Label Distribution Learning Machine – Supplementary Material

Before proving Theorems 1 and 2, we first introduce the following lemma proved in (Wang & Geng, 2019).

Lemma 1. *Let $c1, c2, c3$ and c_4 be real values satisfying $c_1 > c_2$ and $c_3 > c_4$. Then, $c_1 - c_2 < |c_1 - c_4| + |c_2 - c3|$.*

A. Proof of Theorem 1

Theorem 1. *For each $\boldsymbol{x} \in \mathcal{X}$, if the predicted label distribution satisfies the following inequality*

$$\sum_j |d_{\boldsymbol{x}}^{y_j} - \hat{d}_{\boldsymbol{x}}^{y_j}| \leq \alpha_{\boldsymbol{x}},$$

the predicted label satisfies $\hat{y}_{\boldsymbol{x}} = y_{\boldsymbol{x}}$.

Proof. We prove by contradiction. Suppose for the sake of contradiction that $\hat{y}_{\boldsymbol{x}} \neq y_{\boldsymbol{x}}$. Without loss of generality, let $y_{\boldsymbol{x}} = y_j$ and $\hat{y}_{\boldsymbol{x}} = y_i$ for $i \neq j$. Recall the definition of $y_{\boldsymbol{x}} = \arg\max_{\bar{y}} d_{\boldsymbol{x}}^{\bar{y}}$ and $\hat{y}_{\boldsymbol{x}} = \arg\max_{\bar{y}} \hat{d}_{\boldsymbol{x}}^{\bar{y}}$. Then, we have $d_{\boldsymbol{x}}^{y_j} > d_{\boldsymbol{x}}^{y_i}$ and $\hat{d}_{\boldsymbol{x}}^{y_i} > \hat{d}_{\boldsymbol{x}}^{y_j}$. By Lemma 1,

$$d_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_i} < |d_{\boldsymbol{x}}^{y_j} - \hat{d}_{\boldsymbol{x}}^{y_j}| + |d_{\boldsymbol{x}}^{y_i} - \hat{d}_{\boldsymbol{x}}^{y_i}|. \tag{1}$$

Further, observe that $\alpha_{\boldsymbol{x}} \leq d_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_i}$ and $|d_{\boldsymbol{x}}^{y_j} - \hat{d}_{\boldsymbol{x}}^{y_j}| + |d_{\boldsymbol{x}}^{y_i} - \hat{d}_{\boldsymbol{x}}^{y_i}| \leq \sum_l |d_{\boldsymbol{x}}^{y_l} - \hat{d}_{\boldsymbol{x}}^{y_l}|$, which yields

$$\alpha_{\boldsymbol{x}} < \sum_l |d_{\boldsymbol{x}}^{y_l} - \hat{d}_{\boldsymbol{x}}^{y_l}|.$$

The above equation contradicts. Thereby, we must have $y_{\boldsymbol{x}} = \hat{y}_{\boldsymbol{x}}$, which completes the proof. □

B. Proof of Theorem 2

Theorem 2. *For each $\boldsymbol{x} \in \mathcal{X}$, if the predicted label distribution satisfies the following inequality*

$$\sum_{j: y_j \neq y_{\boldsymbol{x}}} |d_{\boldsymbol{x}}^{y_j} - \hat{d}_{\boldsymbol{x}}^{y_j}| \leq \beta_{\boldsymbol{x}}, \tag{2}$$

the predicted label satisfies $\hat{y}_{\boldsymbol{x}} = y_{\boldsymbol{x}}$ or $\hat{y}_{\boldsymbol{x}} = y'_{\boldsymbol{x}}$.

Proof. The theorem holds if $\hat{y}_{\boldsymbol{x}} = y_{\boldsymbol{x}}$. Next, we will prove that $\hat{y}_{\boldsymbol{x}} = y'_{\boldsymbol{x}}$ if $\hat{y}_{\boldsymbol{x}} \neq y_{\boldsymbol{x}}$.

