

Appendix: Improving the Generalization Performance of Multi-class SVM via Angular Regularization

Jianxin Li¹, Haoyi Zhou¹, Pengtao Xie^{2,3}, Yingchun Zhang¹

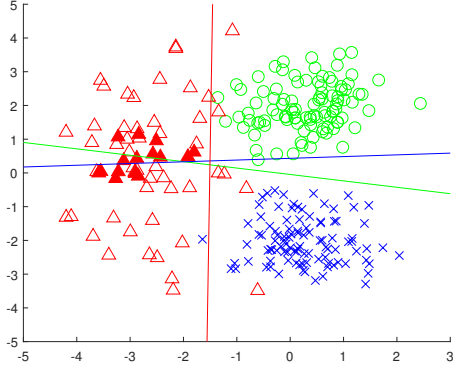
¹ School of Computer Science and Engineering, Beihang University

² Machine Learning Department, Carnegie Mellon University

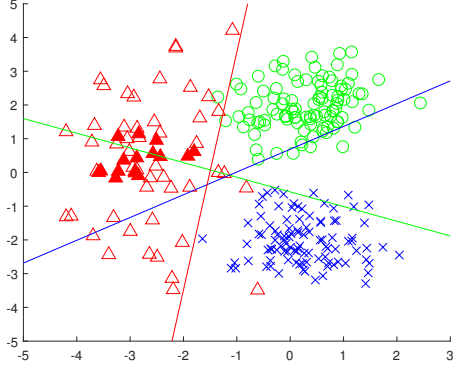
³ Petuum Inc, USA

1 Supplementary for intuitive figure

Recall the intuitive figure Fig. 1 (in main paper), it is better to run an example of how the angular regularization technique taking effects on multi-class SVM. Here, we have a try on randomized toy data to verify its correctness.



(a)



(b)

Figure 1: (a) Without angular regularization; (b) with angular regularization. The toy data is generated on a 2D-plane with 3 categories. Remarking that the solid red triangles are the training samples while the void ones denote test samples.

2 Proof of Lemma 1

To prove Theorem 1 (in main paper), the following lemma is needed.

Lemma 1. *Let the weight vector \mathbf{w}_k of hyperplane k be decomposed into $\mathbf{w}_k = \mathbf{x}_k + l_k \mathbf{e}_k$, where $\mathbf{x}_k = \sum_{j=1, j \neq k}^K \alpha_j \mathbf{w}_j$ lies in the subspace L spanned by $\{\mathbf{w}_1, \dots, \mathbf{w}_K\} \setminus \{\mathbf{w}_k\}$, \mathbf{e}_k is in the orthogonal complement of L , $\|\mathbf{e}_k\| = 1$, $\mathbf{e}_k \cdot \mathbf{w}_k > 0$, l_k is a scalar. Then the gradient of $\hat{\mathcal{R}}(\mathbf{W})$ w.r.t \mathbf{w}_k is $p_k \mathbf{x}_k + q_k \mathbf{e}_k$, where p_k is a positive scalar.*

Proof To prove Lemma 1, the following lemma is needed.

Lemma 2. [Xie et al., 2015a] *Let the weight vector \mathbf{w}_k of hyperplane k be decomposed into $\mathbf{w}_k = \mathbf{x}_k + l_k \mathbf{e}_k$, where $\mathbf{x}_k = \sum_{j=1, j \neq k}^K \alpha_j \mathbf{w}_j$ lies in the subspace L spanned by $\{\mathbf{w}_1, \dots, \mathbf{w}_K\} \setminus \{\mathbf{w}_k\}$, \mathbf{e}_k is in the orthogonal complement of L , $\|\mathbf{e}_k\| = 1$, $\mathbf{e}_k \cdot \mathbf{w}_k > 0$, l_k is a scalar. Then $\det(\mathbf{W}^\top \mathbf{W}) = \det(\mathbf{W}_{-k}^\top \mathbf{W}_{-k}) (l_k \mathbf{e}_k \cdot \mathbf{w}_k)$, where $\mathbf{W}_{-i} = [\mathbf{w}_1, \dots, \mathbf{w}_{i-1}, \mathbf{w}_{i+1}, \dots, \mathbf{w}_K]$ with \mathbf{w}_i excluded.*

According to the chain rule, the gradient of $\mathcal{R}(\mathbf{W})$ w.r.t \mathbf{w}_k can be written as

$$\begin{aligned} \frac{\partial \mathcal{R}(\mathbf{W})}{\partial \mathbf{w}_k} &= g'(\text{tr}(\mathbf{W}^\top \mathbf{W})) \frac{\partial \text{tr}(\mathbf{W}^\top \mathbf{W})}{\partial \mathbf{w}_k} \\ &\quad - \frac{1}{K} g'(\det(\mathbf{W}^\top \mathbf{W})) \frac{\partial \det(\mathbf{W}^\top \mathbf{W})}{\partial \mathbf{w}_k}, \end{aligned}$$

where $g(x) = \log(x)$. It is easy to check that $g(x)$ is an increasing function and $g'(x) = 1/x$. As assumed earlier, the weight vectors in \mathbf{W} are linearly independent and hence $\text{tr}(\mathbf{W}^\top \mathbf{W}) > 0$ and $\det(\mathbf{W}^\top \mathbf{W}) > 0$.

According to Lemma 2, we have

$$\frac{\partial \det(\mathbf{W}^\top \mathbf{W})}{\partial \mathbf{w}_k} = \det(\mathbf{W}_{-k}^\top \mathbf{W}_{-k}) l_k \mathbf{e}_k,$$

where $\det(\mathbf{W}_{-k}^\top \mathbf{W}_{-k}) > 0$ and $l_k > 0$ (knowing from $\det(\mathbf{W}^\top \mathbf{W}) = \det(\mathbf{W}_{-k}^\top \mathbf{W}_{-k}) l_k \mathbf{e}_k \cdot \mathbf{w}_k > 0$ and $\mathbf{e}_k \cdot \mathbf{w}_k > 0$). Besides, we have

$$\frac{\partial \text{tr}(\mathbf{W}^\top \mathbf{W})}{\partial \mathbf{w}_k} = 2\mathbf{w}_k,$$

Substitute above equations into the gradient of $\widehat{\mathcal{R}}(\mathbf{W})$

$$\frac{\partial \widehat{\mathcal{R}}(\mathbf{W})}{\partial \mathbf{w}_k} = \frac{2\mathbf{w}_k}{\text{tr}(\mathbf{W}^\top \mathbf{W})} - \frac{1}{K} \frac{\det(\mathbf{W}_{-k}^\top \mathbf{W}_{-k}) l_k \mathbf{e}_k}{\det(\mathbf{W}^\top \mathbf{W})}.$$

With the weight vector \mathbf{w}_k decomposition, we have

$$\begin{aligned} \frac{\partial \widehat{\mathcal{R}}(\mathbf{W})}{\partial \mathbf{w}_k} &= \frac{2}{\text{tr}(\mathbf{W}^\top \mathbf{W})} \mathbf{x}_k \\ &\quad + \left[\frac{2}{\text{tr}(\mathbf{W}^\top \mathbf{W})} - \frac{\det(\mathbf{W}_{-k}^\top \mathbf{W}_{-k})}{K \det(\mathbf{W}^\top \mathbf{W})} \right] l_k \mathbf{e}_k \\ &= p_k \mathbf{x}_k + q_k \mathbf{e}_k, \end{aligned}$$

where $p_k = \frac{2}{\text{tr}(\mathbf{W}^\top \mathbf{W})} > 0$ and $q_k = \left[\frac{2}{\text{tr}(\mathbf{W}^\top \mathbf{W})} - \frac{\det(\mathbf{W}_{-k}^\top \mathbf{W}_{-k})}{K \det(\mathbf{W}^\top \mathbf{W})} \right] l_k$. \square

3 Supplement to the proof of Theorem 1

In the main paper, we give a brief proof of Theorem 1 and ignore the time stamp t for simplicity. Our aim is to declare that the minimum angle $\mathcal{R}(\mathbf{W})$ will decrease alongside with the negative gradient direction of the regularizer $\widehat{\mathcal{R}}(\mathbf{W})$. We use its cosine similarity to measure the minimum angle (denotes by *):

$$s_{\min}(\mathbf{W}^{(t)}) = \cos(-\mathcal{R}(\mathbf{W}^{(t)})) = \cos(\theta_{i*j*}). \quad (1)$$

