A Interpretation from Objective Functions

In this section, we provide proofs of the Onehot$(\cdot)$ normalization function and the Scale$(\cdot)$ normalization function from the perspective of objective functions.

A.1 Proof for Onehot Normalization

For $K = 0$, we choose the following objective function during training:

$$\begin{aligned} &\max \sum_{i=1}^{c} w_i \log P_i + M\left(\sum_{i=1}^{c} w_i S_i - 1\right) \\ &s.t. \sum_{i}^{c} w_i = 1, w_i \geq 0. \end{aligned} \tag{1}$$

Introduce Lagrange multipliers $\delta_i, i \in [1, c]$ and γ into Eq. 1, we have:

$$\mathcal{L} = \sum_{i=1}^{c} w_i \log P_i + M\left(\sum_{i=1}^{c} w_i S_i - 1\right) + \gamma\left(1 - \sum_{i=1}^{c} w_i\right) + \sum_{i=1}^{c} \delta_i w_i. \tag{2}$$

Combined with the Karush-Kuhn-Tucker (KKT) conditions, the optimal point should satisfy:

$$\log P_i + MS_i - \gamma + \delta_i = 0, \tag{3}$$

$$\sum_{i=1}^{c} w_i = 1, \delta_i \geq 0, w_i \geq 0, \delta_i w_i = 0. \tag{4}$$

Since $S_i \in \{0, 1\}$, we have $MS_i = \log(e^M S_i + (1 - S_i))$. The equivalent equation of Eq. 3 is:

$$\delta_i = \gamma - \log(e^M S_i + (1 - S_i))P_i. \tag{5}$$

Combined with $\delta_i \geq 0$ in Eq. 4, we have:

$$\gamma \geq \max_i \left(\log(e^M S_i + (1 - S_i))P_i\right). \tag{6}$$

$\delta_i > 0$ is true if $\gamma > \max_i \left(\log(e^M S_i + (1 - S_i))P_i\right)$. According to $\delta_i w_i = 0$, we always have $w_i = 0$, which conflicts with $\sum_{i=1}^{c} w_i = 1$. Therefore, we get:

$$\gamma = \max_i \left(\log(e^M S_i + (1 - S_i))P_i\right). \tag{7}$$

We assume that only one $i_0 \in [1, c]$ reaches the maximum γ, then we have $w_i = 0, i \in [1, c]/i_0$. Combined with $\sum_{i=1}^{c} w_i = 1$, we get $w_{i_0} = 1$. Therefore, $w(x)$ should satisfy:

$$w(x) = \text{Onehot}\left(\log(e^M S(x) + (1 - S(x)))P(x)\right). \tag{8}$$

We mark $\lambda = e^{-M}$ and convert Eq. 8 to its equivalent version:

$$w(x) = \text{Onehot}\left((S(x) + \lambda(1 - S(x))P(x)\right). \tag{9}$$

A.2 Proof for Scale Normalization

For $K > 0$, $\log w_i$ ensures that w_i must be positive. Therefore, the constraint $w_i \geq 0$ can be excluded. Then, the objective function can be converted to:

$$\begin{aligned} &\max \sum_{i=1}^{c} w_i \log P_i + M\left(\sum_{i=1}^{c} w_i S_i - 1\right) - K\sum_{i=1}^{c} w_i \log w_i \\ &s.t. \sum_{i}^{c} w_i = 1. \end{aligned} \tag{10}$$

Introduce the Lagrange multiplier γ in Eq. 10, we have:

$$\mathcal{L} = \sum_{i=1}^{c} w_i \log P_i + M\left(\sum_{i=1}^{c} w_i S_i - 1\right) - K \sum_{i=1}^{c} w_i \log w_i + \gamma\left(1 - \sum_{i}^{c} w_i\right). \tag{11}$$

Since the optimal point should satisfy $\nabla_w \mathcal{L} = 0$, we have:

$$\log P_i + MS_i - K\left(1 + \log w_i\right) - \gamma = 0. \tag{12}$$

Since $S_i \in \{0, 1\}$, we have $MS_i = \log(e^M S_i + (1 - S_i))$. The equivalent equation of Eq. 12 is:

$$\log(e^M S_i + (1 - S_i))P_i - (K + \gamma) - K \log w_i = 0, \tag{13}$$

$$w_i^K = \frac{\left(e^M S_i + (1 - S_i)\right) P_i}{e^{K+\gamma}}. \tag{14}$$

We mark $\lambda = e^{-M}$. Then, we have:

$$w_i = \frac{\left((S_i + \lambda(1 - S_i))P_i\right)^{1/K}}{e^{1+(\gamma-M)/K}}. \tag{15}$$

