof.* According to the Definition 1, the probability  $P(s = y_i | x)$  can be expressed as follows:

$$\begin{aligned} P(y = i \neq cl | x) &= \sum_j^K P(s = j, y = i \neq cl | x) + P(s = s_{none}, y = i \neq cl | x) \\ &= P(y = i \neq cl, s = i | x) + P(s = s_{none}, y = i \neq cl | x) \\ &= P(s = i | x) + P(s = s_{none} | x, y = i \neq cl) P(y = i \neq cl | x) \quad (1) \\ &\quad \left( \because P(s = s_{none} | x, y = i \neq cl) = \frac{K - L}{K}, \text{ Definition 1} \right) \\ &= P(s = i | x) + \frac{K - L}{K} P(y = i \neq cl | x). \end{aligned}$$

Hence, we have

$$P(y = i \neq cl | x) = \frac{K}{L} P(s = i | x). \quad (2)$$

In a similar manner, we have

$$\begin{aligned} P(y = cl | x) &= \sum_j^K P(s = j, y = cl | x) + P(s = s_{none}, y = cl | x) \\ &= P(s = s_{none}, y = cl | x) \\ &= \sum_Y P(Y, s = s_{none}, y = cl | x). \end{aligned} \quad (3)$$

On the other hands,

$$\begin{aligned} P(s = s_{none}, Y | x) &= \sum_i^K P(y = i, s = s_{none}, Y | x) + P(y = cl, s = s_{none}, Y | x) \\ &= \sum_i^K [P(Y, y = i | x) - P(Y, s = i, y = i | x)] \\ &\quad + P(y = cl, s = s_{none}, Y | x) \\ &= P(Y | x) - P(Y, y = cl | x) \\ &\quad - \sum_i^K P(Y, s = i, y = i | x) + P(y = cl, s = s_{none}, Y | x). \end{aligned} \quad (4)$$

$$\begin{aligned} P(y = cl, s = s_{none}, Y | x) &= P(s = s_{none}, Y | x) - P(Y | x) + P(Y, y = cl | x) \\ &\quad + \sum_i^K P(Y, s = i, y = i | x). \end{aligned} \quad (5)$$

By substituting eq.(5) into eq.(3), we can obtain

$$\begin{aligned} P(y = cl | x) &= \sum_Y \left[ P(s = s_{none}, Y | x) - P(Y | x) \right. \\ &\quad \left. + P(Y, y = cl | x) + \sum_i^K P(Y, s = i, y = i | x) \right] \\ &= P(s = s_{none} | x) - 1 + P(y = cl | x) \\ &\quad + \sum_i^K P(s = i, y = i | x) \\ &= P(s = s_{none} | x) - 1 + P(y = cl | x) \\ &\quad + \sum_i^K P(s = i | y = i, x) P(y = i | x) \\ &= P(s = s_{none} | x) - 1 + P(y = cl | x) \\ &\quad + \frac{L}{K} \sum_i^K P(y = i | x). \quad (\because \text{Definition 1}) \end{aligned} \quad (6)$$

On the other hands,

$$\begin{aligned} P(y = cl | x) &= P(s = s_{none} | x) - 1 + P(y = cl | x) \\ &\quad + \frac{L}{K} [1 - P(y = cl | x)] \\ &= P(s = s_{none} | x) + p(y = cl | x) \\ &\quad - \frac{L}{K} P(y = cl | x) - \frac{K - L}{K} \\ &= P(s = s_{none} | x) + \frac{K - L}{K} p(y = cl | x) - \frac{K - L}{K} \end{aligned} \quad (7)$$

Hence, we can obtain

$$\begin{aligned} P(y = cl | x) &= \frac{K}{L} P(s = s_{none} | x) - \frac{K - L}{L}, \end{aligned} \quad (8)$$

which proves Lemma 2.  $\square$

## 2 PROOF OF THEOREM 3

**Theorem 3.** Under the concealed labels assumption, for multi-class classifier  $f$ , we have  $R_m(f) = R_{CL}(f)$ , where  $R_{CL}(f)$  is defined as

$$\begin{aligned} R_{CL}(f) = & \mathbb{E}_{(x,s) \sim P(x,s \neq s_{none})} \frac{K}{L} \mathcal{L}(f(x), s) \\ & + \mathbb{E}_{(x,s) \sim P(x,s = s_{none})} \frac{K}{L} \mathcal{L}(f(x), cl) \\ & - \mathbb{E}_M \frac{K-L}{L} \mathcal{L}(f(x), cl) \end{aligned}$$