Equally, we first analysis $\widehat{\mathcal{R}}(\mathbf{W})$'s behaviour on a trivial angle θ_{ij} :

$$s_{ij}(\mathbf{W}^{(t)}) = \cos(\theta_{ij}). \quad (2)$$

Some notations need to be introduced. Let $V = \{(i, j) | 1 \leq i, j \leq K, i \neq j, \mathbf{w}_i^{(t)} \cdot \mathbf{w}_j^{(t)} = 0\}$, $N = \{(i, j) | 1 \leq i, j \leq K, i \neq j, \mathbf{w}_i^{(t)} \cdot \mathbf{w}_j^{(t)} \neq 0\}$, where $\mathbf{w}_i^{(t)}$ is the i -th column of \mathbf{W}_t . Let $x_{ij}^{(t)} = \mathbf{w}_i^{(t)} \cdot \mathbf{w}_j^{(t)}$, $y_{ij}^{(t)} = \|\mathbf{w}_i^{(t)}\|_2 \cdot \|\mathbf{w}_j^{(t)}\|_2$, $x_{ij}^{(t+1)} = \mathbf{w}_i^{(t+1)} \cdot \mathbf{w}_j^{(t+1)}$, $y_{ij}^{(t+1)} = \|\mathbf{w}_i^{(t+1)}\|_2 \cdot \|\mathbf{w}_j^{(t+1)}\|_2$. Following the gradient direction of $\widehat{\mathcal{R}}(\mathbf{W})$ from Lemma 1, we have

$$\begin{aligned} \mathbf{w}_i^{(t+1)} &= \mathbf{w}_i^{(t)} - \eta(p_i \mathbf{x}_i + q_i \mathbf{e}_i) \\ \mathbf{w}_j^{(t+1)} &= \mathbf{w}_j^{(t)} - \eta(p_j \mathbf{x}_j + q_j \mathbf{e}_j), \end{aligned}$$

and acquire some important equations

$$\begin{aligned} \mathbf{e}_i \cdot \mathbf{w}_j^{(t)} &= 0, & \mathbf{e}_j \cdot \mathbf{w}_i^{(t)} &= 0 \\ \mathbf{e}_i \cdot \mathbf{x}_i &= 0, & \mathbf{e}_j \cdot \mathbf{x}_j &= 0 \\ \mathbf{e}_i \cdot \mathbf{x}_j &= \alpha_i \mathbf{e}_i \cdot \mathbf{w}_i^{(t)}, & \mathbf{e}_j \cdot \mathbf{x}_i &= \alpha_j \mathbf{e}_j \cdot \mathbf{w}_j^{(t)} \end{aligned}$$

Thus, we have:

$$\begin{aligned} x_{ij}^{(t+1)} &= [\mathbf{w}_i^{(t)} - \eta(p_i \mathbf{x}_i + q_i \mathbf{e}_i)] [\mathbf{w}_j^{(t)} - \eta(p_j \mathbf{x}_j + q_j \mathbf{e}_j)] \\ &= \mathbf{w}_i^{(t)} \cdot \mathbf{w}_j^{(t)} - \eta(p_j \mathbf{w}_i^{(t)} \cdot \mathbf{x}_j + p_i \mathbf{x}_i \cdot \mathbf{w}_j^{(t)}) \\ &\quad + \eta^2(p_i \mathbf{x}_i + q_i \mathbf{e}_i)(p_j \mathbf{x}_j + q_j \mathbf{e}_j) \\ &= \mathbf{w}_i^{(t)} \cdot \mathbf{w}_j^{(t)} - \eta a + \eta^2 b \end{aligned}$$

where $a = p_j \mathbf{w}_i^{(t)} \cdot \mathbf{x}_j + p_i \mathbf{x}_i \cdot \mathbf{w}_j^{(t)}$;

$$\begin{aligned} y_{ij}^{(t+1)} &= \sqrt{(\mathbf{w}_i^{(t)} - \eta(p_i \mathbf{x}_i + q_i \mathbf{e}_i))^2} \sqrt{(\mathbf{w}_j^{(t)} - \eta(p_j \mathbf{x}_j + q_j \mathbf{e}_j))^2} \\ &= \sqrt{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)} - 2\eta(p_i \mathbf{x}_i \cdot \mathbf{x}_i + q_i l_i) + \eta^2(p_i^2 \mathbf{x}_i \cdot \mathbf{x}_i + q_i^2)} \\ &\quad \cdot \sqrt{\mathbf{w}_j^{(t)} \cdot \mathbf{w}_j^{(t)} - 2\eta(p_j \mathbf{x}_j \cdot \mathbf{x}_j + q_j l_j) + \eta^2(p_j^2 \mathbf{x}_j \cdot \mathbf{x}_j + q_j^2)} \\ &= \sqrt{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)} - 2\eta c + \eta^2 d} \sqrt{\mathbf{w}_j^{(t)} \cdot \mathbf{w}_j^{(t)} - 2\eta e + \eta^2 f} \end{aligned}$$

where $c = p_i \mathbf{x}_i \cdot \mathbf{x}_i + q_i l_i$ and $e = p_i \mathbf{x}_i \cdot \mathbf{x}_i + q_i l_i$.

Thereby, we can represent the cosine similarity s_{ij} as

$$s_{ij}(\mathbf{W}^{(t+1)}) = \frac{x_{ij}^{(t+1)}}{y_{ij}^{(t+1)}}, \quad s_{ij}(\mathbf{W}^{(t)}) = \frac{x_{ij}^{(t)}}{y_{ij}^{(t)}}$$

The following lemma 3 and lemma 4 are needed for proving Theorem 1.

Lemma 3. $\forall (i, j) \in V$, we have $s_{ij}(\mathbf{W}^{(t+1)}) - s_{ij}(\mathbf{W}^{(t)}) = o(\eta)$, where $\lim_{\eta \rightarrow 0} \frac{o(\eta)}{\eta} = 0$.

Proof For $(i, j) \in V$, $\mathbf{w}_i^{(t)} \cdot \mathbf{w}_j^{(t)} = 0$, thereby $x_{ij}^{(t)} = 0$ and

$$s_{ij}(\mathbf{W}^{(t+1)}) - s_{ij}(\mathbf{W}^{(t)}) = x_{ij}^{(t+1)} / y_{ij}^{(t+1)} - 0.$$

Simply we have $x_{ij}^{(t+1)} = -\eta a + \eta^2 b$, where

$$\begin{aligned} a &= p_j \mathbf{w}_i^{(t)} \cdot \mathbf{x}_j + p_i \mathbf{x}_i \cdot \mathbf{w}_j^{(t)} \\ &= p_j \mathbf{w}_i^{(t)} \cdot (\mathbf{w}_j^{(t)} - l_j \mathbf{e}_j) + p_i (\mathbf{w}_i^{(t)} - l_i \mathbf{e}_i) \cdot \mathbf{w}_j^{(t)} \\ &= p_j \mathbf{w}_i^{(t)} \cdot \mathbf{w}_j^{(t)} + p_i \mathbf{w}_i^{(t)} \cdot \mathbf{w}_j^{(t)} \\ &= (p_j + p_i) \mathbf{w}_i^{(t)} \cdot \mathbf{w}_j^{(t)} \\ &= 0 \end{aligned}$$

Thus we can derive $x_{ij}^{(t+1)} = \eta^2 b$.

Next we consider $1/y_{ij}^{(t+1)} = \frac{1}{\sqrt{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)} - 2\eta \cdot c + \eta^2 \cdot d}}$. $\frac{1}{\sqrt{\mathbf{w}_j^{(t)} \cdot \mathbf{w}_j^{(t)} - 2\eta \cdot e + \eta^2 \cdot f}}$, we take $\|\mathbf{w}_i^{(t)}\|_2$ out and discuss the first term in denominator. Using the Taylor expansion of $\frac{1}{\sqrt{1+x}}$ at $x = 0$, we obtain that

$$\begin{aligned} &\frac{1}{\|\mathbf{w}_i^{(t)}\|_2 \sqrt{1 + \frac{-2\eta \cdot c + \eta^2 \cdot d}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)}}}} \\ &= \frac{1}{\|\mathbf{w}_i^{(t)}\|_2} \left[1 - \frac{1}{2} \left(\frac{-2\eta \cdot c + \eta^2 \cdot d}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)}} \right) + o\left(\frac{-2\eta \cdot c + \eta^2 \cdot d}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)}} \right) \right] \\ &= \frac{1}{\|\mathbf{w}_i^{(t)}\|_2} \left[1 + \frac{\eta \cdot c}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)}} + o(\eta) \right] \end{aligned}$$

Thereby, we have

$$\begin{aligned} &1/y_{ij}^{(t+1)} \\ &= \frac{1}{\|\mathbf{w}_i^{(t)}\|_2} \left[1 + \frac{\eta \cdot c}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)}} + o(\eta) \right] \\ &\quad \cdot \frac{1}{\|\mathbf{w}_j^{(t)}\|_2} \left[1 + \frac{\eta \cdot e}{\mathbf{w}_j^{(t)} \cdot \mathbf{w}_j^{(t)}} + o(\eta) \right] \\ &= \frac{1}{\|\mathbf{w}_i^{(t)}\|_2 \|\mathbf{w}_j^{(t)}\|_2} \left[1 + \left(\frac{c}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)}} + \frac{e}{\mathbf{w}_j^{(t)} \cdot \mathbf{w}_j^{(t)}} \right) \eta + o(\eta) \right] \end{aligned}$$

