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

**Lemma 1.** Let \( c_1, c_2, c_3 \) and \( c_4 \) be real values satisfying \( c_1 > c_2 \) and \( c_3 > c_4 \). Then, \( c_1 - c_2 < |c_1 - c_4| + |c_2 - c_3| \).

**A. Proof of Theorem 1**

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

\[
\sum_j |d^{y_j}_x - \hat{d}^{y_j}_x| \leq \alpha_x,
\]

the predicted label satisfies \( \hat{y}_x = y_x \).

**Proof.** We prove by contradiction. Suppose for the sake of contradiction that \( \hat{y}_x \neq y_x \). Without loss of generality, let \( y_x = y_j \) and \( \hat{y}_x = y_i \) for \( i \neq j \). Recall the definition of \( y_x = \arg \max_{y} d^{y}_x \) and \( \hat{y}_x = \arg \max_{y} \hat{d}^{y}_x \). Then, we have \( d^{y_j}_x > d^{y_i}_x \). By Lemma 1,

\[
d^{y_j}_x - d^{y_i}_x < |d^{y_j}_x - \hat{d}^{y_j}_x| + |d^{y_i}_x - \hat{d}^{y_i}_x|.
\]

Further, observe that \( \alpha_x \leq d^{y_j}_x - d^{y_i}_x \) and \( |d^{y_j}_x - \hat{d}^{y_j}_x| + |d^{y_i}_x - \hat{d}^{y_i}_x| \leq \sum_i |d^{y_i}_x - \hat{d}^{y_i}_x| \), which yields

\[
\alpha_x < \sum_i |d^{y_i}_x - \hat{d}^{y_i}_x|.
\]

The above equation contradicts. Thereby, we must have \( y_x = y_i \), which completes the proof. \( \square \)

**B. Proof of Theorem 2**

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

\[
\sum_{j \neq y_x} |d^{y_j}_x - \hat{d}^{y_j}_x| \leq \beta_x,
\]

the predicted label satisfies \( \hat{y}_x = y_x \) or \( \hat{y}_x = y'_x \).

**Proof.** The theorem holds if \( \hat{y}_x = y_x \). Next, we will prove that \( \hat{y}_x = y'_x \) if \( \hat{y}_x \neq y_x \).

We prove by contradiction. Suppose for the sake of contradiction that \( \hat{y}_x \neq y'_x \). Without loss of generality, let \( \hat{y}_x = y_i \neq y_x \) and \( y'_x = y_j \). If \( y_i \neq y_j \), By the definition of \( \hat{y}_x \), we have \( d^{y_i}_x > d^{y_j}_x \). Recall \( \hat{y}_x = \arg \max_{y \neq y_x} d^{y}_x \). Then, we have \( d^{y_i}_x > d^{y_x}_x \) because \( y_i \neq y_x \). By Lemma 1,

\[
d^{y_i}_x - d^{y_x}_x < |d^{y_i}_x - \hat{d}^{y_i}_x| + |d^{y_x}_x - \hat{d}^{y_x}_x|.
\]

If \( y_i = y_j \), the above inequality still holds. Notice that \( y_j \neq y_x \) and \( y_i \neq y_x \), which leads to \( \beta_x \leq d^{y_i}_x - d^{y_x}_x \) and \( |d^{y_i}_x - \hat{d}^{y_i}_x| + |d^{y_x}_x - \hat{d}^{y_x}_x| \leq \sum_i |d^{y_i}_x - \hat{d}^{y_i}_x| - d^{y_x}_x - \hat{d}^{y_x}_x| \). Thereby,

\[
\beta_x < \sum_i |d^{y_i}_x - \hat{d}^{y_i}_x|,
\]

which contradicts. Hence, we must \( y_x = y'_x \), which completes the proof. \( \square \)

**C. Proof of Theorem 3**

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

\[
R(f) \leq \hat{R}_\rho(f) + \frac{2\sqrt{2}\rho \Lambda m}{(1 - \rho)\sqrt{n}} + \frac{\sqrt{\log \frac{1}{\delta}}}{2n},
\]

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

Before presenting the proof, we introduce the following definition.

