

# Appendices for Partial-Label Regression

**Xin Cheng<sup>1</sup>, Dengbao Wang<sup>2</sup>, Lei Feng<sup>1\*</sup>, Minling Zhang<sup>2</sup>, Bo An<sup>3</sup>**

<sup>1</sup>College of Computer Science, Chongqing University, Chongqing, China

<sup>2</sup>School of Computer Science and Engineering, Southeast University, Nanjing, China

<sup>3</sup>School of Computer Science and Engineering, Nanyang Technological University, Singapore  
xincheng@stu.cqu.edu.cn, lfeng@cqu.edu.cn, {wangdb, zhangml}@seu.edu.cn, boan@ntu.edu.sg

## A Proof of Theorem 1

First, we prove that the optimal model  $f^*$  learned from fully labeled data (i.e.,  $f^* = \arg \min R(f)$ ) is also the optimal model for  $R_{\min}(f) = \mathbb{E}_{p(\mathbf{x}, S)}[\ell_{\min}(f(\mathbf{x}), S)]$  as follows.

By substituting the  $f^*$  into  $R_{\min}(f)$ , we obtain:

$$\begin{aligned} R_{\min}(f^*) &= \mathbb{E}_{p(\mathbf{x}, S)}[\ell_{\min}(f^*(\mathbf{x}), S)] \\ &= \int_{\mathcal{X}} \int_S \ell_{\min}(f^*(\mathbf{x}), S) p(S | \mathbf{x}) p(\mathbf{x}) dS d\mathbf{x} \\ &= \int_{\mathcal{X}} \int_S \int_Y \ell_{\min}(f^*(\mathbf{x}), S) p(S, y | \mathbf{x}) p(\mathbf{x}) dy dS d\mathbf{x} \\ &= \int_{\mathcal{X}} \int_S \int_Y \min_{y' \in S} \ell(f^*(\mathbf{x}), y') p(S, y | \mathbf{x}) p(\mathbf{x}) dy dS d\mathbf{x} \\ &= \int_{\mathcal{X}} \int_Y \ell(f^*(\mathbf{x}), y) \int_S p(S | y, \mathbf{x}) p(y | \mathbf{x}) p(\mathbf{x}) dS dy d\mathbf{x} \\ &= \int_{\mathcal{X}} \int_Y \ell(f^*(\mathbf{x}), y) p(\mathbf{x}, y) dy d\mathbf{x} \\ &= R(f^*) = 0, \end{aligned}$$

which indicates that  $f^*$  is the optimal model for  $R_{\min}(f)$ .

On the other hand, we prove that  $f^*$  is the sole optimal model for  $R_{\min}(f)$  by contradiction. Specifically, we assume that there is at least one other model  $g$  that makes  $R_{\min}(g) = 0$  and predicts a label  $y_g \neq y$  for at least one instance  $\mathbf{x}$ . Therefore, for any  $S$  containing  $y$ , we have

$$\min_{y' \in S} \ell(g(\mathbf{x}), y') = \ell(g(\mathbf{x}), y_g) = 0.$$

The above equality implies that  $y_g$  is always included in the candidate label set of  $\mathbf{x}$  (co-occurring with the true label  $y$ ), and in this case, the ambiguity degree is 1. This contradicts the basic PLR assumption that the ambiguity degree should be less than 1. Therefore, there is one, and only one minimizer of  $R_{\min}$ , which is the same as the minimizer  $f^*$  learned from fully labeled data. The proof is completed.  $\square$

## B Proof of Theorem 2

Let us introduce the following notations:

$$\begin{aligned} d &= \text{Pdim}(\{\mathbf{x} \mapsto \ell(f(\mathbf{x}), y) \mid f \in \mathcal{F}\}), \\ d' &= \text{Pdim}(\{\mathbf{x} \mapsto \min_{y \in S} \ell(f(\mathbf{x}), y) \mid f \in \mathcal{F}\}), \end{aligned}$$

where  $\text{Pdim}(\mathcal{F})$  denotes the pseudo-dimension of the functional space  $\mathcal{F}$ . It is worth noting that we may represent  $d'$  by  $d$  with some derivations, while for simplicity and convenience, we directly formulate the expression of  $d'$ .