Now we prove $\lim_{\eta \rightarrow 0} \frac{s_{ij}(\mathbf{W}^{(t+1)}) - s_{ij}(\mathbf{W}^{(t)})}{\eta} = 0$. Consider $x_{ij}^{(t+1)}/y_{ij}^{(t+1)} = \frac{\eta^2 b}{\|\mathbf{w}_i^{(t)}\| \|\mathbf{w}_j^{(t)}\|} [1 + (\frac{c}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)}} + \frac{e}{\mathbf{w}_j^{(t)} \cdot \mathbf{w}_j^{(t)}})\eta + o(\eta)] = o(\eta)$, then $\lim_{\eta \rightarrow 0} \frac{x_{ij}^{(t+1)}/y_{ij}^{(t+1)}}{\eta} = 0$ and $\lim_{\eta \rightarrow 0} \frac{x_{ij}^{(t+1)}}{y_{ij}^{(t+1)}} = 0$. Thereby we have

$$\lim_{\eta \rightarrow 0} \frac{s_{ij}(\mathbf{W}^{(t+1)}) - s_{ij}(\mathbf{W}^{(t)})}{\eta} = \lim_{\eta \rightarrow 0} \frac{x_{ij}^{(t+1)}/y_{ij}^{(t+1)} - 0}{\eta} = 0$$

The proof completes. \square

Lemma 4. $\forall (i, j) \in N$, $\exists \psi_{ij} < 0$, such that $s_{ij}(\mathbf{W}^{(t+1)})/s_{ij}(\mathbf{W}^{(t)}) = 1 + \psi_{ij}\eta + o(\eta)$, where $\lim_{\eta \rightarrow 0} \frac{o(\eta)}{\eta} = 0$.

Proof For $(i, j) \in N$, $\mathbf{w}_i^{(t)} \cdot \mathbf{w}_j^{(t)} \neq 0$, thereby

$$s_{ij}(\mathbf{W}^{(t+1)})/s_{ij}(\mathbf{W}^{(t)}) = \frac{x_{ij}^{(t+1)}/y_{ij}^{(t+1)}}{x_{ij}^{(t)}/y_{ij}^{(t)}}$$

According to the definition of $x_{ij}^{(t+1)}/y_{ij}^{(t+1)}$, we have

$$x_{ij}^{(t+1)}/y_{ij}^{(t+1)} = \frac{x_{ij}^{(t)}}{y_{ij}^{(t)}} \frac{1 + \frac{-\eta a + \eta^2 b}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_j^{(t)}}}{\sqrt{1 + \frac{-2\eta c + \eta^2 d}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)}}} \sqrt{1 + \frac{-2\eta e + \eta^2 f}{\mathbf{w}_j^{(t)} \cdot \mathbf{w}_j^{(t)}}}}$$

Easily $1 + \frac{-\eta a + \eta^2 b}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_j^{(t)}} = 1 - \frac{a}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_j^{(t)}}\eta + o(\eta)$. Recall the analysis in previous proof, $\frac{1}{\sqrt{1 + \frac{-2\eta c + \eta^2 d}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)}}} \sqrt{1 + \frac{-2\eta e + \eta^2 f}{\mathbf{w}_j^{(t)} \cdot \mathbf{w}_j^{(t)}}}} = 1 + (\frac{c}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)}} + \frac{e}{\mathbf{w}_j^{(t)} \cdot \mathbf{w}_j^{(t)}})\eta + o(\eta)$. Substituting the above equations to $x_{ij}^{(t+1)}/y_{ij}^{(t+1)}$, we can obtain that

$$\begin{aligned} & x_{ij}^{(t+1)}/y_{ij}^{(t+1)} \\ &= \frac{x_{ij}^{(t)}}{y_{ij}^{(t)}} [1 - \frac{a}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_j^{(t)}}\eta + o(\eta)] \\ & \quad \cdot [1 + (\frac{c}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)}} + \frac{e}{\mathbf{w}_j^{(t)} \cdot \mathbf{w}_j^{(t)}})\eta + o(\eta)] \\ &= \frac{x_{ij}^{(t)}}{y_{ij}^{(t)}} [1 - (\frac{a}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_j^{(t)}} - \frac{c}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)}} - \frac{e}{\mathbf{w}_j^{(t)} \cdot \mathbf{w}_j^{(t)}})\eta + o(\eta)] \end{aligned}$$

Thereby,

$$\begin{aligned} s_{ij}(\mathbf{W}^{(t+1)})/s_{ij}(\mathbf{W}^{(t)}) &= \frac{x_{ij}^{(t+1)}/y_{ij}^{(t+1)}}{x_{ij}^{(t)}/y_{ij}^{(t)}} \\ &= 1 - (\frac{a}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_j^{(t)}} - \frac{c}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)}} - \frac{e}{\mathbf{w}_j^{(t)} \cdot \mathbf{w}_j^{(t)}})\eta + o(\eta) \end{aligned}$$

Let $\psi = -[\frac{a}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_j^{(t)}} - \frac{c}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)}} - \frac{e}{\mathbf{w}_j^{(t)} \cdot \mathbf{w}_j^{(t)}}]$ and then we prove $\psi < 0$. Consider the first term

$$\begin{aligned} \frac{a}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_j^{(t)}} &= \frac{(p_j + p_i)\mathbf{w}_i^{(t)} \cdot \mathbf{w}_j^{(t)}}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_j^{(t)}} \\ &= p_j + p_i \end{aligned}$$

Actually, we have $p_i = p_j = \frac{2}{\text{tr}(\mathbf{W}^\top \mathbf{W})} > 0$ and it indicates that $\frac{a}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_j^{(t)}} > 0$.

Before moving further, we look into the decomposition $\mathbf{w}_k = \mathbf{x}_k + l_k \mathbf{e}_k$ firstly. If we square both side of the equation,

$$\mathbf{w}_k \cdot \mathbf{w}_k = \mathbf{x}_k \cdot \mathbf{x}_k + l_k^2,$$

notice $\mathbf{w}_k \cdot \mathbf{w}_k$ is square of the norm of \mathbf{w}_k . With simple algebraic geometry, we have $\mathbf{x}_k \cdot \mathbf{x}_k \geq 0$ and $\mathbf{w}_k \mathbf{w}_k \geq l_k^2$. Consider the last two terms

$$\begin{aligned} \frac{c}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)}} &= \frac{p_i \mathbf{x}_i \cdot \mathbf{x}_i + q_i l_i}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)}} > 0, \\ \frac{d}{\mathbf{w}_j^{(t)} \cdot \mathbf{w}_j^{(t)}} &= \frac{p_j \mathbf{x}_j \cdot \mathbf{x}_j + q_j l_j}{\mathbf{w}_j^{(t)} \cdot \mathbf{w}_j^{(t)}} > 0. \end{aligned}$$

Substitute above equation and we have

$$\begin{aligned} \frac{c}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)}} &= \frac{p_i(\mathbf{x}_i \cdot \mathbf{x}_i) + q_i l_i}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)}} \\ &= \frac{p_i(\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)} - l_i^2) + q_i l_i}{\mathbf{w}_i^{(t)} \cdot \mathbf{w}_i^{(t)}} \\ &= p_i + \frac{q_i l_i - p_i l_i^2}{\|\mathbf{w}_i^{(t)}\|^2} \end{aligned}$$

and

$$\frac{d}{\mathbf{w}_j^{(t)} \cdot \mathbf{w}_j^{(t)}} = p_j + \frac{q_j l_j - p_j l_j^2}{\|\mathbf{w}_j^{(t)}\|^2}$$

Thereby, we have

$$\begin{aligned} \psi &= -[(p_i + p_j) - (p_i + \frac{q_i l_i - p_i l_i^2}{\|\mathbf{w}_i^{(t)}\|^2}) - (p_j + \frac{q_j l_j - p_j l_j^2}{\|\mathbf{w}_j^{(t)}\|^2})] \\ &= -[\frac{p_i l_i^2 - q_i l_i}{\|\mathbf{w}_i^{(t)}\|^2} + \frac{p_j l_j^2 - q_j l_j}{\|\mathbf{w}_j^{(t)}\|^2}] \end{aligned}$$

If $\frac{p_i l_i^2 - q_i l_i}{\|\mathbf{w}_i^{(t)}\|^2} > 0$ and $\frac{p_j l_j^2 - q_j l_j}{\|\mathbf{w}_j^{(t)}\|^2} > 0$, we can immediately draw the conclusion that $\psi < 0$.

