Learn the Highest Label and Rest Label Description Degrees – Supplementary Material

A Proof of Theorem 1

Proof. Since the label distribution function is assumed to be the conditional probability distribution function, define the Bayes classifier [Devroye *et al.*, 1996]

$$f^*(\boldsymbol{x}) = \arg\max_{\bar{y}\in\mathcal{Y}} d_{\boldsymbol{x}}^{\bar{y}}$$

expected 0/1 loss of which is the Bayes error, *i.e.*, $L_1^* = \mathbb{P}(f^*(\boldsymbol{x}) \neq y)$. Fix any $\boldsymbol{x}$, then we have

$$\begin{aligned}\mathbb{P}(f(\boldsymbol{x}) \neq y \mid \boldsymbol{x}) &= 1 - \mathbb{P}(f(\boldsymbol{x}) = y \mid \boldsymbol{x}) \\ &= 1 - \sum_{y_j : y_j = f(\boldsymbol{x})} \mathbb{P}(y = y_j | \boldsymbol{x}).\end{aligned}$$

Without loss of generality, let $f(\boldsymbol{x}) = y_l$ and $f^*(\boldsymbol{x}) = y_k$. Then, it follows that

$$\mathbb{P}(f(\boldsymbol{x}) \neq y \mid \boldsymbol{x}) - \mathbb{P}(f^*(\boldsymbol{x}) \neq y \mid \boldsymbol{x}) = d_{\boldsymbol{x}}^{y_k} - d_{\boldsymbol{x}}^{y_l}. \quad (1)$$

If $y_l = y_k$, then the right-hand side of Eq. (1) reduces to 0. If $y_l \neq y_k$, then the right-hand side of Eq. (1) is bounded by $\bar{d}_{\boldsymbol{x}}^{y_k} - \bar{d}_{\boldsymbol{x}}^{y_l}$ by the definition of the degenerated label distribution. Notice that $\bar{d}_{\boldsymbol{x}}^{y_k} \geq \bar{d}_{\boldsymbol{x}}^{y_l}$ and $h_{\boldsymbol{x}}^{y_k} \leq h_{\boldsymbol{x}}^{y_l}$ according to the definitions of f and f^*. Then by [Wang and Geng, 2019, Lemma 10], it follows that

$$\bar{d}_{\boldsymbol{x}}^{y_k} - \bar{d}_{\boldsymbol{x}}^{y_l} \leq |\bar{d}_{\boldsymbol{x}}^{y_k} - h_{\boldsymbol{x}}^{y_k}| + |\bar{d}_{\boldsymbol{x}}^{y_l} - h_{\boldsymbol{x}}^{y_l}|$$

which leads to

$$\mathbb{P}(f(\boldsymbol{x}) \neq y \mid \boldsymbol{x}) - \mathbb{P}(f^*(\boldsymbol{x}) \neq y \mid \boldsymbol{x}) \leq \sum_{j=1}^{m} |h_{\boldsymbol{x}}^{y_j} - \bar{d}_{\boldsymbol{x}}^{y_j}|.$$

Taking expectation on both sides of the preceding equation, we finish the proof. $\square$

B Proof of Theorem 2

Proof. Fix any $\boldsymbol{x}$. Without less of generality, let $f(\boldsymbol{x}) = y_l$ and $f'(\boldsymbol{x}) = y_k$. Then,

$$\mathbb{P}(f(\boldsymbol{x}) \neq y \mid \boldsymbol{x}) - \mathbb{P}(f'(\boldsymbol{x}) \neq y \mid \boldsymbol{x}) = d_{\boldsymbol{x}}^{y_k} - d_{\boldsymbol{x}}^{y_l}. \quad (2)$$

If $y_l = y_{\boldsymbol{x}}$, then the right-hand side of the above equation is less than 0. If $y_l = y_k$, then the right-hand side of the preceding equation reduces to 0. If $y_l \neq y_{\boldsymbol{x}}$ and $y_l \neq y_k$, then $d_{\boldsymbol{x}}^{y_k} \geq d_{\boldsymbol{x}}^{y_l}$ and $h_{\boldsymbol{x}}^{y_l} \geq h_{\boldsymbol{x}}^{y_k}$. By [Wang and Geng, 2019, Lemma 10], the right-hand side of the above equation is bounded by $|d_{\boldsymbol{x}}^{y_k} - h_{\boldsymbol{x}}^{y_k}| + |d_{\boldsymbol{x}}^{y_l} - h_{\boldsymbol{x}}^{y_l}|$, where $y_l \neq y_{\boldsymbol{x}}$ and $y_k \neq y_{\boldsymbol{x}}$. To summarize, we have

$$\mathbb{P}(f(\boldsymbol{x}) \neq y \mid \boldsymbol{x}) - \mathbb{P}(f'(\boldsymbol{x}) \neq y \mid \boldsymbol{x}) \leq \sum_{j : y_j \neq y_{\boldsymbol{x}}} |h_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_j}|.$$

Take expectation on both sides of the above equation, which completes the proof. $\square$

