Multi instance multi label learning in the presence of novel class instances: Supplementary Material

1. Surrogate function calculation

In this section, we show the steps to compute the surrogate function. In our setting, the observed data is $\{\mathbf{Y}_D, \mathbf{X}_D\}$, the parameter is $\mathbf{w}$, and the hidden data $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_B\}$. To compute the surrogate $g(\mathbf{w}, \mathbf{w}')$, we begin with the derivation of the complete log-likelihood. We apply the conditional rule as follows

$$\begin{aligned} p(\mathbf{Y}_D, \mathbf{X}_D, \mathbf{y}|\mathbf{w}) &= p(\mathbf{Y}_D|\mathbf{y}, \mathbf{X}_D, \mathbf{w})p(\mathbf{y}|\mathbf{X}_D, \mathbf{w})p(\mathbf{X}_D|\mathbf{w}) \\ &= p(\mathbf{Y}_D|\mathbf{y})[\prod_{b=1}^{B}\prod_{i=1}^{n_b} p(y_{bi}|\mathbf{x}_{bi}, \mathbf{w})]p(\mathbf{X}_D). \quad (1) \end{aligned}$$

We recall the relation between the instance label and feature vector, including novel class, as follows

$$p(y_{bi}|\mathbf{x}_{bi}, \mathbf{w}) = \frac{\prod_{c=0}^{C} e^{I(y_{bi}=c)\mathbf{w}_c^T\mathbf{x}_{bi}}}{\sum_{c=0}^{C} e^{\mathbf{w}_c^T\mathbf{x}_{bi}}}. \quad (2)$$

Then, the complete log-likelihood can be computed by taking the logarithm of (1), replacing $p(y_{bi}|\mathbf{x}_{bi}, \mathbf{w})$ from (2) into (1), and reorganizing as follows

$$\begin{aligned} \log p(\mathbf{Y}_D, \mathbf{X}_D, \mathbf{y}|\mathbf{w}) &= \sum_{b=1}^{B}\sum_{i=1}^{n_b}\sum_{c=0}^{C} I(y_{bi}=c)\mathbf{w}_c^T\mathbf{x}_{bi} \quad (3) \\ -\sum_{b=1}^{B}\sum_{i=1}^{n_b}\log(\sum_{c=0}^{C} e^{\mathbf{w}_c^T\mathbf{x}_{bi}}) &+ \log p(\mathbf{Y}_D|\mathbf{y}) + \log p(\mathbf{X}_D). \end{aligned}$$

Finally, taking the expectation of (3) w.r.t. $\mathbf{y}$ given $\mathbf{Y}_D, \mathbf{X}_D$, and $\mathbf{w}'$, we obtain the surrogate function $g(\cdot,\cdot)$ as follows

$$\begin{aligned} g(\mathbf{w}, \mathbf{w}') &= E_{\mathbf{y}}[\log p(\mathbf{Y}_D, \mathbf{X}_D, \mathbf{y}|\mathbf{w})|\mathbf{Y}_D, \mathbf{X}_D, \mathbf{w}'] \quad (4) \\ &= \sum_{b=1}^{B}\sum_{i=1}^{n_b}[\sum_{c=0}^{C} p(y_{bi}=c|\mathbf{Y}_b, \mathbf{X}_b, \mathbf{w}')\mathbf{w}_c^T\mathbf{x}_{bi} \\ &\quad - \log(\sum_{c=0}^{C} e^{\mathbf{w}_c^T\mathbf{x}_{bi}})] + \zeta, \end{aligned}$$

where $\zeta = E_{\mathbf{y}}[\log p(\mathbf{Y}_D|\mathbf{y})|\mathbf{Y}_D, \mathbf{X}_D, \mathbf{w}'] + \log p(\mathbf{X}_D)$ is a constant w.r.t. $\mathbf{w}$.

2. Proof for Proposition 1

In this section, we show the detailed proof for Proposition 1 of computing $p(y_{bn_b}, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w})$ from $p(\mathbf{Y}_b^{n_b-1}|\mathbf{X}_b, \mathbf{w})$ and $p(y_{bn_b} = c|\mathbf{x}_{bn_b}, \mathbf{w})$.