Obviously, we only need to discuss one case. From Lemma 2, we have $p_i = \frac{2}{\text{tr}(\mathbf{W}^\top \mathbf{W})} > 0$ and $q_i = \frac{2}{\text{tr}(\mathbf{W}^\top \mathbf{W})} - \frac{\det(\mathbf{W}_{-i}^\top \mathbf{W}_{-i})}{K \det(\mathbf{W}^\top \mathbf{W})} l_i$. Let $\Omega_i = \frac{\det(\mathbf{W}_{-i}^\top \mathbf{W}_{-i})}{K \det(\mathbf{W}^\top \mathbf{W})}$, we have

$$\begin{aligned} q_i &= (p_i - \Omega_i) l_i, \\ \Omega_i &= \frac{\det(\mathbf{W}_{-i}^\top \mathbf{W}_{-i})}{K \cdot \det(\mathbf{W}_{-i}^\top \mathbf{W}_{-i}) (l_i \mathbf{e}_i \cdot \mathbf{w}_i^{(t)})} \\ &= \frac{1}{K l_i \mathbf{e}_i \cdot \mathbf{w}_i^{(t)}} \end{aligned}$$

Thereby,

$$\begin{aligned} \frac{p_i l_i^2 - q_i l_i}{\|\mathbf{w}_i^{(t)}\|^2} &= \frac{p_i l_i^2 - (p_i - \Omega_i) l_i \cdot l_i}{\|\mathbf{w}_i^{(t)}\|^2} \\ &= \frac{\Omega_i l_i^2}{\|\mathbf{w}_i^{(t)}\|^2}, \\ &= \frac{l_i}{K(\mathbf{e}_i \cdot \mathbf{w}_i^{(t)}) \|\mathbf{w}_i^{(t)}\|^2} \end{aligned}$$

where K denotes the number of hyperplanes, $l_i > 0$ and $(\mathbf{e}_i \cdot \mathbf{w}_i^{(t)}) > 0$ (discussed in Lemma 2). Thus,

$$\frac{p_i l_i^2 - q_i l_i}{\|\mathbf{w}_i^{(t)}\|^2} > 0.$$

The proof completes. \square

Given these two lemmas, we can prove Theorem 1 now.

Proof For the cosine similarity $s(\mathbf{W}^{(t)})$ between hyperplanes \mathbf{w}_i and \mathbf{w}_j ,

A. if $s_{ij}(\mathbf{W}^{(t)}) \in V$,

$$\lim_{\eta \rightarrow 0} \frac{s_{ij}(\mathbf{W}^{(t+1)}) - s_{ij}(\mathbf{W}^{(t)})}{\eta} = \lim_{\eta \rightarrow 0} \frac{o(\eta)}{\eta} = 0$$

If $s_{min}(\mathbf{W}^{(t)}) = 0$, we have

$$\lim_{\eta \rightarrow 0} \frac{s_{min}(\mathbf{W}^{(t+1)}) - s_{min}(\mathbf{W}^{(t)})}{\eta} = 0. \quad (3)$$

B. if $s_{ij}(\mathbf{W}^{(t)}) \in N$,

$$\begin{aligned} \lim_{\eta \rightarrow 0} \frac{s_{ij}(\mathbf{W}^{(t+1)}) - s_{ij}(\mathbf{W}^{(t)})}{\eta} \\ = \lim_{\eta \rightarrow 0} \frac{\frac{s_{ij}(\mathbf{W}^{(t+1)})}{s_{ij}(\mathbf{W}^{(t)})} - 1}{\eta} \cdot s_{ij}(\mathbf{W}^{(t)}) \\ = \psi_{ij} \cdot s_{ij}(\mathbf{W}^{(t)}) \end{aligned}$$

Since the minimal angle $\theta_{i^*j^*} \in [0, \frac{\pi}{2}]$, we have $s_{min}(\mathbf{W}^{(t)}) > 0$. Then

$$\lim_{\eta \rightarrow 0} \frac{s_{min}(\mathbf{W}^{(t+1)}) - s_{min}(\mathbf{W}^{(t)})}{\eta} < 0$$

So $\exists \kappa > 0$, such that $\forall \eta \in (0, \kappa)$, we have $\frac{s_{min}(\mathbf{W}^{(t+1)}) - s_{min}(\mathbf{W}^{(t)})}{\eta} < 0$. That is $s_{min}(\mathbf{W}^{(t+1)}) - s_{min}(\mathbf{W}^{(t)}) < 0$.

Combine both **A.** and **B.**, we have:

$$s_{min}(\mathbf{W}^{(t+1)}) - s_{min}(\mathbf{W}^{(t)}) \leq 0.$$

Using the $\cos(\cdot)$'s monotonicity in $[0, \frac{\pi}{2}]$ and $s_{min}(\mathbf{W}^{(t)}) = \cos(-\mathcal{R}(\mathbf{W}^{(t)}))$, we have

$$\mathcal{R}(\mathbf{W}^{(t+1)}) \leq \mathcal{R}(\mathbf{W}^{(t)}), \quad (4)$$

where $\mathbf{W}^{(t+1)} = \mathbf{W}^{(t)} - \eta \nabla \hat{\mathcal{R}}(\mathbf{W}^{(t)})$.

The proof completes. \square

4 Proof of Theorem 2

4.1 Preliminary

We first present the problem setup before further discussion.

- **Task:** Large Margin Machine for Multi-class Learning
- **Input:** $S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$ is a set of m training examples, where \mathbf{x}_i is drawn from a domain $\mathcal{X} \subseteq \mathbb{R}^D$ and each label y_i is an integer from $\mathcal{Y} = \{1, \dots, K\}$.
- **Distribution:** \mathbb{D} represents true distribution of sample S and $\hat{\mathbb{D}}$ is the empirical distribution.
- **Hypothesis set:** Parameterizing with hyperplanes matrix \mathbf{W} , a family of hypotheses is defined

$$H = \{h | h(\mathbf{x}, y) = \mathbf{w}_y^\top \mathbf{x}, \mathbf{x} \in \mathcal{X}, y \in \mathcal{Y}\}, \quad (5)$$

which maps $\mathcal{X} \times \mathcal{Y}$ to \mathbb{R} . For each hypothesis $h \in H$, its multi-class margin for the input-output pair (\mathbf{x}, y) is

$$\rho_h(\mathbf{x}, y) = h(\mathbf{x}, y) - \max_{r \neq y} h(\mathbf{x}, r). \quad (6)$$

- **Loss set:** The hinge loss function is

$$\mathcal{A} = \{l | l(\mathbf{x}, y) = \Phi_p(\rho_h(\mathbf{x}, y))\}, \quad (7)$$

where $\Phi_p(x) = \max(0, 1 - x/p)$ is a p -margin hinge loss taking values in $[0, +\infty)$. Typically, p is set to 1.

- **Error:** The generalization error of a hypothesis h is

$$L(h) = E_{(\mathbf{x}, y) \sim \mathbb{D}}[\Phi_p(\rho_h(\mathbf{x}, y))]. \quad (8)$$

And the training error of a hypothesis h is

$$\hat{L}(h) = E_{(\mathbf{x}, y) \sim \hat{\mathbb{D}}}[\Phi_p(\rho_h(\mathbf{x}, y))]. \quad (9)$$

Moreover, we need extra bounds in error analysis.

- Let input vector $\mathbf{x}_i \in \mathbb{R}^n$ bounded with $\|\mathbf{x}_i\|_2 \leq C_1$.
- Let each hyperplane $\mathbf{w}_r \in \mathbb{R}^n$ w.r.t class r bounded with $\|\mathbf{w}_r\| \leq C_2$. The minimal pairwise error is $-\mathcal{R}(\mathbf{W}) = \min_{i \neq j} \theta_{ij}$, denoting by θ_{min} for short.

The Rademacher complexity $R_m(\mathcal{A})$ of the loss function set \mathcal{A} is defined as $R_m(\mathcal{A}) = \mathbb{E}[\sup_{l \in \mathcal{A}} \frac{1}{m} \sum_{i=1}^m \sigma_i \cdot l(\mathbf{x}_i, y_i)]$, where σ_i is uniform over $\{-1, 1\}$ and $(\mathbf{x}_i, y_i)_{i=1}^m$ are i.i.d samples drawn from \mathbb{D} . The following analysis is based on Bartlett and Mendelson; Percy [2002; 2015]'s Corollary.

Lemma 5. Fix $p > 0$. With probability at least $1 - \delta$

$$L(\hat{h}) - L(h^*) \leq 4R_m(\mathcal{A}) + B \sqrt{\frac{2 \log(2/\delta)}{m}} \quad (10)$$

for $B \geq \sup_{\mathbf{x}, y, l} |l(\mathbf{x}, y)|$.

Regarding to two situations H^0 and H^1 , further sequentially boundings on $R_m(\mathcal{A})$ and B complete the proof.