We prove by contradiction. Suppose for the sake of contradiction that $\hat{y}_{\boldsymbol{x}} \neq y'_{\boldsymbol{x}}$. Without loss of generality, let $\hat{y}_{\boldsymbol{x}} = y_i \neq y_{\boldsymbol{x}}$ and $y'_{\boldsymbol{x}} = y_j$. If $y_i \neq y_j$. By the definition of $\hat{y}_{\boldsymbol{x}}$, we have $\hat{d}_{\boldsymbol{x}}^{y_i} > \hat{d}_{\boldsymbol{x}}^{y_j}$. Recall $y'_{\boldsymbol{x}} = \arg\max_{\bar{y} \neq y_{\boldsymbol{x}}} d_{\boldsymbol{x}}^{\bar{y}}$. Then, we have $d_{\boldsymbol{x}}^{y_j} > d_{\boldsymbol{x}}^{y_i}$ because $y_i \neq y_{\boldsymbol{x}}$. By Lemma 1,

$$d_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_i} < |d_{\boldsymbol{x}}^{y_j} - \hat{d}_{\boldsymbol{x}}^{y_j}| + |d_{\boldsymbol{x}}^{y_i} - \hat{d}_{\boldsymbol{x}}^{y_i}|. \tag{3}$$

If $y_i = y_j$, the above inequality still holds. Notice that $y_j \neq y_{\boldsymbol{x}}$ and $y_i \neq y_{\boldsymbol{x}}$, which leads to $\beta_{\boldsymbol{x}} \leq d_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_i}$ and $|d_{\boldsymbol{x}}^{y_j} - \hat{d}_{\boldsymbol{x}}^{y_j}| + |d_{\boldsymbol{x}}^{y_i} - \hat{d}_{\boldsymbol{x}}^{y_i}| \leq \sum_{l: y_l \neq y_{\boldsymbol{x}}} |d_{\boldsymbol{x}}^{y_l} - \hat{d}_{\boldsymbol{x}}^{y_l}|$. Thereby,

$$\beta_{\boldsymbol{x}} < \sum_{l: y_l \neq y_{\boldsymbol{x}}} |d_{\boldsymbol{x}}^{y_l} - \hat{d}_{\boldsymbol{x}}^{y_l}|,$$

which contradicts. Hence, we must $\hat{y}_{\boldsymbol{x}} = y'_{\boldsymbol{x}}$, which completes the proof. □

C. Proof of Theorem 3

Theorem 3. *Let $\mathcal{F} = \{\boldsymbol{x} \mapsto \boldsymbol{W}^\top \cdot \boldsymbol{x} : \|\boldsymbol{w}_j\|_2 \leq \Lambda\}$ be the hypothesis space. Fix $1 > \rho > 0$. For any $\delta > 0$, with probability at least $1 - \delta$, the bounds hold for all $f \in \mathcal{F}$,*

$$R(f) \leq \hat{R}_\rho(f) + \frac{2\sqrt{2} r \Lambda m}{(1-\rho)\sqrt{n}} + \sqrt{\frac{\log {}^{1}/_{\delta}}{2n}},$$

$$R(f) \leq \min\left\{\hat{R}_\rho(f) + \frac{2\sqrt{2} r \Lambda m}{(1-\rho)\sqrt{n}}, \tilde{R}_\rho(f) + \frac{4 r \Lambda m}{\rho\sqrt{n}}\right\} + \sqrt{\frac{\log {}^{2}/_{\delta}}{2n}}.$$

Before presenting the proof, we introduce the following definition.

Definition. *For any $\rho < 1$, define the ρ-margin loss Φ_ρ*

$$\Phi_\rho(x) = \begin{cases} 0 & \text{if } x \leq \rho \\ \frac{x-\rho}{1-\rho} & \text{if } \rho < x \leq 1 \\ 1 & \text{otherwise.} \end{cases}$$

Fig. 1 shows the ρ-insensitive loss and the ρ-margin loss. It's trivial hat Φ_ρ satisfies $1/(1-\rho)$-Lipschitzness.