**Definition.** For any \( \rho < 1 \), define the \( \rho \)-margin loss \( \Phi_\rho \)

\[
\Phi_\rho(x) = \begin{cases} 0 & \text{if } x \leq \rho \\ \frac{x - \rho}{\rho - \rho^2} & \text{if } \rho < x \leq 1 \\ 1 & \text{otherwise.} \end{cases}
\]

Fig. 1 shows the \( \rho \)-insensitive loss and the \( \rho \)-margin loss. It’s trivial that \( \Phi_\rho \) satisfies \( 1/(1 - \rho) \)-Lipschitzness.

**Proof.** Recall \( L = \{l^{y_j}_{x_i}, \ldots, l^{y_m}_{x_i}\} \), where \( l^{y_j}_{x_i} \) equals 1 if \( y_j = \hat{y}_x \) and 0 otherwise. Let \( \mathcal{H} = \{ z = (x, y_x) \mapsto \sum_j |f_j(x) - l^{y_j}_{x_i}| : f \in \mathcal{F} \} \). Consider the family of functions taking values in \([0, 1] \)

\[
\hat{H} = \{ \Phi_\rho \circ h : h \in \mathcal{H} \}.
\]
Applying a standard Rademacher bound \cite{Mohri2018} to $\mathcal{H}$, for any $\delta > 0$, with probability at least $1 - \delta$, the following bound holds for all $g \in \mathcal{H}$,

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

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

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

By Corollary 1, $E[\Phi_\rho(\|f(x) - L\|_1)] \geq 0$ if $\|f(x) - L\|_1 \leq 1$. Moreover, $E[\Phi_\rho(\|f(x) - L\|_1)] = 1$ if $\|f(x) - L\|_1 \geq 1$. Hence, $R(f) \leq E[\Phi_\rho(\|f(x) - L\|_1)]$, which leads to

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

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

$$R_n(\Phi_\rho \circ \mathcal{H}) \leq \frac{1}{1 - \rho} R_n(\mathcal{H}) \leq \frac{\sqrt{2}}{1 - \rho} \sum_{j=1}^{m} R_n(f_j),$$ 

where the second inequality is according to \cite{Maurer2016}, and $f_j = \{x \mapsto w_j \cdot x : \|w_j\|_2 \leq \Lambda\}$. According to \cite{Mohri2018}, $R_n(f_j) \leq \Lambda r/\sqrt{n}$, which yields

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

Thus, we have the following bound

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

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 $f \in \mathcal{F}$,

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

Besides, \cite{Mohri2018} 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 $f \in \mathcal{F}$,

$$R(f) < \hat{R}_n(f) + \frac{4m\Lambda r}{\rho\sqrt{n}} + \sqrt{\frac{\log 2/\delta}{2n}}.$$ 

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

\section{D. Proof of Theorem 5}

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

$$P(\hat{y}_x \neq y) - L_1^* \leq E_{x \sim D_{\mathcal{N} \cap \mathcal{M}}} \left[ \sum_y |\tilde{d}_{xy}^\rho - \hat{d}_{xy}^\rho| \right].$$

Before proving the theorem, we introduce the following lemma.