4.2 A1. Upper bound the term $R_m(\mathcal{A})$

Firstly, we should find an upper bound of the Rademacher complexity of the hypothesis set H , that is

$$R_m(\mathcal{A}) = \mathbb{E}[\sup_{l \in \mathcal{A}} \frac{1}{m} \sum_{i=1}^m \sigma_i \cdot l(\mathbf{x}_i, y_i)]. \quad (11)$$

However, the hypothesis set H has no direct connection with the loss function set (no matter $0 \cdot 1$ or hinge loss), defined on the multi-class margin. Let \tilde{H} be the family of hypothesis mapping $\mathcal{X} \times \mathcal{Y}$ to \mathbb{R} defined by $\tilde{H} = \{z = (\mathbf{x}, y) \mapsto \rho_h(\mathbf{x}, y) : \mathbf{x} \in \mathcal{X}, y \in \mathcal{Y}, h \in H\}$. Then we expand the intractable Rademacher complexity term as

$$R_m(\mathcal{A}) = R_n(l \circ \tilde{H}). \quad (12)$$

To employ the conclusion from [Bartlett and Mendelson, 2002], we rewrite the Rademacher complexity $R_m(\mathcal{A})$ in Eq.(11)'s definiton in absolute form

$$R_m^\parallel(\mathcal{A}) = \mathbb{E}[\sup_{l \in \mathcal{A}} \frac{2}{m} \sum_{i=1}^m |\sigma_i \cdot l(\mathbf{x}_i, y_i)|]. \quad (13)$$

Considering previous definition with the loss function $l(\cdot) \geq 0$ in Eq.(11), we have

$$R_m(\mathcal{A}) \leq \frac{1}{2} R_m^\parallel(\mathcal{A}). \quad (14)$$

From Eq.(12), it is natural that

$$R_m^\parallel(\mathcal{A}) = R_m^\parallel(l \circ \tilde{H}). \quad (15)$$

Bounding $R_m^\parallel(\mathcal{A})$ equals bounding $R_m^\parallel(\tilde{H})$ through discussing the L-Lipschitz property on two kinds of loss function, and the analysis of $R_m^\parallel(\tilde{H})$ is put in advance for brevity. It requires us to extend the proof of Theorem 1 in Cortes *et al.* [2013]'s work.

Let $H_{\mathcal{X}}$ denotes a set of functions defined over \mathcal{X} and derived from H as follows: $H_{\mathcal{X}} = \{\mathbf{x} \mapsto h(\mathbf{x}, y) : y \in \mathcal{Y}, h \in H\}$. For any fixed $y \in \mathcal{Y}$, $H_{\mathcal{X}}$ only takes \mathbf{x} into consideration and its empirical Rademacher complexity becomes an upper bound of \tilde{H} , which is given in Lemma 6 (proof is given in Sec. 5).

Lemma 6. *With the definition of $\rho_h(\mathbf{x}, y)$ and \tilde{H} , the empirical Rademacher complexity:*

$$R_m^\parallel(\tilde{H}) \leq K^2 R_m^\parallel(H_{\mathcal{X}}). \quad (16)$$

Next, we bound $R_m^\parallel(H_{\mathcal{X}})$ [Xie *et al.*, 2015b]. For any $y \in \mathcal{Y}$ we have following key steps

$$\begin{aligned} R_m^\parallel(H_{\mathcal{X}}) &= \mathbb{E}[\sup_{h \in H_{\mathcal{X}}} \frac{2}{m} |\sum_{i=1}^m \sigma_i \mathbf{w}_y^\top \mathbf{x}_i|] \\ &\leq \frac{2C_2}{m} \mathbb{E}[\|\sum_{i=1}^m \sigma_i \mathbf{x}_i\|_2] \quad (\|\mathbf{w}_r\| \leq C_2) \\ &= \frac{2C_2}{m} \mathbb{E}_{\mathbb{D}}[\mathbb{E}_{\sigma}[\|\sum_{i=1}^n \sigma_i \mathbf{x}_i\|_2 \mid \mathbf{x}_i \sim \mathbb{D}]] \\ &\quad (\text{the definition of Rademacher complexity, expanding Expectation}) \\ &= \frac{2C_2}{m} \mathbb{E}_{\mathbb{D}}[\sqrt{\mathbb{E}_{\sigma}[\sum_{i=1}^n \sigma_i^2 (\mathbf{x}_i)^2 \mid \mathbf{x}_i \sim \mathbb{D}]}] \\ &\quad (\forall i \neq j, \sigma_i \perp \sigma_j) \\ &= \frac{2C_2}{m} \mathbb{E}_{\mathbb{D}}[\sqrt{\sum_{i=1}^n (\mathbf{x}_i)^2}] \\ &\quad (\mathbf{x}_i \text{ are i.i.d samples from } \mathbb{D}, \text{ same to } \mathbf{x}_i) \\ &\leq \frac{2C_1 C_2}{\sqrt{m}} \quad (\|\mathbf{x}_i\| \leq C_1) \end{aligned}$$

Plug above results into Lemma 6,

$$R_m^\parallel(\tilde{H}) \leq \frac{2K^2 C_1 C_2}{\sqrt{m}}. \quad (17)$$

Ultimately, we discuss the Lipschitz connectivity of the $0 \cdot 1$ loss and p -margin hinge loss. It's easy to find that both of them are differentiable almost everywhere, that is, differentiable at every point outside a set of Lebesgue measure zero (e.x. $x = 0$ for $\mathbb{I}(\cdot)$, $x = p$ for $\Phi_p(\cdot)$). We have Lipschitz constants

$$\mathcal{L}_{0 \cdot 1} = 1, \quad \mathcal{L}_{hinge} = \frac{1}{p}. \quad (18)$$

Let $l'(\cdot) = l(\cdot) - l(0)$, then $l'(0) = 0$ and l' is also \mathcal{L} -Lipschitz. Then

$$\begin{aligned} R_m^\parallel(l \circ \tilde{H}) &= R_m^\parallel(l' \circ \tilde{H} + l(0)) \\ &\leq R_m^\parallel(l' \circ \tilde{H}) + \frac{\|l(0)\|_\infty}{\sqrt{m}} \\ &\leq 2 \cdot \mathcal{L} R_m^\parallel(\tilde{H}) + \frac{\|l(0)\|_\infty}{\sqrt{m}} \end{aligned} \quad (19)$$

The last two inequations are derived from Theorem 12.5 and Theorem 12.4 in Bartlett and Mendelson [2002]' work. Note that the $l(0)$ is required to be uniformly bounded function by Bartlett's work. For the $0 \cdot 1$ loss function, $\mathbb{I}(0)$ can be assigned any value t while maintains its measurable property. For the p -margin hinge loss, $\Phi_p(0) = \max(0, 1 - 0/p) = 1$. Without lose of generality, we set $t = 1$ and $l(0)$ is uniformly bounded within $[1 - \epsilon, 1 + \epsilon]$ for all ϵ . So we have $\|l(0)\|_\infty = 1$ with $\epsilon \rightarrow 0$ for both of them. This is different from previous work [Xie *et al.*, 2015b].

So far, we find an upper bound for $4R_m(\mathcal{A})$ in RHS of Eq.(10) through combing Eq.(14, 15, 19, 17) sequentially.

4.3 A2. Find the upper bound B

Finding the maximum of $|l(\mathbf{x}, y)|$ completes our prove. Given the formulation of p -margin hinge loss as $l(\cdot) = \Phi_p(\cdot)$ and its $\frac{1}{p}$ -Lipschitz property, we have

$$\begin{aligned} |\Phi_p(\mathbf{z}) - \Phi_p(\mathbf{0})| &\leq \frac{1}{p} \|\mathbf{z} - \mathbf{0}\| \\ \Rightarrow |\Phi_p(\mathbf{z}) - 1| &\leq \frac{1}{p} \|\mathbf{z}\| \\ \Rightarrow |\Phi_p(\mathbf{z})| &\leq \frac{1}{p} \|\mathbf{z}\| + 1 \end{aligned}$$

With the margin condition $\rho_h(\mathbf{x}, y) \leq p$ and $p \geq 0$ we let $z = \rho_h(\mathbf{x}, y)$:

$$|l(\mathbf{x}, y)| = |\Phi_p(\rho_h(\mathbf{x}, y))| \leq \frac{1}{p} |\rho_h(\mathbf{x}, y)| + 1, \quad (20)$$

where the only remaining issue is to bound $|\Phi_p(\rho_h(\mathbf{x}, y))|$. Due to the different formulation of cost function, this proof is different from [Xie *et al.*, 2015b].

With the hyperplanes set \mathbf{W} , we define an adjunction one $\mathbf{W}_y = \{\mathbf{w}_y \dots \mathbf{w}_y \dots \mathbf{w}_y\}$ for a fixed $y \in \mathcal{Y}$. Let $\mathbf{w}_R^\top \mathbf{x} = \max_{r \in \mathcal{Y} \setminus y} \mathbf{w}_r^\top \mathbf{x}$, $\mathbf{w}_{r^*}^\top \mathbf{x} = \{\mathbf{w}_r^\top \mathbf{x} \mid r \in \mathcal{Y} \setminus \{y, R\}\}$ and we have

$$\mathbf{w}_R^\top \mathbf{x} > \text{all}(\mathbf{w}_{r^*}^\top \mathbf{x}).$$

Moreover, we define $\mathbf{w}_{r^+}^\top \mathbf{x} = \{\mathbf{w}_r^\top \mathbf{x} \mid \mathbf{w}_r^\top \mathbf{x} \geq 0, r \in \mathcal{Y} \setminus \{y, R\}\}$ and $\mathbf{w}_{r^-}^\top \mathbf{x}$ reversely, the inequalities hold

$$\mathbf{w}_R^\top \mathbf{x} > \text{all}(\mathbf{w}_{r^+}^\top \mathbf{x}) \quad , \quad \mathbf{w}_R^\top \mathbf{x} > \text{all}(\mathbf{w}_{r^-}^\top \mathbf{x}).$$

Let $g(\cdot)$ be an auxiliary function as

$$g(z) = |\mathbf{w}_y^\top \mathbf{x} - z|. \quad (21)$$

From the analysis in Sec.6, we have that condition A1c, A2c and B1a all share an important property

$$\mathbf{w}_R^\top \mathbf{x} \leq \mathbf{w}_y^\top \mathbf{x} \Leftrightarrow \max_{r \in \mathcal{Y} \setminus y} \mathbf{w}_r^\top \mathbf{x} \leq \mathbf{w}_y^\top \mathbf{x},$$

which indicates that the labeled input data (\mathbf{x}, y) has the biggest activation other than wrong labels $\mathcal{Y} \setminus y$ in weighted matrix \mathbf{W} . The analysis is given in the extreme situation from the worst \mathbf{W} to perfect \mathbf{W} .