Proof. Recall $L = \{l_{\boldsymbol{x}}^{y_1}, \cdots, l_{\boldsymbol{x}}^{y_m}\}$, where $l_{\boldsymbol{x}}^{y_j}$ equals 1 if $y_j = y_{\boldsymbol{x}}$ and 0 otherwise. Let $\mathcal{H} = \{z = (\boldsymbol{x}, y_{\boldsymbol{x}}) \mapsto \sum_j |f_j(\boldsymbol{x}) - l_{\boldsymbol{x}}^{y_j}| : f \in \mathcal{F}\}$. Consider the family of functions taking values in [0, 1]

$$\tilde{\mathcal{H}} = \{\Phi_\rho \circ h : h \in \mathcal{H}\}.$$

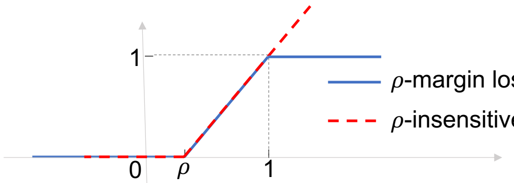

Figure 1. Illustration of the ρ-insensitive loss and ρ-margin loss.

Applying a standard Rademacher bound (Mohri et al., 2018) to $\tilde{\mathcal{H}}$, for any $\delta > 0$, with probability at least $1-\delta$, the following bound holds for all $g \in \tilde{\mathcal{H}}$,

$$\mathbb{E}\left[g(z)\right] \leq \frac{1}{n}\sum_{i=1}^{n} g(z_i) + 2\mathcal{R}_n(\tilde{\mathcal{H}}) + \sqrt{\frac{\log 1/\delta}{2n}},$$

and the following bound holds for all $f \in \mathcal{F}$

$$\mathbb{E}\left[\Phi_\rho(\|f(\boldsymbol{x}) - L\|_1)\right] \leq \hat{R}_\rho(f) + 2\mathcal{R}_n(\Phi_\rho \circ \mathcal{H}) + \sqrt{\frac{\log 1/\delta}{2n}}.$$

By Corollary 1, $\mathbb{E}\left[\Phi_\rho(\|f(\boldsymbol{x}) - L\|_1)\right] \geq \mathbb{I}(\hat{y}_{\boldsymbol{x}} \neq y_{\boldsymbol{x}}) = 0$ if $\|f(\boldsymbol{x}) - L\|_1 \leq 1$. Moreover, $\mathbb{E}\left[\Phi_\rho(\|f(\boldsymbol{x}) - L\|_1)\right] = 1$ if $\|f(\boldsymbol{x}) - L\|_1 \geq 1$. Hence, $R(f) \leq \mathbb{E}\left[\Phi_\rho(\|f(\boldsymbol{x}) - L\|_1)\right]$, which leads to

$$R(f) \leq \hat{R}_\rho(f) + 2\mathcal{R}_n(\Phi_\rho \circ \mathcal{H}) + \sqrt{\frac{\log 1/\delta}{2n}}.$$

By the $1/(1-\rho)$-Lipschitzness of Φ_ρ, we have

$$\mathcal{R}_n(\Phi_\rho \circ \mathcal{H}) \leq \frac{1}{1-\rho}\mathcal{R}_n(\mathcal{H}) \leq \frac{\sqrt{2}}{1-\rho}\sum_{j=1}^{m}\mathcal{R}_n(\mathcal{F}_j),$$

where the second inequality is according to (Maurer, 2016), and $\mathcal{F}_j = \{\boldsymbol{x} \mapsto \boldsymbol{w}_j \cdot \boldsymbol{x} : \|\boldsymbol{w}_j\|_2 \leq \Lambda\}$. According to (Mohri et al., 2018), $\mathcal{R}_n(\mathcal{F}_j) \leq \Lambda r/\sqrt{n}$, which yields

$$\mathcal{R}_n(\Phi_\rho \circ \mathcal{H}) \leq \frac{\sqrt{2}m\Lambda r}{(1-\rho)\sqrt{n}}.$$

Thus, we have the following bound

$$R(f) \leq \hat{R}_\rho(f) + \frac{2\sqrt{2}m\Lambda r}{(1-\rho)\sqrt{n}} + \sqrt{\frac{\log 1/\delta}{2n}}, \qquad (4)$$

which completes the proof for the first part.