From the assumptions in Theorem 2, using the discussion in Theorem 10.6 of Mohri, Rostamizadeh, and Talwalkar (2012), with probability  $1 - \delta$  for all  $f \in \mathcal{F}$ ,

$$\begin{aligned} \left| \mathbb{E}_{p(\mathbf{x}, S)}[\ell_{\min}(\mathbf{x}, S)] - \sum_{i=1}^n \ell_{\min}(\mathbf{x}_i, S_i) \right| \\ \leq M \sqrt{\frac{2d' \log \frac{ne}{d'}}{n}} + M \sqrt{\frac{\log \frac{1}{\delta}}{2n}}. \end{aligned}$$

Hence we can know that  $|R_{\min}(f) - \hat{R}_{\min}(f)|$  converges in the order of  $\mathcal{O}(1/\sqrt{n})$  for all  $f \in \mathcal{F}$ .

Then, we have the following derivations:

$$\begin{aligned} R_{\min}(\hat{f}_{\min}) - R_{\min}(f^*) &\leq R_{\min}(\hat{f}_{\min}) - \hat{R}_{\min}(\hat{f}_{\min}) + \hat{R}_{\min}(\hat{f}_{\min}) - \hat{R}_{\min}(f^*) \\ &\quad + \hat{R}_{\min}(f^*) - R_{\min}(f^*) \\ &\leq R_{\min}(\hat{f}_{\min}) - \hat{R}_{\min}(\hat{f}_{\min}) + \hat{R}_{\min}(f^*) - R_{\min}(f^*) \\ &\leq |R_{\min}(\hat{f}_{\min}) - \hat{R}_{\min}(\hat{f}_{\min})| \\ &\quad + |\hat{R}_{\min}(f^*) - R_{\min}(f^*)|, \end{aligned}$$

where the first inequality holds because  $\hat{R}_{\min}(\hat{f}_{\min}) - \hat{R}_{\min}(f^*) \leq 0$  since  $\hat{f}_{\min} = \arg \min_{f \in \mathcal{F}} \hat{R}_{\min}(f)$ . As both the last two terms in the last line converges in the order of  $\mathcal{O}(1/\sqrt{n})$ , we can know that  $R_{\min}(\hat{f}_{\min}) - R_{\min}(f^*)$  converges in the order of  $\mathcal{O}(1/\sqrt{n})$ . The proof is completed.  $\square$

\*Corresponding author: Lei Feng <lfeng@cqu.edu.cn>.

### C Proof of Theorem 3

First, we prove that the optimal model  $f^*$  learned from fully labeled data (i.e.,  $f^* = \arg \min R(f)$ ) is also the optimal model for  $R_{\text{wet}}(f) = \mathbb{E}_{p(\mathbf{x}, S)}[\ell_{\text{wet}}(f(\mathbf{x}), S)]$  as follows.

By substituting the  $f^*$  into  $R_{\text{wet}}(f)$ , we obtain:

$$\begin{aligned}
R_{\text{wet}}(f^*) &= \mathbb{E}_{p(\mathbf{x}, S)}[\ell_{\text{wet}}(f^*(\mathbf{x}), S)] \\
&= \int_{\mathcal{X}} \int_S \ell_{\text{wet}}(f^*(\mathbf{x}), S) p(S | \mathbf{x}) p(\mathbf{x}) dS d\mathbf{x} \\
&= \int_{\mathcal{X}} \int_S \int_Y \ell_{\text{wet}}(f^*(\mathbf{x}), S) p(S, y | \mathbf{x}) p(\mathbf{x}) dy dS d\mathbf{x} \\
&= \int_{\mathcal{X}} \int_S \int_Y \sum_{y' \in S} w(\mathbf{x}, y') \ell(f^*(\mathbf{x}), y') \\
&\quad \cdot p(S, y | \mathbf{x}) p(\mathbf{x}) dy dS d\mathbf{x} \\
&= \int_{\mathcal{X}} \int_Y \ell(f^*(\mathbf{x}), y) \int_S p(S | y, \mathbf{x}) p(y | \mathbf{x}) p(\mathbf{x}) dS dy d\mathbf{x} \\
&= \int_{\mathcal{X}} \int_Y \ell(f^*(\mathbf{x}), y) p(\mathbf{x}, y) dy d\mathbf{x} \\
&= R(f^*) = 0,
\end{aligned}$$

where we used the equality  $\sum_{y' \in S} w(\mathbf{x}, y') \ell(f^*(\mathbf{x}), y') = \ell(f^*(\mathbf{x}), y)$ . This is because when the true label  $y \in S$  can make  $\ell(f^*(\mathbf{x}), y) = 0$ , and thus the weighting function would be  $w(\mathbf{x}, y) = 1$  and  $w(\mathbf{x}, y') = 0$  for  $y' \neq y$  since we try to minimize  $\ell_{\text{wet}}$ .

