A Summary of Proof of Theorem 1

Here, we summarized the main steps and ideas of our complete proof in order to help the readers understand the whole strategy.

1 A Summary of the Proof of Theorem 1(i)

The proof of Theorem 1(i) is consisted of five steps. The whole strategy is to relate our problem to the problem considered by Shivaswamy and Jebara (2010) (specifically, the error bound analysis on the function class of $\sum-$SVM, i.e., Definition 6 by Shivaswamy and Jebara (2010)) and follow their methods. Their key idea is to use a landmark set U, an i.i.d. sample of size n, to remove the dependence of the hypothesis class on the training data X. We need some definitions first. $D := \overline{D} := 1/2$ and $\hat{w}_l := d_l w_l$. We define $\mathcal{H}_2$ as

$$\mathcal{H}_2(B,\overline{E},X) := \Bigg\{ \left(x^l\right)_{l=1}^{m+1} \to \sum_{l=1}^{m+1} \left(\hat{w}_l + d_l v_l\right)^T x^l \Big| \ \hat{w}_l, v_l \in \mathcal{X}_l, 1 \le l \le m+1;\ \hat{w}_{m+1} = \mathbf{0};$$
$$(1.a)\ \|\boldsymbol{d}\|_p^p \le 1,\ \boldsymbol{d} \ge 0;\ (1.b)\ \frac{1}{2}\sum_{l=1}^{m+1} d_l \|v_l\|^2 \le B;\ (1.c)\ \frac{\overline{D}}{2}\sum_{l=1}^{m} \|\hat{w}_l\|^2 + \frac{D}{2n}\sum_{i=1}^{n}\Big(\sum_{l=1}^{m} \hat{w}_l^T x_i^l\Big)^2 \le \overline{E} \Bigg\} \tag{1}$$

Step 1: Associating $\mathcal{H}_1$ with $\mathcal{H}_2$. (For notation convenience, we switch the symbols v_{m+1} and w_{m+1} in the original $\mathcal{H}_1$.) This step is our major contribution in this proof while other steps are adapted from the approach by Shivaswamy and Jebara (2010) to our specific problem. This step provides the basics for later steps because the data-dependent constraint (1.c) in $\mathcal{H}_2$ is similar to the hypothesis class of $\sum-$SVM analyzed by Shivaswamy and Jebara (2010). We relax $\mathcal{H}_1$ to $\mathcal{H}_2$ by the following lemma (Recall the definition of $\mathcal{H}_1$ and E_1 in section 2.3 and section 2.4).

Lemma 1 *If $B > 0$, $C_l > 0$, $1 \le l \le m$, and the training set X is a random draw of n i.i.d. data, then* $\mathcal{H}_1(B, X, \boldsymbol{C_l}) \subseteq \mathcal{H}_2(B, E_1, X)$

Step 2: Introducing the landmark set. We define $\mathcal{H}_3$ relying on the landmark set U independent of the training data X as: for any $B, \overline{E} > 0$,

$$\mathcal{H}_3(B,\overline{E},U) := \Bigg\{ \left(x^l\right)_{l=1}^{m+1} \to \sum_{l=1}^{m+1} \left(\hat{w}_l + d_l v_l\right)^T x^l \Big| \hat{w}_l, v_l \in \mathcal{X}_l, 1 \le l \le m+1;\ \hat{w}_{m+1} = \mathbf{0};$$
$$(2.a)\ \|\boldsymbol{d}\|_p^p \le 1,\ \boldsymbol{d} \ge 0; (2.b)\ \frac{1}{2}\sum_{l=1}^{m+1} d_l \|v_l\|^2 \le B; (2.c)\ \frac{\overline{D}}{2}\sum_{l=1}^{m} \|\hat{w}_l\|^2 + \frac{D}{2n}\sum_{i=1}^{n}\Big(\sum_{l=1}^{m} \hat{w}_l^T u_i^l\Big)^2 \le \overline{E} \Bigg\} \tag{2}$$

Step 3: Obtaining a true risk bound for $\mathcal{H}_3$. Following Lanckriet et al. (2004), Theorem 11 and Theorem 15(i) from Shivaswamy and Jebara (2010), we can bound the empirical Rademacher complexity of $\mathcal{H}_3$ as: with probability at least $1-\delta$, for any positive $\overline{E}$ and B,

$$\widehat{\mathcal{R}}_n(\mathcal{H}_3(B,\overline{E},U)) \le \left(\frac{2\sqrt{2\overline{E}}}{n}\boldsymbol{E}_U[T(U,X)] + \frac{2m^{3/2}\sqrt{\ln(1/\delta)\overline{E}}}{\overline{D}n}\right) + \frac{\sqrt{2B\mathcal{C}(\mathcal{K}^+)}}{\sqrt{n}} \tag{3}$$

where $T(U,X)$ is defined in Theorem 1(i) and $\mathcal{C}(\mathcal{K}^+)$ is defined in section 2.4. Since $\mathcal{H}_3$ is independent of the training data, we can bound the true risk of functions from $\mathcal{H}_3$ with the empirical Rademacher complexity of $\mathcal{H}_3$.

Step 4: Associating $\mathcal{H}_2$ with $\mathcal{H}_3$. We have a corollary of Theorem 16 by Shivaswamy and Jebara (2010)

Corollary 1 *With probability at least $1-2\delta$, $\mathcal{H}_2(B, E_1, X) \subseteq \mathcal{H}_3(B, E'(E_1, 2), U)$.*

Step 5: Obtaining the true risk bound for $\mathcal{H}_1$. Using a union bound to combine Lemma 1, Corollary 1, inequality (3), and Theorem 14 from Shivaswamy and Jebara (2010), we finally obtain a true risk bound for functions from $\mathcal{H}_1$ in Theorem 1(i).

References

G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M.I. Jordan. Learning the Kernel Matrix with Semidefinite Programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

P.K. Shivaswamy, and T. Jebara. Maximum Relative Margin and Data-Dependent Regularization. *Journal of Machine Learning Research*, 11:747–788, 2010.