Since $\sum_i^c w_i = 1$, we have:

$$\sum_{i=1}^{c} \frac{\left((S_i + \lambda(1 - S_i))P_i\right)^{1/K}}{e^{1+(\gamma-M)/K}} = 1, \tag{16}$$

$$e^{1+(\gamma-M)/K} = \sum_{i=1}^{c} \left((S_i + \lambda\left(1 - S_i\right))P_i\right)^{1/K}. \tag{17}$$

Combine Eq. 15 and Eq. 17 and we have:

$$w_i = \frac{\left((S_i + \lambda(1 - S_i))P_i\right)^{1/K}}{\sum_{i=1}^{c} \left((S_i + \lambda(1 - S_i))P_i\right)^{1/K}}. \tag{18}$$

Combined with the definition of $\mathrm{Scale}(\cdot)$, this equation can be converted to:

$$w(x) = \mathrm{Scale}\left((S(x) + \lambda(1 - S(x)))P(x)\right). \tag{19}$$

B EM Perspective of ALIM

EM aims to maximize the likelihood of the dataset $\mathcal{D}$:

$$\begin{aligned}
\max_\theta \sum_{x \in \mathcal{D}} \log P\left(x, S(x); \theta\right) &= \max_\theta \sum_{x \in \mathcal{D}} \log \sum_{i=1}^{c} P\left(x, S(x), y(x) = i; \theta\right) \\
&= \max_\theta \sum_{x \in \mathcal{D}} \log \sum_{i=1}^{c} w_i(x) \frac{P\left(x, S(x), y(x) = i; \theta\right)}{w_i(x)} \\
&\geq \max_\theta \sum_{x \in \mathcal{D}} \sum_{i=1}^{c} w_i(x) \log \frac{P\left(x, S(x), y(x) = i; \theta\right)}{w_i(x)},
\end{aligned} \tag{20}$$

where θ is the trainable parameter. The last step of Eq. 20 utilizes Jensen's inequality. Since the $\log(\cdot)$ function is strictly concave, the equal sign takes when $P(x, S(x), y(x) = i; \theta)/w_i(x)$ is some constant C, i.e.,

$$w_i(x) = \frac{1}{C} P(x, S(x), y(x) = i; \theta). \tag{21}$$

Considering that $\sum_{i=1}^{c} w_i(x) = 1$, we can further get:

$$C = \sum_{i=1}^{c} P(x, S(x), y(x) = i; \theta). \tag{22}$$

Then, we have:

$$w_i(x) = \frac{P(x, S(x), y(x) = i; \theta)}{\sum_{i=1}^{c} P(x, S(x), y(x) = i; \theta)} = \frac{P(x, S(x), y(x) = i; \theta)}{P(x, S(x); \theta)} = P(y(x) = i | x, S(x); \theta). \tag{23}$$

In the EM algorithm, the E-step aims to calculate $w_i(x)$ and the M-step aims to maximize the lower bound of Eq. 20:

$$\begin{aligned} &\operatorname*{argmax}_{\theta} \sum_{x \in \mathcal{D}} \sum_{i=1}^{c} w_i(x) \log \frac{P(x, S(x), y(x) = i; \theta)}{w_i(x)} \\ = &\operatorname*{argmax}_{\theta} \sum_{x \in \mathcal{D}} \sum_{i=1}^{c} w_i(x) \log P(x, S(x), y(x) = i; \theta). \end{aligned} \tag{24}$$

E-Step. In this step, we aim to predict the ground-truth label for each sample:

$$\begin{aligned} w_i(x) = P(y(x) = i | x, S(x); \theta) &= \frac{P(S(x) | y(x) = i, x; \theta) P(y(x) = i | x; \theta)}{P(S(x) | x; \theta)} \\ &= \frac{P(S(x) | y(x) = i, x; \theta) P(y(x) = i | x; \theta)}{\sum_{i=1}^{c} P(S(x) | y(x) = i, x; \theta) P(y(x) = i | x; \theta)}. \end{aligned} \tag{25}$$

According to Assumption 1, we have:

$$P(S(x) | y(x), x) = \begin{cases} \alpha(x), S_{y(x)}(x) = 1 \\ \beta(x), S_{y(x)}(x) = 0. \end{cases} \tag{26}$$

It can be converted to:

$$P(S(x) | y(x), x) = \alpha(x) S_{y(x)}(x) + \beta(x) \left(1 - S_{y(x)}(x)\right). \tag{27}$$

