A Complete Proof of Theorem 1

1 Statement of Theorem 1

Let's recall some definitions and Theorem 1 from the submission.
$C_{min} := \min\{C_1, C_2, \ldots, C_m\}$.
$\mathcal{C}(\mathcal{K}^+) := \min\big(m+1, \|((\lambda_l)_{l=1}^{m+1}\|_q\big)$, where $\lambda_l, 1 \le l \le m+1$, is the maximum eigenvalue of K_l. When $p > 1$, $1/p + 1/q = 1$. When $p = 1$ we define $q = \infty$.
$E_1 := B/(2C_{min}) + m\big(B + m^{1/(2q)} n\sqrt{2B/C_{min}}\,\big)/(2n)$.
For any $\overline{E}, c > 0$, $E'(\overline{E}, c) := \overline{E} + \Big(4\sqrt{4m\overline{E}/n} + 6\sqrt{\ln(c/\delta)/2n}\Big)\Big(\overline{E}/2 + \overline{E}m/2\Big)$.

Theorem 1 *Fix $\gamma > 0$, and $C_l > 0$, $1 \le l \le m$. Let X be a training set of n i.i.d. data drawn from a distribution $\mathcal{P}$, and U be a landmark set of size n, for any $h \in \mathcal{H}_1(B, X, \boldsymbol{C_l})$ with $B > 0$:*

(i) With probability at least $1-\delta$ over a random draw of X, we have

$$\Pr_{\mathcal{P}}\Big[y^{m+1} \neq \operatorname{sign}\Big(h\big((x^l)_{l=1}^{m+1}\big)\Big)\Big] \le \frac{1}{n\gamma}\sum_{i=1}^{n}\xi_i^{m+1} + 3\sqrt{\frac{\ln(8/\delta)}{2n}} + \frac{8m^{\frac{3}{2}}\sqrt{\ln(4/\delta)E'(E_1,8)}}{\gamma n} +$$
$$\frac{2\sqrt{2B\mathcal{C}(\mathcal{K}^+)}}{\gamma\sqrt{n}} + \frac{4\sqrt{2E'(E_1,8)}}{\gamma n}\boldsymbol{E}_U\underbrace{\left[\sum_{i=1}^{n}\big((x_i^l)_{l=1}^m\big)^T\Big(\frac{1}{2}\boldsymbol{I} + \frac{1}{2n}\sum_{j=1}^{n}\big((u_j^l)_{l=1}^m\big)\big((u_j^l)_{l=1}^m\big)^T\Big)^{-1}\big((x_i^l)_{l=1}^m\big)\right]^{\frac{1}{2}}}_{:=T(U,X)}$$

(ii) With probability at least $1-\delta$ over a random draw of X, we have

$$\Pr_{\mathcal{P}}\Big[y^{m+1} \neq \operatorname{sign}\Big(h\big((x^l)_{l=1}^{m+1}\big)\Big)\Big] \le \frac{1}{n\gamma}\sum_{i=1}^{n}\xi_i^{m+1} + 3\sqrt{\frac{\ln(8/\delta)}{2n}} + \frac{8m^{\frac{3}{2}}\sqrt{\ln(4/\delta)E'(E_1,8)}}{\gamma n}$$
$$+ \frac{2\sqrt{2B(m+1)}}{\gamma\sqrt{n}} + \frac{4\sqrt{mE'(E_1,8)}}{\gamma\sqrt{n}}$$

where $\xi_i^{m+1} = \max\Big(0, \gamma - y_i^{m+1}h\big((x_i^l)_{l=1}^{m+1}\big)\Big)$ are the so-called slack variables.

2 A Complete Proof of Theorem 1(i)

Here we present the complete of Theorem 1(i). The true risk in this proof always refers to the true risk on the main task. The goal here is to follow approaches proposed by Shivaswamy and Jebara (2010) and Lanckriet et al. (2004) to derive a bound on the true risk that can guide the development of a practical algorithm. We want to understand how the parameter B in $\mathcal{H}_1(B, X, \boldsymbol{C_l})$ influences the true risk. The obtained bound may not be the tightest possible but it is informative, i.e., the complexity term in the bound will vanish when the number of training data goes to infinite.