Proposition 1 *The probability $p(y_{bn_b} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w})$ for all $c \in \mathbf{L}\bigcup\{0\}$ can be computed using*

- *If $c = 0$,*

$$\begin{aligned} &p(y_{bn_b} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w}) = p(y_{bn_b} = c|\mathbf{x}_{bn_b}, \mathbf{w})\times \\ &[p(\mathbf{Y}_b^{n_b-1} = \mathbf{L}|\mathbf{X}_b, \mathbf{w}) + p(\mathbf{Y}_b^{n_b-1} = \mathbf{L}\bigcup\{0\}|\mathbf{X}_b, \mathbf{w})]. \end{aligned}$$

- *Else if $c \neq 0$,*

$$\begin{aligned} &p(y_{bn_b} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w}) = p(y_{bn_b} = c|\mathbf{x}_{bn_b}, \mathbf{w})\times \\ &[p(\mathbf{Y}_b^{n_b-1} = \mathbf{L}|\mathbf{X}_b, \mathbf{w}) + p(\mathbf{Y}_b^{n_b-1} = \mathbf{L}_{\backslash c}|\mathbf{X}_b, \mathbf{w}) + \\ &p(\mathbf{Y}_b^{n_b-1} = \mathbf{L}\bigcup\{0\}|\mathbf{X}_b, \mathbf{w}) + \\ &p(\mathbf{Y}_b^{n_b-1} = \mathbf{L}_{\backslash c}\bigcup\{0\}|\mathbf{X}_b, \mathbf{w})]. \end{aligned}$$

Proof. Denote the power set of $\mathbf{L}\bigcup\{0\}$ excluding the empty set as $\mathbf{P}$. We compute $p(y_{bn_b} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w})$ by marginalizing $p(y_{bn_b}, \mathbf{Y}_b = \mathbf{L}, \mathbf{Y}_b^{n_b} = \mathbf{L}'|\mathbf{X}_b, \mathbf{w})$ over $\mathbf{Y}_b^{n_b}$ as follows

$$\begin{aligned} &p(y_{bn_b} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w}) \\ &= \sum_{\mathbf{L}'\subseteq\mathbf{P}} p(y_{bn_b} = c, \mathbf{Y}_b = \mathbf{L}, \mathbf{Y}_b^{n_b} = \mathbf{L}'|\mathbf{X}_b, \mathbf{w}). \quad (5) \end{aligned}$$

Using conditional probability rule for the right hand side of (5) we obtain

$$\begin{aligned} &p(y_{bn_b} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w}) \quad (6) \\ &= \sum_{\mathbf{L}'\subseteq\mathbf{P}} p(y_{bn_b} = c, \mathbf{Y}_b^{n_b} = \mathbf{L}'|\mathbf{X}_b, \mathbf{w})p(\mathbf{Y}_b = \mathbf{L}|\mathbf{Y}_b^{n_b} = \mathbf{L}'). \end{aligned}$$

From the proposed model, $p(\mathbf{Y}_b = \mathbf{L}|\mathbf{Y}_b^{n_b} = \mathbf{L}') = I(\mathbf{L} = \mathbf{L}') + I(\mathbf{L}\bigcup\{0\} = \mathbf{L}')$. Replacing $p(\mathbf{Y}_b = \mathbf{L}|\mathbf{Y}_b^{n_b} = \mathbf{L}')$ into (6) we obtain

$$\begin{aligned} p(y_{bn_b} = c, \mathbf{Y}_b = \mathbf{L}|\mathbf{X}_b, \mathbf{w}) &= p(y_{bn_b} = c, \mathbf{Y}_b^{n_b} = \mathbf{L}|\mathbf{X}_b, \mathbf{w}) \\ &+ p(y_{bn_b} = c, \mathbf{Y}_b^{n_b} = \mathbf{L}\bigcup\{0\}|\mathbf{X}_b, \mathbf{w}). \quad (7) \end{aligned}$$

• For $c \neq 0$: The first term in the right hand side of (7), $p(y_{bn_b} = c, \mathbf{Y}_b^{n_b} = \mathbf{L}|\mathbf{X}_b, \mathbf{w})$, is computed by marginalizing $p(y_{bn_b} = c, \mathbf{Y}_b^{n_b} = \mathbf{L}, \mathbf{Y}_b^{n_b-1} = \mathbf{L}'|\mathbf{X}_b, \mathbf{w})$ over $\mathbf{Y}_b^{n_b-1}$ as follows

$$\begin{aligned} &p(y_{bn_b} = c, \mathbf{Y}_b^{n_b} = \mathbf{L}|\mathbf{X}_b, \mathbf{w}) \\ &= \sum_{\mathbf{L}' \subseteq \mathbf{P}} p(y_{bn_b} = c, \mathbf{Y}_b^{n_b} = \mathbf{L}, \mathbf{Y}_b^{n_b-1} = \mathbf{L}'|\mathbf{X}_b, \mathbf{w}). \end{aligned} \tag{8}$$