Then, we get the equivalent equation of Eq. 25:

$$w_i(x) = \frac{(\alpha(x) S_i(x) + \beta(x)(1 - S_i(x))) P(y(x) = i | x; \theta)}{\sum_{i=1}^{c} (\alpha(x) S_i(x) + \beta(x)(1 - S_i(x))) P(y(x) = i | x; \theta)}. \tag{28}$$

We mark $\lambda(x) = \beta(x)/\alpha(x)$ and $P_i(x) = P(y(x) = i | x; \theta)$. Then, we get:

$$w_i(x) = \frac{(S_i(x) + \lambda(x)(1 - S_i(x))) P_i(x)}{\sum_{i=1}^{c} (S_i(x) + \lambda(x)(1 - S_i(x))) P_i(x)}. \tag{29}$$

It connects traditional PLL and noisy PLL. In traditional PLL, we assume that the ground-truth label must be in the candidate set, i.e., $\beta(x) = 0$. Since $\lambda(x) = \beta(x)/\alpha(x) = 0$, Eq. 29 degenerates to:

$$w_i(x) = \frac{S_i(x) P_i(x)}{\sum_{i=1}^{c} S_i(x) P_i(x)}, \tag{30}$$

which is identical to the classic PLL method, RC.

M-Step. The objective function of this step is:

$$\begin{aligned} &\operatorname*{argmax}_{\theta} \sum_{x \in \mathcal{D}} \sum_{i=1}^{c} w_i(x) \log P(x, S(x), y(x) = i; \theta) \\ =&\operatorname*{argmax}_{\theta} \sum_{x \in \mathcal{D}} \sum_{i=1}^{c} w_i(x) \log P(x; \theta) P(y(x) = i | x; \theta) P(S(x) | y(x) = i, x; \theta). \end{aligned} \tag{31}$$

Considering that $P(x;\theta) = P(x)$ and $P(S(x)|y(x) = i, x; \theta) = P(S(x)|y(x) = i, x)$, the equivalent version of Eq. 31 is:

$$\underset{\theta}{\operatorname{argmax}} \sum_{x \in \mathcal{D}} \sum_{i=1}^{c} w_i(x) \log P(y(x) = i | x; \theta). \tag{32}$$

Therefore, the essence of the M-step is to minimize the classification loss.

C Adaptively Adjusted λ

Since η controls the noise level of the dataset, we have:

$$P(S_{y(x)}(x) = 0) = \eta. \tag{33}$$

After the warm-up training, we assume that the predicted label generated by ALIM $\hat{y}(x) = \arg\max_{1 \leq i \leq c} w(x)$ is accurate, i.e., $\hat{y}(x) = y(x)$. Then we have:

$$P(S_{\hat{y}(x)}(x) = 0) = \eta. \tag{34}$$

To estimate the value of λ, we first study the equivalent meaning of $S_{\hat{y}(x)}(x) = 0$:

$$\max_{S_i(x)=0} \left(S_i(x) + \lambda \left(1 - S_i(x)\right)\right) P_i(x) \geq \max_{S_i(x)=1} \left(S_i(x) + \lambda \left(1 - S_i(x)\right)\right) P_i(x). \tag{35}$$

We simplify the left and right sides of Eq.35 as follows:

$$\begin{aligned} &\max_{S_i(x)=0} (S_i(x) + \lambda(1 - S_i(x)))P_i(x) \\ =& \max_{S_i(x)=0} \lambda(1 - S_i(x))P_i(x) \\ =& \max_i \lambda(1 - S_i(x))P_i(x), \end{aligned} \tag{36}$$

$$\begin{aligned} &\max_{S_i(x)=1} (S_i(x) + \lambda(1 - S_i(x)))P_i(x) \\ =& \max_{S_i(x)=1} S_i(x)P_i(x) \\ =& \max_i S_i(x)P_i(x). \end{aligned} \tag{37}$$

Then, we have:

$$\max_i \lambda(1 - S_i(x))P_i(x) \geq \max_i S_i(x)P_i(x), \tag{38}$$

$$\lambda \geq \frac{\max_i S_i(x)P_i(x)}{\max_i (1 - S_i(x))P_i(x)}. \tag{39}$$

Therefore, $P(S_{\hat{y}(x)}(x) = 0) = \eta$ can be converted to:

$$P\left(\lambda \geq \frac{\max_i \ S_i(x)P_i(x)}{\max_i \ (1 - S_i(x))P_i(x)}\right) = \eta. \tag{40}$$

It means that λ is the η-quantile of

$$\left\{\frac{\max_i S_i(x)P_i(x)}{\max_i (1 - S_i(x))P_i(x)}\right\}_{x \in \mathcal{D}}. \tag{41}$$