*Proof.* According to the Lemma 2, we have

$$\begin{aligned} R_m(f) &= \mathbb{E}_{x \sim M} \left\{ \sum_{i=1}^K P(y = i | x) \mathcal{L}(f(x), i) \right. \\ &\quad \left. + P(y = cl | x) \mathcal{L}(f(x), cl) \right\} \\ &= \mathbb{E}_{x \sim M} \sum_{i=1}^K \frac{K}{L} P(s = i | x) + \\ &\quad \left[ \frac{K}{L} P(s = s_{none} | x) - \frac{K-L}{L} \mathcal{L}(f(x), cl) \right] \\ &= \mathbb{E}_{(x,s) \sim P(s, s \neq s_{none})} \frac{K}{L} \mathcal{L}(f(x), s) \\ &\quad + \mathbb{E}_{(x,s) \sim P(s, s = s_{none})} \frac{K}{L} \mathcal{L}(f(x), cl) \\ &\quad - \mathbb{E}_M \frac{K-L}{L} \mathcal{L}(f(x), cl) \end{aligned} \quad (9)$$

which concludes the proof.  $\square$

## 3 PROOF OF LEMMA 4

In this section, We analyze the generalization error bounds for the proposed approach, which is implemented using deep neural networks with the OVR strategy. Let  $\mathbf{f} = [f_1, \dots, f_K, f_{cl}]$  denote the classification vector function in the hypothesis set  $\mathcal{F}$ . We assume that there exist a constant  $C_\phi > 0$ , such that  $\sup_{z \in \mathcal{Z}} \phi(z) \leq C_\phi$ . Let  $L_\phi$  be the Lipschitz constant of  $\phi$ , we can introduce the following lemma.

**Lemma 4.** For any  $\delta > 0$ , with the probability at least  $1 - \delta$ ,

$$\begin{aligned} \sup_{\mathbf{f} \in \mathcal{F}} |R_s(\mathbf{f}) - \widehat{R}_s(\mathbf{f})| &\leq 2L_\phi \mathfrak{R}_{n_s}(\mathcal{F}) + 2 \frac{C_\phi K}{L} \sqrt{\frac{\ln(2/\delta)}{2n_s}} \end{aligned} \quad (10)$$

$$\begin{aligned} \sup_{\mathbf{f} \in \mathcal{F}} |R_{none}(\mathbf{f}) - \widehat{R}_{none}(\mathbf{f})| &\leq 2L_\phi \mathfrak{R}_{n_{none}}(\mathcal{F}) + 2 \frac{C_\phi(K-L)}{L} \sqrt{\frac{\ln(2/\delta)}{2n_{none}}} \\ \sup_{\mathbf{f} \in \mathcal{F}} |R_c(\mathbf{f}) - \widehat{R}_c(\mathbf{f})| &\leq 2L_\phi \mathfrak{R}_{n_c}(\mathcal{F}) + 2 \frac{C_\phi K}{L} \sqrt{\frac{\ln(2/\delta)}{2n_c}} \end{aligned}$$

where  $R_s(\mathbf{f}) = \mathbb{E}_{(x,s) \sim P(x,s \neq y_{none})} \frac{K}{L} \mathcal{L}(f(x), s)$ ,  $R_c(\mathbf{f}) = \mathbb{E}_M \frac{K-L}{L} \mathcal{L}(f(x), cl)$  and  $R_{none}(\mathbf{f}) = \mathbb{E}_{(x,s) \sim P(x,s = y_{none})} \frac{K}{L} \mathcal{L}(f(x), cl)$  and  $\widehat{R}_s(\mathbf{f})$  denote the empirical risk estimator to  $R_s(\mathbf{f})$ ,  $R_{none}(\mathbf{f})$  and  $R_c(\mathbf{f})$  respectively,  $\mathfrak{R}_{n_s}(\mathcal{F})$ ,  $\mathfrak{R}_{n_{none}}(\mathcal{F})$  and  $\mathfrak{R}_{n_c}(\mathcal{F})$  are the Rademacher complexities[1] of  $\mathcal{F}$  for the sampling of size  $n_s$  from  $P(x, s \neq y_{none})$ , the sampling of size  $n_{none}$  from  $P(x, s = y_{none})$  and the sampling of size  $n_c$  from  $P(x)$ .