The proof of Theorem 1(i) is consisted of five steps. The whole strategy is to relate our problem to the problem considered by 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 then we can use the technique of landmark set introduced by Shivaswamy and Jebara (2010) to overcome the difficulty of data-dependent regularization.

In this proof, Theorem 2, Theorem 3, Theorem 4, and Theorem 5 are existing results, or are trivial adaptations of existing results to fit our settings. These four theorems are not new contributions of this proof.

Recall that the definition of $\mathcal{H}_1$ in the submitted paper is: for any positive B, C_l, $1 \le l \le m$, and a training set

X of n i.i.d. data, (NOTE: for notation convenience, we switch the symbols v_{m+1} and w_{m+1} in the original $\mathcal{H}_1$.)

$$\mathcal{H}_1(B, X, \boldsymbol{C_l}) := \Bigg\{ \left(x^l\right)_{l=1}^{m+1} \to \sum_{l=1}^{m+1} \left(d_l(w_l + v_l)\right)^T x^l \Big| w_l, v_l \in \mathcal{X}_l, 1 \le l \le m+1;\ w_{m+1} = \mathbf{0};$$
$$(1.a)\ \|\boldsymbol{d}\|_p^p \le 1,\ \boldsymbol{d} \ge 0;\ \text{and}$$
$$\underbrace{\frac{1}{2}\sum_{l=1}^{m+1} d_l \|v_l\|^2}_{(1.b)} + \underbrace{\frac{1}{2}\sum_{l=1}^{m} C_l d_l \|w_l\|^2}_{(1.c)} + \underbrace{\frac{1}{2}\sum_{l=1}^{m}\sum_{i=1}^{n} d_l (w_l^T x_i^l - y_i^l)^2}_{(1.d)} \le B \Bigg\} \tag{1}$$

We first relax $\mathcal{H}_1$ to a more convenient form for the purpose of error bound analysis. Let's define $\mathcal{H}_{1'}$ as: for any positive B, C_l, $1 \le l \le m$, and a training set X of n i.i.d. data

$$\mathcal{H}_{1'}(B, X, \boldsymbol{C_l}) := \Bigg\{ \left(x^l\right)_{l=1}^{m+1} \to \sum_{l=1}^{m+1} \left(d_l(w_l + v_l)\right)^T x^l \Big| \tag{2}$$
$$w_l, v_l \in \mathcal{X}_l, 1 \le l \le m+1;\ w_{m+1} = \mathbf{0},$$
$$(2.a)\ \|\boldsymbol{d}\|_p^p \le 1,\ \boldsymbol{d} \ge 0;$$
$$(2.b)\ \frac{1}{2}\sum_{l=1}^{m+1} d_l \|v_l\|^2 \le B;$$
$$(2.c)\ \frac{1}{2}\sum_{l=1}^{m} C_l d_l \|w_l\|^2 \le B;$$
$$(2.d)\ \frac{1}{2}\sum_{l=1}^{m}\sum_{i=1}^{n} d_l (w_l^T x_i^l - y_i^l)^2 \le B \Bigg\}$$

We can further relax $\mathcal{H}_{1'}$ by eliminating the variable d_l's appearing in (2.c) and (2.d), and converting these two constraints to the form of the hypothesis class considered by Shivaswamy and Jebara (2010) (Definition 6 (Shivaswamy and Jebara, 2010)). Let $D := \overline{D} := 1/2$[1], $\hat{w}_l := d_l w_l$. In Step 1 of our proof, we will relax $\mathcal{H}_{1'}$ to $\mathcal{H}_2$ which is defined as: for any positive B, $\overline{E}$, and a training set X of n i.i.d. data,