Proposition 1. *If $g(\mathbf{w}_R^\top \mathbf{x}) < g(\text{any}(\mathbf{w}_{r*}^\top \mathbf{x}))$, then we can put the alongside*

$$\begin{aligned} |\mathbf{w}_y^\top \mathbf{x} - \mathbf{w}_R^\top \mathbf{x}| &< |\mathbf{w}_y^\top \mathbf{x} - \text{any}(\mathbf{w}_{r+}^\top \mathbf{x})| \\ |\mathbf{w}_y^\top \mathbf{x} - \mathbf{w}_R^\top \mathbf{x}| &< |\mathbf{w}_y^\top \mathbf{x} - \text{any}(\mathbf{w}_{r-}^\top \mathbf{x})| \\ |\mathbf{w}_y^\top \mathbf{x} - \mathbf{w}_R^\top \mathbf{x}| &= |\mathbf{w}_y^\top \mathbf{x} - \mathbf{w}_R^\top \mathbf{x}| \\ 0 &= |\mathbf{w}_y^\top \mathbf{x} - \mathbf{w}_y^\top \mathbf{x}| \end{aligned}$$

We take square on both sides, add them all and have

$$\begin{aligned} |\mathbf{w}_y^\top \mathbf{x} - \mathbf{w}_R^\top \mathbf{x}|^2 &\leq \frac{1}{K} \sum_{r \in \mathcal{Y}} |\mathbf{w}_y^\top \mathbf{x} - \mathbf{w}_r^\top \mathbf{x}|^2 \\ \Rightarrow |\mathbf{w}_y^\top \mathbf{x} - \mathbf{w}_R^\top \mathbf{x}|^2 &\leq \frac{1}{K} \|\mathbf{W}^\top \mathbf{x} - \mathbf{W}_y^\top \mathbf{x}\|_2^2 \end{aligned}$$

Building further on above proposition, we should bound B by two separated situations.

B1. For hypothesis set $H^0 = \{h(\mathbf{x}, y) = \mathbf{w}_y^\top \cdot \mathbf{x}, \mathbf{w}_y^\top \mathbf{x} < \mathbf{w}_R^\top \mathbf{x}\}$, then $|\rho_h(\mathbf{x}, y)|$ has a natural upper bound by expanding $\rho_h(\mathbf{x}, y)$'s definition on hypothesis set H

$$\begin{aligned} |\rho_h(\mathbf{x}, y)| &= |\mathbf{w}_y^\top \mathbf{x} - \max_{r \in \mathcal{Y} \setminus y} \mathbf{w}_r^\top \mathbf{x}| \\ &\leq |\mathbf{w}_y^\top \mathbf{x}| + |\mathbf{w}_R^\top \mathbf{x}| \\ &\leq \|\mathbf{w}_y\| \cdot \|\mathbf{x}\| + \|\mathbf{w}_R\| \cdot \|\mathbf{x}\| \\ &\leq 2C_1 C_2 \end{aligned} \quad (22)$$

B2. For hypothesis set $H^1 = \{h(\mathbf{x}, y) = \mathbf{w}_y^\top \cdot \mathbf{x}, \mathbf{w}_y^\top \mathbf{x} \geq \mathbf{w}_R^\top \mathbf{x}\}$, $|\rho_h(\mathbf{x}, y)|$ has a tighter upper bound

$$\begin{aligned} |\rho_h(\mathbf{x}, y)|^2 &= |\mathbf{w}_y^\top \mathbf{x} - \max_{r \in \mathcal{Y} \setminus y} \mathbf{w}_r^\top \mathbf{x}|^2 \\ &\leq \frac{1}{K} \|\mathbf{W}^\top \mathbf{x} - \mathbf{W}_y^\top \mathbf{x}\|_2^2 \\ &\leq \frac{1}{K} \|(\mathbf{W} - \mathbf{W}_y)^\top\|_{op}^2 \|\mathbf{x}\|_2^2 \\ &= \frac{1}{K} \|\mathbf{W} - \mathbf{W}_y\|_{op}^2 \|\mathbf{x}\|_2^2 \end{aligned} \quad (23)$$

Following the property of $\|\cdot\|_{op}$ (operator norm)

$$\|\mathbf{W} - \mathbf{W}_y\|_{op}^2 \leq \|\mathbf{W}\|_{op}^2 + \|\mathbf{W}_y\|_{op}^2. \quad (24)$$

Next we can make use of the lower bound of θ_{ij} between weights \mathbf{w}_i and \mathbf{w}_j ($i \neq j$), which is θ_{min} , to get the bound of $\|\mathbf{W}\|_{op}$ (see in Xie *et al.* [2015b]). Here only gives main

steps:

$$\begin{aligned} \|\mathbf{W}\|_{op}^2 &= \sup_{\|\mathbf{u}\|_2=1} \|\mathbf{W}\mathbf{u}\|_2^2 \\ &= \sup_{\|\mathbf{u}\|_2=1} (\mathbf{u}^\top \mathbf{W}^\top \mathbf{W} \mathbf{u}) \\ &= \sup_{\|\mathbf{u}\|_2=2} \sum_{p=1}^K \sum_{q=1}^K \mathbf{u}_p \mathbf{u}_q \mathbf{w}_p \cdot \mathbf{w}_q \\ &\leq \sup_{\|\mathbf{u}\|_2=2} \sum_{p=1}^K \sum_{q=1}^K |\mathbf{u}_p| |\mathbf{u}_q| \|\mathbf{w}_p\| \|\mathbf{w}_q\| \cos(\theta_{pq}) \\ &\leq C_2^2 \sup_{\|\mathbf{u}\|_2=2} \sum_{p=1}^K \sum_{q=1}^K |\mathbf{u}_p| |\mathbf{u}_q| \cos(\theta_{pq}) \\ &\quad (\|\mathbf{w}_r\| \leq C_2) \\ &\leq C_2^2 \sup_{\|\mathbf{u}\|_2=2} \left[\sum_{p=1}^K \sum_{q=1}^K |\mathbf{u}_p| |\mathbf{u}_q| \cos(\theta_{min}) \cdot \right. \\ &\quad \left. \mathbb{I}(p \neq q) + \sum_{p=1}^K |\mathbf{u}_p|^2 \right] \end{aligned}$$

Define $\mathbf{u}' = [|\mathbf{u}_1|, \dots, |\mathbf{u}_K|]^\top$, $Q \in \mathbb{R}^{K \times K}$: $Q_{pq} = \cos \theta_{pq}$ for $p \neq q$ and $Q_{pp} = 1$, then we have $\|\mathbf{u}'\|_2 = \|\mathbf{u}\|$ and

$$\begin{aligned} \|\mathbf{W}\|_{op}^2 &\leq C_2^2 \sup_{\|\mathbf{u}\|_2=2} [\mathbf{u}'^\top Q \mathbf{u}'] \\ &\leq C_2^2 \sup_{\|\mathbf{u}\|_2=2} [\lambda_1(Q) \|\mathbf{u}'\|_2^2] \\ &\leq C_2^2 \lambda_1(Q) \end{aligned}$$

where $\lambda_1(Q)$ denotes the largest eigenvalue of Q and we can derive $\lambda_1(Q) = (K-1) \cos \theta_{min} + 1$. Thus

$$\|\mathbf{W}\|_{op}^2 \leq ((K-1) \cos \theta_{min} + 1) C_2^2. \quad (25)$$

Since columns in \mathbf{W}_y are same, we get

$$\|\mathbf{W}_y\|_{op}^2 \leq K C_2^2. \quad (26)$$

Put above equations together, we have the upper bound:

$$\begin{aligned} |\rho_h(\mathbf{x}, y)| &\leq \sqrt{\frac{1}{K} (\|\mathbf{W}\|_{op}^2 + \|\mathbf{W}_y\|_{op}^2) \|\mathbf{x}\|_2^2} \\ &\leq \sqrt{\frac{1}{K} ((K-1) \cos \theta_{min} + K + 1) C_2^2 \|\mathbf{x}\|_2^2} \\ &\leq \sqrt{(1 - \frac{1}{K}) (\cos \theta_{min} + \frac{K+1}{K-1})} C_1 C_2 \end{aligned} \quad (27)$$