*Proof.* Since the surrogate loss  $\phi(z)$  is bounded by  $\sup_z \phi(z) \leq C_\phi$ , let function  $\Phi_s$  defined for any unconcealed labels sample  $S_s$  by  $\Phi(S_s) = \sup_{\mathbf{f} \in \mathcal{F}} R_s(\mathbf{f}) - \widehat{R}_s(\mathbf{f})$ . If  $x_i$  in unconcealed labels dataset is replaced with  $x'_i$ , the change of  $\Phi_s(S_s)$  does not exceed the supremum of the difference, we have

$$\Phi_s(S'_s) - \Phi_s(S_s) \leq \frac{C_\phi K}{n_s L} \quad (11)$$

Then, by McDiarmid's inequality, for any  $\delta > 0$ , with probability at least  $1 - \delta$ , the following holds:

$$\sup_{\mathbf{f} \in \mathcal{F}} |\widehat{R}_s(\mathbf{f}) - R_s(\mathbf{f})| \leq \mathbb{E}_{S_s} \Phi_s(S_s) + \frac{C_\phi K}{L} \sqrt{\frac{\ln(2/\delta)}{2n_s}} \quad (12)$$

Hence, by using the Rademacher complexity [1], we can obtain

$$\sup_{\mathbf{f} \in \mathcal{F}} |\widehat{R}_s(\mathbf{f}) - R_s(\mathbf{f})| \leq 2\mathfrak{R}_{n_s}(\widehat{l}_s \circ \mathcal{F}) + 2 \frac{C_\phi K}{L} \sqrt{\frac{\ln(2/\delta)}{2n_s}} \quad (13)$$

where  $\mathfrak{R}_{n_s}(\widehat{l}_s \circ \mathcal{F})$  is the Rademacher complexity of the composite function class  $(\widehat{l}_s \circ \mathcal{F})$  for examples size  $n_s$ . As  $L_\phi$  is the Lipschitz constant of  $\phi$ , we have  $\mathfrak{R}_{n_s}(\widehat{l}_s \circ \mathcal{F}) \leq L_\phi \mathfrak{R}_{n_s}(\mathcal{F})$  by Talagrand's contraction Lemma [1]. Then, we can obtain the

$$\sup_{\mathbf{f} \in \mathcal{F}} |R_s(\mathbf{f}) - \widehat{R}_s(\mathbf{f})| \leq 2L_\phi \mathfrak{R}_{n_s}(\mathcal{F}) + 2 \frac{C_\phi K}{L} \sqrt{\frac{\ln(2/\delta)}{2n_s}} \quad (14)$$

Then,  $\sup_{\mathbf{f} \in \mathcal{F}} |R_{none}(\mathbf{f}) - \widehat{R}_{none}(\mathbf{f})|$  and  $\sup_{\mathbf{f} \in \mathcal{F}} |R_c(\mathbf{f}) - \widehat{R}_c(\mathbf{f})|$  can be proven using the same proof technique, which proves Lemma 4.  $\square$

## 4 PROOF OF THEOREM 5

Based on the Lemma 4, we can obtain the generalization error bound as follows.

**Theorem 5.** For any  $\delta > 0$ , with the probability at least  $1 - \delta$ ,

$$\begin{aligned} R_{CL}(\hat{\mathbf{f}}_{CL}) - \min_{\mathbf{f} \in \mathcal{F}} R_{CL}(\mathbf{f}) &\leq \\ &4L_\phi \mathfrak{R}_{n_s}(\mathcal{F}) + 4L_\phi \mathfrak{R}_{n_{none}}(\mathcal{F}) + 4L_\phi \mathfrak{R}_{n_c}(\mathcal{F}) \\ &+ 4 \frac{C_\phi K}{L} \sqrt{\frac{\ln(2/\delta)}{2n_s}} + 4 \frac{C_\phi(K-L)}{L} \sqrt{\frac{\ln(2/\delta)}{2n_{none}}} \\ &+ 4 \frac{C_\phi K}{L} \sqrt{\frac{\ln(2/\delta)}{2n_c}} \end{aligned} \quad (15)$$

where  $\hat{\mathbf{f}}_{CL}$  is trained by minimizing the classification risk  $R_{CL}$ .

*Proof.* According to Lemma 4, the estimation error bound is proven through

$$\begin{aligned} R_{CL}(\hat{\mathbf{f}}_{CL}) - R_{CL}(\mathbf{f}^*) &= (\hat{R}_{CL}(\hat{\mathbf{f}}_{CL}) - \hat{R}_{CL}(\hat{\mathbf{f}}^*)) \\ &\quad + (R(\hat{\mathbf{f}}_{CL}) - \hat{R}_{CL}(\hat{\mathbf{f}}_{CL})) \\ &\quad + (\hat{R}_{CL}(\hat{\mathbf{f}}^*) - R(\hat{\mathbf{f}}^*)) \\ &\leq 0 + 2\sup_{\mathbf{f} \in \mathcal{F}} |R_{CL}(\mathbf{f}) - \hat{R}_{CL}(\mathbf{f})| \end{aligned} \quad (16)$$

where  $\mathbf{f}^* = \arg \min_{\mathbf{f} \in \mathcal{F}} R(\mathbf{f})$ .