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

$$P_y[\hat{y}_x \neq y \mid x] - P_y[y = \hat{y}_x \mid x] = d_{xy}^u - d_{xy}^\rho.$$ 

\textbf{Proof of Lemma 2.} First, we have

$$P_y[\hat{y}_x \neq y \mid x] = 1 - P_y[y = \hat{y}_x \mid x] = 1 - d_{xy}^u,$$

and

$$P_y[y = \hat{y}_x \mid x] = 1 - P_y[y = y_x \mid x] = 1 - d_{xy}^\rho,$$

which yields

$$P_y[\hat{y}_x \neq y \mid x] - P_y[y_x \neq y \mid x] = d_{xy}^u - d_{xy}^\rho.$$ 

\textbf{Proof of Theorem 5.} First, notice that

$$P(\hat{y}_x \neq y) - L_1^* = E_{y,x \sim D_{\mathcal{N} \cap \mathcal{M}}} \left[ I[y_x \neq \hat{y}_x] - I[y_x = y_x] \right]$$

(7)

$$+ E_{y,x \sim D_{\mathcal{N} \cap \mathcal{M}}} \left[ I[y_x \neq y_x] - I[y_x = y_x] \right],$$

where $\mathcal{N} = \mathcal{X} \setminus \mathcal{N}$ is the complementary set of $\mathcal{N}$. By the definitions of $\mathcal{N}$ and $\mathcal{M}$, for any $x \in \mathcal{N} \cup \mathcal{M}$, $y_x = y_x$. 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

$$E_{x \sim D_{\mathcal{N} \cap \mathcal{M}}} \left[ d_{xy}^u - d_{xy}^\rho \right].$$

If $y_x \neq \hat{y}_x$, according to Eq. (1), it follows that

$$d_{xy}^u - d_{xy}^\rho \leq \sum_j |d_{xy}^u - d_{xy}^l|.$$
If \( y_{x} = \hat{y}_{x} \), the above inequality still holds. Thereby,
\[
P(\hat{y}_{x} \neq y) - L_{1}^{*} \leq \mathbb{E}_{x \sim \mathcal{N} \cap \mathcal{M}} \left[ d_{y\hat{x}} - d_{y} \right] \\
\leq \mathbb{E}_{x \sim \mathcal{D} \cap \mathcal{M}} \left[ \sum_{j} |d_{y\hat{x}} - d_{y}^{j}| \right],
\]
which completes the proof.

\[\square\]

### 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 \( \beta \leq \beta_{x} \) for all \( x \in \mathcal{X} \). Then, for any \( \delta > 0 \), with probability at least \( 1 - \delta \), the following bound holds for all \( f \in \mathcal{F} \)
\[
P(\hat{y}_{x} \neq y) \leq \min \left\{ L_{1}^{*} + \hat{R}_{f}(f) + \frac{2 \sqrt{2} r \Lambda_{m}}{(1 - \rho) \sqrt{n}}, L_{2}^{*} + \hat{R}_{f}(f) + \frac{2 \sqrt{2} m \Lambda_{r}}{\sqrt{n}} + \sqrt{\frac{\log \frac{1}{\delta}}{2n}} \right\}.
\]

To prove Theorem 6, we first establish the 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 bound holds for all \( f \in \mathcal{F} \)
\[
P(y_{x} \neq y) - L_{1}^{*} \leq \hat{R}_{f}(f) + \frac{2 \sqrt{2} r \Lambda_{m}}{(1 - \rho) \sqrt{n}} + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.
\]

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

Take expectation on both sides of the above inequality,
\[
P(\hat{y}_{x} \neq y) - L_{1}^{*} \leq \mathbb{E}[\Phi_{p}(\| f(x) - L \|_{1})].
\]

According to proof of Theorem 3, the right-hand side of above inequality is bounded by
\[
\mathbb{E}[\Phi_{p}(\| f(x) - L \|_{1})] \leq \hat{R}_{f}(f) + \frac{2 \sqrt{2} m \Lambda_{r}}{\sqrt{n}} + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.
\]

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

\[\square\]

**Lemma 4.** Let \( \beta \) be defined in Theorem 6. Let \( \hat{\beta} \) be a learned LDL function. Then, the following bound holds
\[
\mathbb{E}_{y,x}[\| \hat{y}_{x} \neq y \|] - L_{2}^{*} \leq \mathbb{E} \left[ \ell_{\beta}(\sum_{j:y_{j} \neq y_{x}}|d_{y\hat{x}}^{j} - d_{y}^{j}|) + \beta \right].
\]