From the above **B1** and **B2**, if the hypothesis set $H \rightsquigarrow H^0$ (such as $W_r = 0$ at the start of the training), then

$$|l(\mathbf{x}, y)| \leq \frac{2}{p} C_1 C_2 + 1. \quad (28)$$

Otherwise the hypothesis set $H \rightsquigarrow H^1$ (while training), then

$$|l(\mathbf{x}, y)| \leq \frac{1}{p} \sqrt{(1 - \frac{1}{K}) (\cos(-\mathcal{R}(\mathbf{W})) + \frac{K+1}{K-1})} C_1 C_2 + 1. \quad (29)$$

4.4 A3. Combination

Finally, combining **A1**'s results and Eq.(20, 28, 29) sequently in **A2**, we acquire desired results in Theorem 2. \square

5 Proof of Lemma 6

Proof For any fixed $y \in \mathcal{Y}$ and any $i \in [1, m]$, define ϵ_i as $2(\mathbb{I}(y = y_i)) - 1$. Naturally $\epsilon_i \in \{-1, +1\}$, the Rademacher random variables σ_i and $\sigma_i \epsilon_i$ follow the same distribution. Thus

$$\begin{aligned}
R_m^\parallel(\tilde{H}) &= \mathbb{E}[\sup_{h \in H} \frac{2}{m} \sum_{i=1}^m |\sigma_i \rho_h(\mathbf{x}_i, y_i)|] \\
&= \frac{2}{m} \mathbb{E}[\sup_{h \in H} \sum_{i=1}^m \sum_{y \in \mathcal{Y}} |\sigma_i \rho_h(\mathbf{x}_i, y) \mathbb{I}(y = y_i)|] \\
&\leq \frac{2}{m} \sum_{y \in \mathcal{Y}} \mathbb{E}[\sup_{h \in H} \sum_{i=1}^m |\sigma_i \rho_h(\mathbf{x}_i, y) \mathbb{I}(y = y_i)|] \\
&\quad (\text{sub-additivity of sup}) \\
&= \frac{2}{m} \sum_{y \in \mathcal{Y}} \mathbb{E}[\sup_{h \in H} \sum_{i=1}^m |\sigma_i \rho_h(\mathbf{x}_i, y) \cdot (\frac{2 \cdot \mathbb{I}(y=y_i)-1}{2} + \frac{1}{2})|] \\
&\leq \frac{1}{m} \sum_{y \in \mathcal{Y}} \mathbb{E}[\sup_{h \in H} \sum_{i=1}^m |\sigma_i \epsilon_i \rho_h(\mathbf{x}_i, y)|] \\
&\quad + \frac{1}{m} \sum_{y \in \mathcal{Y}} \mathbb{E}[\sup_{h \in H} \sum_{i=1}^m |\sigma_i \rho_h(\mathbf{x}_i, y)|] \\
&\quad (\text{substituting } \epsilon \text{ and } |a+b| \leq |a| + |b|) \\
&= \frac{2}{m} \sum_{y \in \mathcal{Y}} \mathbb{E}[\sup_{h \in H} \sum_{i=1}^m |\sigma_i \rho_h(\mathbf{x}_i, y)|] \\
&\quad (\text{merging the variables from same distribution})
\end{aligned}$$

Let $H_{\mathcal{X}}^{(\setminus y)} = \{\max(h_1, \dots) : h_j \in H_{\mathcal{X}}, j \in [1, \dots, k] \setminus y\}$, and we have $R_m^\parallel(H_{\mathcal{X}}^{(\setminus y)}) \leq (k-1)R_m^\parallel(H_{\mathcal{X}})$ from Ledoux and Talagrand [2013]'s conclusion similar to the empirical case.

Next, we can rewrite $\rho_h(\mathbf{x}_i, y_i)$ explicitly

$$\begin{aligned}
R_m^\parallel(\tilde{H}) &\leq \frac{2}{m} \sum_{y \in \mathcal{Y}} \mathbb{E}[\sup_{h \in H} \sum_{i=1}^m |\sigma_i (h(\mathbf{x}_i, y) - \max_{r \neq y} h(\mathbf{x}_i, r))|] \\
&\leq \sum_{y \in \mathcal{Y}} [\frac{2}{m} \mathbb{E}[\sup_{h \in H} \sum_{i=1}^m |\sigma_i h(\mathbf{x}_i, y)|] \\
&\quad + \frac{2}{m} \mathbb{E}[\sup_{h \in H} \sum_{i=1}^m |-\sigma_i \max_{r \neq y} h(\mathbf{x}_i, r)|]] \\
&= \sum_{y \in \mathcal{Y}} [\frac{2}{m} \mathbb{E}[\sup_{h \in H} \sum_{i=1}^m |\sigma_i h(\mathbf{x}_i, y)|] \\
&\quad + \frac{2}{m} \mathbb{E}[\sup_{h \in H} \sum_{i=1}^m |\sigma_i \max_{r \neq y} h(\mathbf{x}_i, r)|]] \\
&= \sum_{y \in \mathcal{Y}} [\frac{2}{m} \mathbb{E}[\sup_{h \in H_{\mathcal{X}}} \sum_{i=1}^m |\sigma_i h(\mathbf{x}_i)|] \\
&\quad + \frac{2}{m} \mathbb{E}[\sup_{h \in H_{\mathcal{X}}^{(\setminus y)}} \sum_{i=1}^m |\sigma_i \max_{r \neq y} h(\mathbf{x}_i)|]] \\
&\leq k [\frac{2k}{m} \mathbb{E}[\sup_{h \in H_{\mathcal{X}}} \sum_{i=1}^m |\sigma_i h(\mathbf{x}_i)|]] \\
&= k^2 R_m^\parallel(H_{\mathcal{X}})
\end{aligned}$$

That concludes the proof. \square

6 Analysis of the function $g(\cdot)$

Depending on the positive and negative property of $\mathbf{w}_y^\top \mathbf{x}$, our results are divided into several situations. For the sake of

simplifying, some shorthands are defined as

$$\begin{aligned}
R &\triangleq g(\mathbf{w}_R^\top \mathbf{x}), \\
r+ &\triangleq g(\text{any}(\mathbf{w}_{r+}^\top \mathbf{x})), \quad r- \triangleq g(\text{any}(\mathbf{w}_{r-}^\top \mathbf{x}))
\end{aligned}$$

If A: $\mathbf{w}_R^\top \mathbf{x} \geq -p$ with margin $p > 0$:

1.	$0 < \mathbf{w}_y^\top \mathbf{x} \leq \mathbf{w}_R^\top \mathbf{x} + p$	
	a) $2\mathbf{w}_y^\top \mathbf{x} \leq \mathbf{w}_R^\top \mathbf{x}$	$R > r+, \quad R ? r-$
	b) $\mathbf{w}_y^\top \mathbf{x} < \mathbf{w}_R^\top \mathbf{x} < 2\mathbf{w}_y^\top \mathbf{x}$	$R ? r+, \quad R > r-$
	c) $-p \leq \mathbf{w}_R^\top \mathbf{x} < \mathbf{w}_y^\top \mathbf{x}$	$R < g(\text{any}(\mathbf{w}_{r*}^\top \mathbf{x}))$
2.	$-p < \mathbf{w}_y^\top \mathbf{x} \leq 0 \leq \mathbf{w}_R^\top \mathbf{x} + p$	
	a) $\mathbf{w}_y^\top \mathbf{x} < 0 \leq \mathbf{w}_R^\top \mathbf{x}$	$R > r+, \quad R ? r-$
	b) $\mathbf{w}_y^\top \mathbf{x} \leq \mathbf{w}_R^\top \mathbf{x} < 0$	$\{\mathbf{w}_{r+}^\top \mathbf{x}\} = \emptyset, R ? r-$
	c) $-p \leq \mathbf{w}_R^\top \mathbf{x} < \mathbf{w}_y^\top \mathbf{x}$	$\{\mathbf{w}_{r+}^\top \mathbf{x}\} = \emptyset, R < r-$
3.	$\mathbf{w}_y^\top \mathbf{x} \leq -p < 0 \leq \mathbf{w}_R^\top \mathbf{x} + p$	
	a) $\mathbf{w}_R^\top \mathbf{x} \geq -p$	$\{\mathbf{w}_{r+}^\top \mathbf{x}\} = \emptyset, R ? r-$

If B: $\mathbf{w}_R^\top \mathbf{x} < -p$ with margin $p > 0$:

1.	$\mathbf{w}_y^\top \mathbf{x} < 0$, because $\mathbf{w}_y^\top \mathbf{x} \leq \mathbf{w}_R^\top \mathbf{x} + p < 0$	
	a) $\mathbf{w}_R^\top \mathbf{x} \leq \mathbf{w}_y^\top \mathbf{x}$	$\{\mathbf{w}_{r+}^\top \mathbf{x}\} = \emptyset, R ? r-$
	b) $\mathbf{w}_y^\top \mathbf{x} < \mathbf{w}_R^\top \mathbf{x} < -p$	$\{\mathbf{w}_{r+}^\top \mathbf{x}\} = \emptyset, R ? r-$

The Fig. 3 below gives the reference sketch used to analyse above conditions.