Next, we prove the second part. The first part can be equivalently re-written as, for any $\delta > 0$, with probability at least $1-\delta/2$, the following bound holds for all $f \in \mathcal{F}$,

$$R(f) \leq \hat{R}_\rho(f) + \frac{2\sqrt{2}m\Lambda r}{(1-\rho)\sqrt{n}} + \sqrt{\frac{\log 2/\delta}{2n}}. \qquad (5)$$

Besides, Mohri et al. (2018) showed that for a multi-class SVM, the generalization bound is as follows: for any $\delta > 0$, with probability at least $1-\delta/2$, the following bound holds for all $f \in \mathcal{F}$,

$$R(f) < \tilde{R}_\rho(f) + \frac{4m\Lambda r}{\rho\sqrt{n}} + \sqrt{\frac{\log 2/\delta}{2n}}. \qquad (6)$$

Combine Eqs. (5) and (6), which completes the proof for the second part. □

D. Proof of Theorem 5

Theorem 5.. *Let $\hat{d}$ be a learned LDL function. Let $\mathcal{N}$ and $\mathcal{M}$ be defined above. Then, the following bound holds*

$$\mathbb{P}(\hat{y}_{\boldsymbol{x}} \neq y) - L_1^* \leq \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{N}\cap\mathcal{M}}}\left[\sum_{\bar{y}} |\hat{d}_{\boldsymbol{x}}^{\bar{y}} - d_{\boldsymbol{x}}^{\bar{y}}|\right].$$

Before proving the theorem, we introduce the following lemma.

Lemma 2. *Fix an $\boldsymbol{x}$. Then,*

$$\mathbb{P}_y[\hat{y}_{\boldsymbol{x}} \neq y \mid \boldsymbol{x}] - \mathbb{P}_y[y_{\boldsymbol{x}} \neq y \mid \boldsymbol{x}] = d_{\boldsymbol{x}}^{y_{\boldsymbol{x}}} - d_{\boldsymbol{x}}^{\hat{y}_{\boldsymbol{x}}}.$$

Proof of Lemma 2. First, we have

$$\mathbb{P}_y[\hat{y}_{\boldsymbol{x}} \neq y \mid \boldsymbol{x}] = 1 - \mathbb{P}_y[y = \hat{y}_{\boldsymbol{x}} \mid \boldsymbol{x}] = 1 - d_{\boldsymbol{x}}^{\hat{y}_{\boldsymbol{x}}},$$

and

$$\mathbb{P}_y[y_{\boldsymbol{x}} \neq y \mid \boldsymbol{x}] = 1 - \mathbb{P}_y[y = y_{\boldsymbol{x}} \mid \boldsymbol{x}] = 1 - d_{\boldsymbol{x}}^{y_{\boldsymbol{x}}},$$

which yields

$$\mathbb{P}_y[\hat{y}_{\boldsymbol{x}} \neq y \mid \boldsymbol{x}] - \mathbb{P}_y[y_{\boldsymbol{x}} \neq y \mid \boldsymbol{x}] = d_{\boldsymbol{x}}^{y_{\boldsymbol{x}}} - d_{\boldsymbol{x}}^{\hat{y}_{\boldsymbol{x}}}.$$