On the other hand, we prove that  $f^*$  is the sole optimal model for  $R_{\text{wet}}(f)$  by contradiction. Specifically, we assume that there is at least one other model  $g$  that makes  $R_{\text{wet}}(g) = 0$  and predicts a label  $y_g \neq y$  for at least one instance  $\mathbf{x}$ . Therefore, for any  $S$  containing  $y$  we have

$$\sum_{y' \in S} w(\mathbf{x}, y') \ell(g(\mathbf{x}), y') = \ell(g(\mathbf{x}), y_g) = 0.$$

The above equality implies that  $y_g$  is always included in the candidate label set of  $\mathbf{x}$  (co-occurring with the true label  $y$ ), and in this case, the ambiguity degree is 1. This contradicts the basic PLR assumption that the ambiguity degree should be less than 1. Therefore, there is one, and only one minimizer of  $R_{\text{wet}}$ , which is the same as the minimizer  $f^*$  learned from fully labeled data. The proof is completed.  $\square$

### D Proof of Theorem 4

Let us introduce the following notations:

$$\begin{aligned}
d &= \text{Pdim}(\{\mathbf{x} \mapsto \ell(f(\mathbf{x}), y) \mid f \in \mathcal{F}\}), \\
\tilde{d} &= \text{Pdim}(\{\mathbf{x} \mapsto \sum_{y \in S} w(\mathbf{x}, y) \ell(f(\mathbf{x}), y) \mid f \in \mathcal{F}\}),
\end{aligned}$$

where  $\text{Pdim}(\mathcal{F})$  denotes the pseudo-dimension of the functional space  $\mathcal{F}$  and  $w(\mathbf{x}, y)$  satisfy the basic conditions described in the main text. It is worth noting that we may represent  $\tilde{d}$  by  $d$  with some derivations, while for simplicity and convenience, we directly formulate the expression of  $\tilde{d}$ .

From the assumptions in Theorem 2, using the discussion in Theorem 10.6 of Mohri, Rostamizadeh, and Talwalkar (2012), with probability  $1 - \delta$  for all  $f \in \mathcal{F}$ ,

$$\begin{aligned}
\left| \mathbb{E}_{p(\mathbf{x}, S)}[\ell_{\text{wet}}(\mathbf{x}, S)] - \sum_{i=1}^n \ell_{\text{wet}}(\mathbf{x}_i, S_i) \right| \\
\leq M \sqrt{\frac{2\tilde{d} \log \frac{ne}{d}}{n}} + M \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.
\end{aligned}$$

Hence we can know that  $|R_{\text{wet}}(f) - \hat{R}_{\text{wet}}(f)|$  converges in the order of  $\mathcal{O}(1/\sqrt{n})$  for all  $f \in \mathcal{F}$ .

Then, we have the following derivations:

$$\begin{aligned}
R_{\text{wet}}(\hat{f}_{\text{wet}}) - R_{\text{wet}}(f^*) &\leq R_{\text{wet}}(\hat{f}_{\text{wet}}) - \hat{R}_{\text{wet}}(\hat{f}_{\text{wet}}) + \hat{R}_{\text{wet}}(\hat{f}_{\text{wet}}) - \hat{R}_{\text{wet}}(f^*) \\
&\quad + \hat{R}_{\text{wet}}(f^*) - R_{\text{wet}}(f^*) \\
&\leq R_{\text{wet}}(\hat{f}_{\text{wet}}) - \hat{R}_{\text{wet}}(\hat{f}_{\text{wet}}) + \hat{R}_{\text{wet}}(f^*) - R_{\text{wet}}(f^*) \\
&\leq |R_{\text{wet}}(\hat{f}_{\text{wet}}) - \hat{R}_{\text{wet}}(\hat{f}_{\text{wet}})| \\
&\quad + |\hat{R}_{\text{wet}}(f^*) - R_{\text{wet}}(f^*)|,
\end{aligned}$$

where the first inequality holds because  $\hat{R}_{\text{wet}}(\hat{f}_{\text{wet}}) - \hat{R}_{\text{wet}}(f^*) \leq 0$  since  $\hat{f}_{\text{wet}} = \arg \min_{f \in \mathcal{F}} \hat{R}_{\text{wet}}(f)$ . As both the last two terms in the last line converges in the order of  $\mathcal{O}(1/\sqrt{n})$ , we can know that  $R_{\text{wet}}(\hat{f}_{\text{wet}}) - R_{\text{wet}}(f^*)$  converges in the order of  $\mathcal{O}(1/\sqrt{n})$ . The proof is completed.  $\square$

### References

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A. 2012. *Foundations of Machine Learning*. MIT Press.