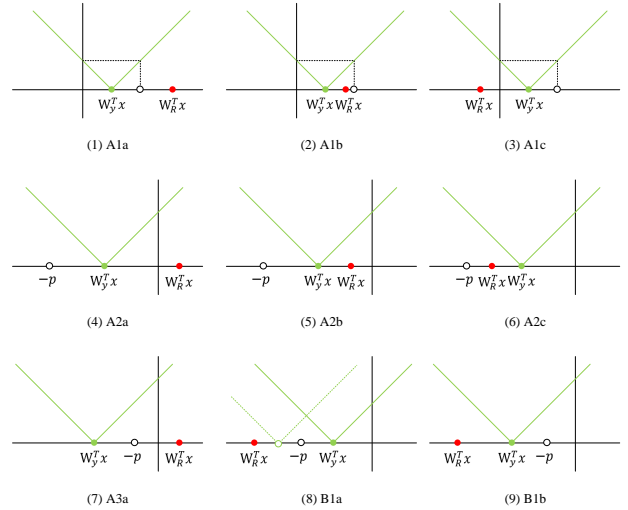


Figure 2: Reference sketch

7 Calculation of the gradient

The primal objective function is defined as

$$\begin{aligned}
P_i(\mathbf{W}) &= \frac{\lambda}{2} \|\mathbf{W}\|_2^2 + L(\mathbf{W}; (\mathbf{x}_i, \mathbf{y}_i)) \\
&\quad + \frac{\beta}{2} \log \text{tr}(\mathbf{W}^\top \mathbf{W}) - \frac{\beta \cdot \mathbb{I}}{2K} \log |\mathbf{W}^\top \mathbf{W}|.
\end{aligned}$$

Using SGD to solve this problem, we have to calculate the gradient of $P_i(\mathbf{W})$. Though the \mathbf{W} is a $n \times k$ (n denotes the feature vectors' dimension and k is the number of classes) matrix, the results of $P_i(\mathbf{W})$ applied with a matrix variable is a scoring $\in \mathbb{R}$. To our knowledge, there is a matrices expression of $\frac{\partial P_i(\mathbf{W})}{\partial \mathbf{W}}$ for a typical formulation.

To make it clearly, we define

$$\begin{aligned} \frac{\partial P_i(\mathbf{W})}{\partial \mathbf{W}} &= \frac{\lambda}{2} \frac{\partial \|\mathbf{W}\|_2^2}{\partial \mathbf{W}} + \frac{\partial L(\mathbf{W}; (\mathbf{x}_i, y_i))}{\partial \mathbf{W}} \\ &\quad + \frac{\beta}{2} \frac{\partial \log \text{tr}(\mathbf{W}^\top \mathbf{W})}{\partial \mathbf{W}} - \frac{\beta \cdot \mathbb{I}}{2K} \frac{\partial \log |\mathbf{W}^\top \mathbf{W}|}{\partial \mathbf{W}}. \\ &= \frac{\lambda}{2} \nabla_{L2} + \nabla_{Loss} + \frac{\beta}{2} \nabla_{logtr} - \frac{\beta \cdot \mathbb{I}}{2K} \nabla_{logdet} \end{aligned} \quad (30)$$

Then, we do the derivation partly.

(1) Firstly, we have

$$\begin{aligned} \nabla_{L2} &= \frac{\partial [\|\mathbf{W}\|_2^2]_{1 \times 1}}{\partial [\mathbf{W}]_{n \times k}} \\ &= \frac{\partial [\mathbf{1}^\top \mathbf{W}^\top \mathbf{W} \mathbf{1}]_{1 \times 1}}{\partial [\mathbf{W}]_{n \times k}} \\ &= \mathbf{W}(\mathbf{1}\mathbf{1}^\top + \mathbf{1}\mathbf{1}^\top) \\ &= [\mathbf{W}]_{n \times k} [(\mathbf{1}\mathbf{1}^\top + \mathbf{1}\mathbf{1}^\top)]_{k \times k} \\ &= \mathbf{W} \cdot 2\mathbf{I} \\ &= 2\mathbf{W} \end{aligned} \quad (31)$$

(2) Secondly, the ∇_{Loss} needs the updating rules with sub-gradient and

$$\nabla_{Loss} = \frac{1}{m} \sum_{i=1}^m \max(0, 1 + \mathbf{w}_{r_i}^\top \mathbf{x}_i - \mathbf{w}_{y_i}^\top \mathbf{x}_i). \quad (32)$$

- If hingloss Eq.(32) equals zero:

$$\nabla_{Loss}^j = 0,$$

- If hingloss Eq.(32) is above zero:

$$\nabla_{Loss}^j = \begin{cases} [-\mathbf{x}_i]_{n \times 1}, & \text{if } j = y_i \\ [\mathbf{x}_i]_{n \times 1}, & \text{if } j = R_i \\ 0, & \text{otherwise} \end{cases}$$

And all the $[\nabla_{Loss}^1 \dots \nabla_{Loss}^j \dots \nabla_{Loss}^k]$ make the ∇_{Loss} .

(3) Thirdly, we have

$$\begin{aligned} \nabla_{logtr} &= \frac{\partial [\log \text{tr}(\mathbf{W}^\top \mathbf{W})]_{1 \times 1}}{\partial [\mathbf{W}]_{n \times k}} \\ &= \frac{1}{\text{tr}(\mathbf{W}^\top \mathbf{W})} \cdot \frac{\partial [\text{tr}(\mathbf{W}^\top \mathbf{W})]_{1 \times 1}}{\partial [\mathbf{W}]_{n \times k}} \\ &= \frac{2\mathbf{W}}{\text{tr}(\mathbf{W}^\top \mathbf{W})} \end{aligned} \quad (33)$$

(4) Fourthly, the ∇_{logdet} term's results are divided based on the gradient direction we pursuit as

$$\nabla_{logdet} = \frac{\partial [\log |\mathbf{W}^\top \mathbf{W}|]_{1 \times 1}}{\partial [\mathbf{W}]_{n \times k}}. \quad (34)$$

If we regard each coordinate's partial direction $\mathbf{W}(i, j)$ as a whole to replace the gradient direction, we have

$$\begin{aligned} &\nabla_{logdet}(i, j) \\ &= \frac{\partial [\log |\mathbf{W}^\top \mathbf{W}|]_{1 \times 1}}{\partial [W_{ij}]_{n \times 1}} \\ &= \text{tr} \left(\frac{\partial \log |\mathbf{W}^\top \mathbf{W}|}{\partial \mathbf{W}^\top \mathbf{W}} \frac{\partial (\mathbf{W}^\top \mathbf{W})}{\partial W_{ij}} \right) \\ &= \text{tr} \left(((\mathbf{W}^\top \mathbf{W})^\top)^{-1} \frac{\partial (\mathbf{W}^\top \mathbf{W})}{\partial W_{ij}} \right) \\ &= \text{tr} \left((\mathbf{W}^\top \mathbf{W})^{-1} \frac{\partial \mathbf{W}^\top \mathbf{W}}{\partial W_{ij}} \right) \\ &= \text{tr} \left((\mathbf{W}^\top \mathbf{W})^{-1} \left(\frac{\partial \mathbf{W}^\top}{\partial W_{ij}} \mathbf{W} + \mathbf{W}^\top \frac{\partial \mathbf{W}}{\partial W_{ij}} \right) \right), \quad (35) \\ &= \sum_{p=1}^k (\mathbf{W}^\top \mathbf{W})_{pi}^{-1} \mathbf{W}_{jp} \\ &\quad + \sum_{p=1}^k (\mathbf{W}^\top \mathbf{W})_{jp}^{-1} \mathbf{W}_{pi}^\top \\ &= 2 \cdot \sum_{p=1}^k (\mathbf{W}^\top \mathbf{W})_{jp}^{-1} \mathbf{W}_{pi}^\top \\ &= 2((\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}^\top)_{ji} \end{aligned}$$

where $W_{ij} \in \mathbb{R}^{n \times k}$ which requires $n \leq k$ which is common in ordinary dataset and \mathbf{W}^+ denotes the pseudo inverse of \mathbf{W} . Thus we have the gradient

$$\nabla_{logdet} = 2\mathbf{W}(\mathbf{W}^\top \mathbf{W})^{-1}, \quad (36)$$

where the symmetric matrix $\mathbf{W}^\top \mathbf{W}$ is invertible.

References

- [Bartlett and Mendelson, 2002] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3(Nov):463–482, 2002.
- [Cortes *et al.*, 2013] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. Multi-class classification with maximum margin multiple kernel. In *ICML*, pages 46–54, 2013.
- [Ledoux and Talagrand, 2013] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.
- [Percy, 2015] Liang Percy. Lecture notes of statistical learning theory. <https://web.stanford.edu/class/cs229t/notes.pdf>, 2015.
- [Xie *et al.*, 2015a] Pengtao Xie, Yuntian Deng, and Eric Xing. Diversifying restricted boltzmann machine for document modeling. In *SIGKDD*, pages 1315–1324. ACM, 2015.
- [Xie *et al.*, 2015b] Pengtao Xie, Yuntian Deng, and Eric Xing. On the generalization error bounds of neural networks under diversity-inducing mutual angular regularization. *arXiv*, 2015.