Now, we have seen the definition of  $R_{CL}(\mathbf{f})$  and  $\hat{R}_{CL}(\mathbf{f})$  that can also be decomposed into:

$$\begin{aligned} R_{CL}(f) &= \mathbb{E}_{(x,s) \sim P(x,s \neq y_{none})} \frac{K}{L} \mathcal{L}(f(x), s) \\ &\quad + \mathbb{E}_{(x,s) \sim P(x,s = y_{none})} \frac{K}{L} \mathcal{L}(f(x), cl) \\ &\quad - \mathbb{E}_M \frac{K-L}{L} \mathcal{L}(f(x), cl) \end{aligned} \quad (17)$$

and

$$\begin{aligned} \hat{R}_{CL}(f) &= \frac{1}{\#\{\mathcal{X}_s\}_s^K} \sum_{s=1}^K \sum_{x_j \in \mathcal{X}_s} \frac{K}{L} \mathcal{L}(f(x_j), s) \\ &\quad + \frac{1}{\#\mathcal{X}_{none}} \sum_{x_j \in \mathcal{X}_{none}} \frac{K}{L} \mathcal{L}(f(x_j), cl) \\ &\quad - \frac{1}{\#\mathcal{X}_c} \sum_{x_j \in \mathcal{X}_c} \frac{K-L}{L} \mathcal{L}(f(x_j), cl). \end{aligned} \quad (18)$$

Due to the sub-additivity of the supremum operators, it holds that

$$\begin{aligned} \sup_{\mathbf{f} \in \mathcal{F}} |\hat{R}_{CL}(\mathbf{f}) - R_{CL}(\mathbf{f})| &\leq \sup_{\mathbf{f} \in \mathcal{F}} |\hat{R}_s(\mathbf{f}) - R_s(\mathbf{f})| \\ &\quad + \sup_{\mathbf{f} \in \mathcal{F}} |\hat{R}_{none}(\mathbf{f}) - R_{none}(\mathbf{f})| \\ &\quad + \sup_{\mathbf{f} \in \mathcal{F}} |\hat{R}_c(\mathbf{f}) - R_c(\mathbf{f})| \end{aligned} \quad (19)$$

where

$$\begin{aligned} R_s(\mathbf{f}) &= \mathbb{E}_{(x,s) \sim P(x,s \neq y_{none})} \frac{K}{L} \mathcal{L}(f(x), s), R_{none}(\mathbf{f}) \\ &= \mathbb{E}_{(x,s) \sim P(x,s = y_{none})} \frac{K}{L} \mathcal{L}(f(x), cl), \end{aligned} \quad (20)$$

$$\hat{R}_s(\mathbf{f}) = \frac{1}{\#\{\mathcal{X}_s\}_s^K} \sum_{s=1}^K \sum_{x_j \in \mathcal{X}_s} \frac{K}{L} \mathcal{L}(f(x_j), s), \quad (21)$$

$$\begin{aligned} R_{none}(\mathbf{f}) &= \mathbb{E}_{(x,s) \sim P(x,s = y_{none})} \frac{K}{L} \mathcal{L}(f(x), cl) \hat{R}_{none}(\mathbf{f}) \\ &= \frac{1}{\#\mathcal{X}_{none}} \sum_{x_j \in \mathcal{X}_{none}} \frac{K}{L} \mathcal{L}(f(x_j), cl) \end{aligned} \quad (22)$$

$$\text{and } R_c(\mathbf{f}) = \mathbb{E}_M \frac{K-L}{L} \mathcal{L}(f(x), cl), \hat{R}_c(\mathbf{f}) = \frac{1}{\#\mathcal{X}_c} \sum_{x_j \in \mathcal{X}_c} \frac{K-L}{L} \mathcal{L}(f(x_j), cl).$$

According to the Lemma 4, we can get the generalization bound that

$$\begin{aligned} R_{CL}(\hat{\mathbf{f}}_{CL}) - \min_{\mathbf{f} \in \mathcal{F}} R_{CL}(\mathbf{f}) &\leq \\ &\quad 4L_\phi \mathfrak{R}_{n_s}(\mathcal{F}) + 4L_\phi \mathfrak{R}_{n_{none}}(\mathcal{F}) + 4L_\phi \mathfrak{R}_{n_c}(\mathcal{F}) \\ &\quad + 4 \frac{C_\phi K}{L} \sqrt{\frac{\ln(2/\delta)}{2n_s}} + 4 \frac{C_\phi (K-L)}{L} \sqrt{\frac{\ln(2/\delta)}{2n_{none}}} \\ &\quad + 4 \frac{C_\phi K}{L} \sqrt{\frac{\ln(2/\delta)}{2n_c}} \end{aligned} \quad (23)$$

with probability at least  $1 - \delta$ , which finishes the proof.  $\square$

## REFERENCES

- [1] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. 2018. *Foundations of machine learning*.