□

Proof of Theorem 5. First, notice that

$$\begin{aligned}&\mathbb{P}(\hat{y}_{\boldsymbol{x}} \neq y) - L_1^* \\ &= \mathbb{E}_{y,\boldsymbol{x}\sim\mathcal{D}_{\mathcal{N}\cap\mathcal{M}}}\left[\mathbb{I}(\hat{y}_{\boldsymbol{x}} \neq y) - \mathbb{I}(y_{\boldsymbol{x}} \neq y)\right] \\ &+ \mathbb{E}_{y,\boldsymbol{x}\sim\mathcal{D}_{\bar{\mathcal{N}}\cup\bar{\mathcal{M}}}}\left[\mathbb{I}(\hat{y}_{\boldsymbol{x}} \neq y) - \mathbb{I}(y_{\boldsymbol{x}} \neq y)\right],\end{aligned} \qquad (7)$$

where $\bar{\mathcal{N}} = \mathcal{X} \setminus \mathcal{N}$ is the complementary set of $\mathcal{N}$. By the definitions of $\mathcal{N}$ and $\mathcal{M}$, for any $\boldsymbol{x} \in \bar{\mathcal{N}} \cup \bar{\mathcal{M}}$, $\hat{y}_{\boldsymbol{x}} = y_{\boldsymbol{x}}$. According to Lemma 2, the second item on the right-hand side of Eq. (7) reduces to 0. Similarly, according to Lemma 2, the first item on the right-hand side of Eq. (7) equals

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}\cap\mathcal{M}}\left[d_{\boldsymbol{x}}^{y_{\boldsymbol{x}}} - d_{\boldsymbol{x}}^{\hat{y}_{\boldsymbol{x}}}\right].$$

If $y_{\boldsymbol{x}} \neq \hat{y}_{\boldsymbol{x}}$, according to Eq. (1), it follows that

$$d_{\boldsymbol{x}}^{y_{\boldsymbol{x}}} - d_{\boldsymbol{x}}^{\hat{y}_{\boldsymbol{x}}} \leq \sum_j |\hat{d}_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_j}|.$$

If $y_{\boldsymbol{x}} = \hat{y}_{\boldsymbol{x}}$, the above inequality still holds. Thereby,

$$\mathbb{P}(\hat{y}_{\boldsymbol{x}} \neq y) - L_1^* \leq \mathbb{E}_{\boldsymbol{x}\sim\mathcal{N}\cap\mathcal{M}}\left[d_{\boldsymbol{x}}^{y_{\boldsymbol{x}}} - d_{\boldsymbol{x}}^{\hat{y}_{\boldsymbol{x}}}\right] \leq \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{\mathcal{N}\cap\mathcal{M}}}\left[\sum_j |\hat{d}_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_j}|\right],$$

which completes the proof. □

E. Proof of Theorem 6

Theorem 6.. *Let $\mathcal{F}$ be the hypothesis space defined in Theorem 3. Fix $1 > \rho > 0$ and $\beta \geq 0$ such that $\beta \leq \beta_{\boldsymbol{x}}$ for all $\boldsymbol{x} \in \mathcal{X}$. Then, for any $\delta > 0$, with probability at least $1-\delta$, the following bound holds for all $f \in \mathcal{F}$*

$$\mathbb{P}(\hat{y}_{\boldsymbol{x}} \neq y) \leq \min\left\{L_1^* + \hat{R}_\rho(f) + \frac{2\sqrt{2}r\Lambda m}{(1-\rho)\sqrt{n}}, L_2^* + \hat{R}_\beta(f) + \frac{2\sqrt{2}m\Lambda r}{\sqrt{n}}\right\} + \sqrt{\frac{\log 2/\delta}{2n}}.$$

To prove Theorem 6, we first establish following lemmas.

Lemma 3. *Let $\mathcal{F}$ be the hypothesis space defined in Theorem 3. Fix $1 > \rho > 0$. Then, for any $\delta > 0$, with probability at least $1-\delta$, the following bounds for all $f \in \mathcal{F}$*

$$\mathbb{P}(\hat{y}_{\boldsymbol{x}} \neq y) - L_1^* \leq \hat{R}_\rho(f) + \frac{2\sqrt{2}r\Lambda m}{(1-\rho)\sqrt{n}} + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