C Proof of Theorem 3

Proof. By Theorem 1, to bound $R(h)$, it's suffices to bound the expected L_1-norm loss. By a standard Rademacher bound [Mohri *et al.*, 2012], for any $\delta > 0$, with probability at least $1 - \delta$, the following bound holds for all $h \in \mathcal{H}$

$$\mathbb{E}\left[\sum_{j=1}^{m} |h_{\boldsymbol{x}}^{y_j} - \bar{d}_{\boldsymbol{x}}^{y_j}|\right] \leq \hat{R}(h) + 2\mathcal{R}(\ell_1 \circ \mathcal{H}) + \sqrt{\frac{\log 1/\delta}{2n}}, \quad (3)$$

where $\circ$ is the function combination operator, and ℓ_1 is the L_1-norm loss. Notice that $\ell_1 \circ \mathcal{H}$ can be re-written as $\{\ell_1 \circ \mathrm{SF}(f) : f \in \mathcal{F}\}$. Wang and Geng [2019] show that $\ell_1 \circ \mathrm{SF}$ satisfies $2m$-Lipschitz. Then, by [Maurer, 2016], we have

$$\mathcal{R}(\ell_1 \circ \mathcal{H}) \leq 2\sqrt{2}m \sum_{j=1}^{m} \mathcal{R}(\mathcal{F}_j), \quad (4)$$

where $\mathcal{F}_j$ is defined by $\mathcal{F}_j = \{\boldsymbol{x} \mapsto \boldsymbol{w}_j \cdot \boldsymbol{x} : \|\boldsymbol{w}_j\|_2 \leq \Lambda_2\}$. According to [Kakade *et al.*, 2009],

$$\mathcal{R}(\mathcal{F}_j) \leq \Lambda_2 \frac{\sup_{\boldsymbol{x}} \|\boldsymbol{x}\|}{\sqrt{n}} = \frac{\Lambda_1 \Lambda_2}{\sqrt{n}},$$

which leads to

$$\mathcal{R}(\ell_1 \circ \mathcal{H}) \leq 2\sqrt{2}m^2 \frac{\Lambda_1 \Lambda_2}{\sqrt{n}}.$$

Plug the above equation into Eq. (3), which yields that for any $\delta > 0$, with probability at least $1 - \delta$, the following bound holds for all $h \in \mathcal{H}$

$$\mathbb{E}\left[\sum_{j=1}^{m} |h_{\boldsymbol{x}}^{y_j} - \bar{d}_{\boldsymbol{x}}^{y_j}|\right] \leq \hat{R}(h) + \frac{4\sqrt{2}m^2 \Lambda_1 \Lambda_2}{\sqrt{n}} + \sqrt{\frac{\log 1/\delta}{2n}}. \quad (5)$$

Next, define the empirical estimation of the Bayes error by $L_1' = \frac{1}{n}\sum_{i=1}^{n}(1 - d_{\boldsymbol{x}_i}^{y_{\boldsymbol{x}_i}})$. By the Hoeffding's inequality, for any $\delta > 0$, with probability at least $1-\delta$ such that

$$|L_1' - L_1^*| \le \sqrt{\frac{\log 2/\delta}{2n}}. \tag{6}$$

Combine Eq. (5), Eq. (6) and Theorem 1, which completes the proof. □

D Proof of Theorem 4

Proof. To start, define the loss function ℓ_1' by

$$\ell_1'(\hat{D}, D) = \sum_{j:y_j \neq y_{\boldsymbol{x}}} |\hat{d}_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_j}|.$$

It's easy to see that ℓ_1' satisfies 1-Lipschitz since

$$\ell_1'(\hat{D}, D) - \ell_1'(\tilde{D}, D) \le \sum_{j:y_j \neq y_{\boldsymbol{x}}} |\hat{d}_{\boldsymbol{x}}^{y_j} - \tilde{d}_{\boldsymbol{x}}^{y_j}| \le \|\hat{D} - \tilde{D}\|_1.$$

Similar to the Proof of Theorem 3, with the 1-Lipschitz of ℓ_1', we can prove that for any $\delta > 0$, with probability at least $1-\delta$, the following bound holds for all $h \in \mathcal{H}$

$$\mathbb{E}\left[\sum_{j:y_j \neq y_{\boldsymbol{x}}} |h_{\boldsymbol{x}}^{y_j} - \bar{d}_{\boldsymbol{x}}^{y_j}|\right] \le \bar{R}(h) + \frac{4\sqrt{2}m^2\Lambda_1\Lambda_2}{\sqrt{n}} + \sqrt{\frac{\log 1/\delta}{2n}}. \tag{7}$$

Define the empirical estimation of L_2^* by $L_2' = \frac{1}{n}\sum_{i=1}^{n}(1 - d_{\boldsymbol{x}_i}^{y'_{\boldsymbol{x}_i}})$. By the Hoeffding's inequality, for any $\delta > 0$, with probability at least $1-\delta$ such that

$$|L_2' - L_2^*| \le \sqrt{\frac{\log 2/\delta}{2n}}. \tag{8}$$

Combine Eq. (7), Eq. (8) and Theorem 2, which completes the proof. □

References

[Devroye *et al.*, 1996] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Stochastic Modelling and Applied Probability*. Springer, 1996.

[Kakade *et al.*, 2009] Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems 21*, pages 793–800. December 2009.

[Maurer, 2016] Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *Proceedings of Algorithmic Learning Theory*, pages 3–17, 2016.

[Mohri *et al.*, 2012] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012.

[Wang and Geng, 2019] Jing Wang and Xin Geng. Theoretical analysis of label distribution learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5256–5263, January - February 2019.