**Proof of Lemma 4.** Fix an \( x \). By Lemma 2, we have
\[
P_{y}[\hat{y}_{x} \neq y | x] - P_{y}[y_{x} \neq y | x] = d_{x}^{\hat{y}_{x}} - d_{x}^{y_{x}}.
\]

If \( \hat{y}_{x} = y_{x} \), the above inequality still holds. Thereby,
\[
P_{y}[\hat{y}_{x} \neq y | x] - P_{y}[y_{x} \neq y | x] \leq \sum_{j:y_{j} \neq y_{x}}|d_{x}^{\hat{y}_{x}} - d_{x}^{y_{x}}|.
\]

Recall the definition of \( \ell_{\beta} \), we have
\[
P_{y}[\hat{y}_{x} \neq y | x] - P_{y}[y_{x} \neq y | x] \leq \ell_{\beta}(\sum_{j:y_{j} \neq y_{x}}|d_{x}^{\hat{y}_{x}} - d_{x}^{y_{x}}|) + \beta.
\]

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

\[\square\]

**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 \( f \in \mathcal{F} \)
\[
\mathbb{E}[I(\hat{y}_{x} \neq y)] - L_{2}^{*} \leq (\hat{R}_{f}(f) + \beta) + \frac{2 \sqrt{2} m \Lambda_{r}}{\sqrt{n}} + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.
\]

**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
\[
\mathbb{E}_{y,x}[I(\hat{y}_{x} \neq y)] - L_{2}^{*} \leq \mathbb{E} \left[ \ell_{\beta}(\sum_{j:y_{j} \neq y_{x}}|d_{x}^{\hat{y}_{x}} - d_{x}^{y_{x}}|) \right],
\]

because \( \mathbb{E}_{y,x}[I(\hat{y}_{x} \neq y)] - L_{2}^{*} \leq 1 \). It suffices to bound the right-hand side of the above equation.

Define \( \mathcal{H} = \{ z = (x, D) \mapsto \sum_{j:y_{j} \neq y_{x}}|f_{j}(x) - d_{x}^{y_{x}}| : f \in \mathcal{F} \} \). Applying a standard Rademacher bound (Mohri et al., 2018) to \( \ell_{\beta} \circ \mathcal{H} \), for any \( \delta > 0 \), with probability at least \( 1 - \delta \), the following bound holds for all \( f \in \mathcal{F} \)
\[
\mathbb{E}\left[ \ell_{\beta}(\sum_{j:y_{j} \neq y_{x}}|d_{x}^{\hat{y}_{x}} - d_{x}^{y_{x}}|) \right] \leq \hat{R}_{f}(f) + 2 \mathcal{R}_{n}(\ell_{\beta} \circ \mathcal{H}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.
\]

By the 1-Lipschitzness of \( \ell_{\beta} \), it follows that
\[
\mathcal{R}_{n}(\ell_{\beta} \circ \mathcal{H}) \leq \mathcal{R}_{n}(\mathcal{H}).
\]
Define $\ell(D, \hat{D}) = \sum_{j: y_j \neq \hat{y}_j} |\hat{d}_{x_j} - d_{x_j}|$. Then, $\mathcal{H}$ can be equivalently re-written as $\ell \circ \mathcal{F}$. Notice that $\ell$ 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
\[
\mathcal{R}_n(\mathcal{H}) \leq \frac{\sqrt{2m\Lambda r}}{\sqrt{n}},
\]
which leads to
\[
\mathbb{E}_{y,x}[|\hat{y}_x \neq y|] - L^*_2 \leq \hat{R}_\beta(f) + \frac{2\sqrt{2m\Lambda r}}{\sqrt{n}} + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}.
\]

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

References