Proof of Lemma 3. Fix an $\boldsymbol{x}$. If $\|f(\boldsymbol{x}) - L\|_1 \leq 1$, $\hat{y}_{\boldsymbol{x}} = y_{\boldsymbol{x}}$, which implies that $\mathbb{P}_y[\hat{y}_{\boldsymbol{x}} \neq y \mid \boldsymbol{x}] - \mathbb{P}_y[y_{\boldsymbol{x}} \neq y \mid \boldsymbol{x}] = 0$. Besides, $\mathbb{P}_y[\hat{y}_{\boldsymbol{x}} \neq y \mid \boldsymbol{x}] - \mathbb{P}_y[y_{\boldsymbol{x}} \neq y \mid \boldsymbol{x}] \leq 1$. By the definition of Φ_ρ, $\Phi_\rho(\|f(\boldsymbol{x}) - L\|_1)$ is larger than or equal to 0 if $\|f(\boldsymbol{x}) - L\|_1 \leq 1$ and is larger than 1 otherwise. Thereby, we have

$$\mathbb{P}_y[\hat{y}_{\boldsymbol{x}} \neq y \mid \boldsymbol{x}] - \mathbb{P}_y[y_{\boldsymbol{x}} \neq y \mid \boldsymbol{x}] \leq \Phi_\rho(\|f(\boldsymbol{x}) - L\|_1).$$

Take expectation on both sides of the above inequality,

$$\mathbb{P}(\hat{y}_{\boldsymbol{x}} \neq y) - L_1^* \leq \mathbb{E}\left[\Phi_\rho(\|f(\boldsymbol{x}) - L\|_1)\right]. \tag{8}$$

According to proof of Theorem 3, the right-hand side of above inequality is bounded by

$$\mathbb{E}\left[\Phi_\rho(\|f(\boldsymbol{x}) - L\|_1)\right] \leq \hat{R}_\rho(f) + \frac{2\sqrt{2}m\Lambda r}{(1-\rho)\sqrt{n}} + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

Combine the above inequality and Eq. (8), which completes the proof. □

Lemma 4. *Let β be defined in Theorem 6. Let $\hat{d}$ be a learned LDL function. Then, the following bound holds*

$$\mathbb{E}_{y,\boldsymbol{x}}\left[\mathbb{I}(\hat{y}_{\boldsymbol{x}} \neq y)\right] - L_2^* \leq \mathbb{E}\left[\ell_\beta(\sum_{j:y_j\neq y_{\boldsymbol{x}}} |\hat{d}_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_j}|) + \beta\right].$$

Proof of Lemma 4. Fix an $\boldsymbol{x}$. By Lemma 2, we have

$$\mathbb{P}_y[\hat{y}_{\boldsymbol{x}} \neq y \mid \boldsymbol{x}] - \mathbb{P}_y[y'_{\boldsymbol{x}} \neq y \mid \boldsymbol{x}] = d_{\boldsymbol{x}}^{y'_{\boldsymbol{x}}} - d_{\boldsymbol{x}}^{\hat{y}_{\boldsymbol{x}}}.$$

If $\hat{y}_{\boldsymbol{x}} \neq y_{\boldsymbol{x}}$, by Eq. (3), it follows that

$$d_{\boldsymbol{x}}^{y'_{\boldsymbol{x}}} - d_{\boldsymbol{x}}^{\hat{y}_{\boldsymbol{x}}} \leq \sum_{j:y_j\neq y_{\boldsymbol{x}}} |\hat{d}_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_j}|.$$

If $\hat{y}_{\boldsymbol{x}} = y_{\boldsymbol{x}}$, the above inequality still holds. Thereby,

$$\mathbb{P}_y[\hat{y}_{\boldsymbol{x}} \neq y\boldsymbol{x}] - \mathbb{P}_y[y'_{\boldsymbol{x}} \neq y \mid \boldsymbol{x}] \leq \sum_{j:y_j\neq y_{\boldsymbol{x}}} |\hat{d}_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_j}|.$$

Recall the definition of ℓ_β, we have

$$\mathbb{P}_y[\hat{y}_{\boldsymbol{x}} \neq y|\boldsymbol{x}] - \mathbb{P}_y[y'_{\boldsymbol{x}} \neq y|\boldsymbol{x}] \leq \ell_\beta(\sum_{j:y_j\neq y_{\boldsymbol{x}}} |\hat{d}_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_j}|) + \beta.$$