Using the conditional probability rule we have

$$\begin{aligned} &p(y_{bn_b} = c, \mathbf{Y}_b^{n_b} = \mathbf{L}, \mathbf{Y}_b^{n_b-1} = \mathbf{L}'|\mathbf{X}_b, \mathbf{w}) \\ &= p(y_{bn_b} = c, \mathbf{Y}_b^{n_b-1} = \mathbf{L}'|\mathbf{X}_b, \mathbf{w})\times \\ &\quad p(\mathbf{Y}_b^{n_b} = \mathbf{L}|y_{bn_b} = c, \mathbf{Y}_b^{n_b-1} = \mathbf{L}'). \end{aligned} \tag{9}$$

Replacing $p(y_{bn_b} = c, \mathbf{Y}_b^{n_b} = \mathbf{L}, \mathbf{Y}_b^{n_b-1} = \mathbf{L}'|\mathbf{X}_b, \mathbf{w})$ into (8) we obtain

$$\begin{aligned} &p(y_{bn_b} = c, \mathbf{Y}_b^{n_b} = \mathbf{L}|\mathbf{X}_b, \mathbf{w}) \\ &= \sum_{\mathbf{L}' \subseteq \mathbf{P}} [p(y_{bn_b} = c, \mathbf{Y}_b^{n_b-1} = \mathbf{L}'|\mathbf{X}_b, \mathbf{w})\times \\ &\qquad p(\mathbf{Y}_b^{n_b} = \mathbf{L}|y_{bn_b} = c, \mathbf{Y}_b^{n_b-1} = \mathbf{L}')]. \end{aligned} \tag{10}$$

From the proposed model we have $p(\mathbf{Y}_b^{n_b} = \mathbf{L}|y_{bn_b} = c, \mathbf{Y}_b^{n_b-1} = \mathbf{L}') = I(\mathbf{L} = \mathbf{L}' \bigcup \{c\})$. Moreover, given instance features, instance labels are independent. Consequently, from (10), we obtain

$$\begin{aligned} &p(y_{bn_b} = c, \mathbf{Y}_b^{n_b} = \mathbf{L}|\mathbf{X}_b, \mathbf{w}) \\ &= \sum_{\mathbf{L}' \subseteq \mathbf{P}} [p(y_{bn_b} = c|\mathbf{x}_{bn_b}, \mathbf{w})\times \\ &\qquad p(\mathbf{Y}_b^{n_b-1} = \mathbf{L}'|\mathbf{X}_b, \mathbf{w}) I(\mathbf{L} = \mathbf{L}' \bigcup \{c\})] \\ &= p(y_{bn_b} = c|\mathbf{x}_{bn_b}, \mathbf{w})\times \\ &\quad [p(\mathbf{Y}_b^{n_b-1} = \mathbf{L}|\mathbf{X}_b, \mathbf{w}) + p(\mathbf{Y}_b^{n_b-1} = \mathbf{L}_{\setminus c}|\mathbf{X}_b, \mathbf{w})]. \end{aligned} \tag{11}$$

Deriving similar steps from (8) to (11) for the second term of (7), $p(y_{bn_b} = c, \mathbf{Y}_b^{n_b} = \mathbf{L} \bigcup \{0\}|\mathbf{X}_b, \mathbf{w})$, we obtain

$$\begin{aligned} &p(y_{bn_b} = c, \mathbf{Y}_b^{n_b} = \mathbf{L} \bigcup \{0\}|\mathbf{X}_b, \mathbf{w}) \\ &= p(y_{bn_b} = c|\mathbf{x}_{bn_b}, \mathbf{w})\times \\ &[p(\mathbf{Y}_b^{n_b-1} = \mathbf{L} \bigcup \{0\}|\mathbf{X}_b, \mathbf{w}) + p(\mathbf{Y}_b^{n_b-1} = \mathbf{L}_{\setminus c} \bigcup \{0\}|\mathbf{X}_b, \mathbf{w})]. \end{aligned} \tag{12}$$

Replacing probabilities obtained in (11) and (12) into (7), we obtain the proof for the case $c \neq 0$.

• For $c = 0$: Since the bag label $\mathbf{L}$ does not contain novel label 0 and $y_{bn_b} \in \mathbf{Y}_b^{n_b}$, the first term in the right hand side of (7), $p(y_{bn_b} = c, \mathbf{Y}_b^{n_b} = \mathbf{L}|\mathbf{X}_b, \mathbf{w}) = 0$. Replacing probabilities obtained in (12) into (7), we obtain the proof for the case $c = 0$.

□