Taking expectation on both sides of the above equation, we completes the proof. □

Lemma 5. *Let $\mathcal{F}$ be the hypothesis space defined in Theorem 3. Fix $\beta > 0$ as Theorem 6 does. Then, for any $\delta > 0$, with probability at least $1-\delta$, the bounds for all $f \in \mathcal{F}$*

$$\mathbb{E}\left[\mathbb{I}(\hat{y}_{\boldsymbol{x}} \neq y)\right] - L_2^* \leq (\hat{R}_\beta(f) + \beta) + \frac{2\sqrt{2}m\Lambda r}{\sqrt{n}} + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

Proof of Lemma 5. To start, define

$$\ell'_\beta(x) = \min\{1, \ell_\beta(x) + \beta\}$$

According to Lemma 4, it's trivial to see that

$$\mathbb{E}_{y,\boldsymbol{x}}\left[\mathbb{I}(\hat{y}_{\boldsymbol{x}} \neq y)\right] - L_2^* \leq \mathbb{E}\left[\ell'_\beta(\sum_{j:y_j\neq y_{\boldsymbol{x}}} |\hat{d}_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_j}|)\right],$$

because $\mathbb{E}_{y,\boldsymbol{x}}\left[\mathbb{I}(\hat{y}_{\boldsymbol{x}} \neq y)\right] - L_2^* \leq 1$. It suffices to bound the right-hand side of the above equation.

Define $\mathcal{H} = \{z = (\boldsymbol{x}, D) \mapsto \sum_{j:y_j\neq y_{\boldsymbol{x}}} |f_j(\boldsymbol{x}) - d_{\boldsymbol{x}}^{y_j}| : f \in \mathcal{F}\}$. Applying a standard Rademacher bound (Mohri et al., 2018) to $\ell'_\beta \circ \mathcal{H}$, for any $\delta > 0$, with probability at $1-\delta$, the following bound holds for all $f \in \mathcal{F}$

$$\mathbb{E}\left[\ell'_\beta(\sum_{j:y_j\neq y_{\boldsymbol{x}}} |\hat{d}_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_j}|)\right] \leq \hat{R}_\beta(f) + 2\mathcal{R}_n(\ell'_\beta \circ \mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

By the 1-Lipschitzness of ℓ'_β, it follows that

$$\mathcal{R}_n(\ell'_\beta \circ \mathcal{H}) \leq \mathcal{R}_n(\mathcal{H}).$$

Define $\ell(D, \hat{D}) = \sum_{j: y_j \neq y_{\boldsymbol{x}}} |\hat{d}_{\boldsymbol{x}}^{y_j} - d_{\boldsymbol{x}}^{y_j}|$. Then, $\mathcal{H}$ can be equivalently re-written as $\ell \circ \mathcal{F}$. Notice that ℓ satisfies 1-Lipschitzness since

$$\ell(D, \hat{D}) - \ell(D, \bar{D}) \leq \|\hat{D} - \bar{D}\|_1.$$

Similar to the proof of Theorem 3, we have

$$\mathcal{R}_n(\mathcal{H}) \leq \frac{\sqrt{2} m \Lambda r}{\sqrt{n}},$$

which leads to

$$\mathbb{E}_{y,\boldsymbol{x}} \left[\mathbb{I}(\hat{y}_{\boldsymbol{x}} \neq y) \right] - L_2^* \leq \hat{R}_\beta(f) + \frac{2\sqrt{2} m \Lambda r}{\sqrt{n}} + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.$$

□

Proof of Theorem 6. The proof of Theorem 6 comes naturally by combining Lemmas 3 and 5. □

References

Maurer, A. A vector-contraction inequality for rademacher complexities. In *Proceedings of International Conference on Algorithmic Learning Theory*, pp. 3–17, 2016.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2 edition, 2018. ISBN 978-0-262-03940-6.

Wang, J. and Geng, X. Theoretical analysis of label distribution learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 5256–5263, 2019.