$$\mathcal{H}_2(B, \overline{E}, X) := \Bigg\{ \left(x^l\right)_{l=1}^{m+1} \to \sum_{l=1}^{m+1} \left(\hat{w}_l + d_l v_l\right)^T x^l \Big| \tag{3}$$
$$\hat{w}_l, v_l \in \mathcal{X}_l, 1 \le l \le m+1;\ \hat{w}_{m+1} = \mathbf{0},$$
$$(3.a)\ \|\boldsymbol{d}\|_p^p \le 1,\ \boldsymbol{d} \ge 0;$$
$$(3.b)\ \frac{1}{2}\sum_{l=1}^{m+1} d_l \|v_l\|^2 \le B;$$
$$(3.c)\ \frac{\overline{D}}{2}\sum_{l=1}^{m} \|\hat{w}_l\|^2 + \frac{D}{2n}\sum_{i=1}^{n}\Big(\sum_{l=1}^{m} \hat{w}_l^T x_i^l\Big)^2 \le \overline{E} \Bigg\}$$

Step 1: Associating $\mathcal{H}_1$ with $\mathcal{H}_2$.

Firstly, the relation between $\mathcal{H}_1$ and $\mathcal{H}_{1'}$ can be obviously stated as:

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

Proof. For any $(\boldsymbol{d}, \boldsymbol{w}, \boldsymbol{v}) \in \mathcal{H}_1(B, X, \boldsymbol{C_l})$, (1.a) implies (2.a). The constraint for (1.b),(1.c), and (1.d) implies (2.b), (2.c), and (2.d). Thus, $(\boldsymbol{d}, \boldsymbol{w}, \boldsymbol{v}) \in \mathcal{H}_{1'}(B, X, \boldsymbol{C_l})$. ∎

[1] In the paper by Shivaswamy and Jebara (2010), $\overline{D} := 1 - D$ and $0 < D < 1$, which is more general. In our model, we fix $\overline{D}$ and D to be 1/2.

In the following, we show that the relation between $\mathcal{H}_{1'}$ and $\mathcal{H}_2$ is

Lemma 2 *If $B > 0$, $C_l > 0$, $1 \le l \le m$, and X is a random draw of n training data, then $\mathcal{H}_{1'}(B, X, \boldsymbol{C_l}) \subseteq \mathcal{H}_2(B, E_1, X)$*

Before proving Lemma 2, we need the following

Lemma 3 *Given fixed positive B, C_l with $1 \le l \le m$ and an i.i.d. training sample X of size n, for any $(\boldsymbol{d}, \boldsymbol{w}, \boldsymbol{v}) \in \mathcal{H}_{1'}$, we have $\left|\sum_{l=1}^{m}\sum_{i=1}^{n} d_l y_i^l w_l^T x_i^l\right| \le n m^{1/(2q)}\sqrt{2B/C_{min}}$.*

Proof.

$$\left|\sum_{l=1}^{m}\sum_{i=1}^{n} d_l y_i^l w_l^T x_i^l\right| = \left|\sum_{l=1}^{m}\sqrt{C_l d_l} w_l^T \sum_{i=1}^{n}\sqrt{\frac{d_l}{C_l}} y_i^l x_i^l\right|$$

$$= \left|\left\langle \left(\sqrt{C_1 d_1} w_1, \sqrt{C_2 d_2} w_2, \ldots, \sqrt{C_m d_m} w_m\right), \left(\sum_{i=1}^{n}\sqrt{\frac{d_1}{C_1}} y_i^1 x_i^1, \sum_{i=1}^{n}\sqrt{\frac{d_2}{C_2}} y_i^2 x_i^2, \ldots, \sum_{i=1}^{n}\sqrt{\frac{d_m}{C_m}} y_i^m x_i^m\right)\right\rangle\right|$$

$$\le \left\|\left(\sqrt{C_1 d_1} w_1, \sqrt{C_2 d_2} w_2, \ldots, \sqrt{C_m d_m} w_m\right)\right\|_2 \cdot \left\|\left(\sum_{i=1}^{n}\sqrt{\frac{d_1}{C_1}} y_i^1 x_i^1, \sum_{i=1}^{n}\sqrt{\frac{d_2}{C_2}} y_i^2 x_i^2, \ldots, \sum_{i=1}^{n}\sqrt{\frac{d_m}{C_m}} y_i^m x_i^m\right)\right\|_2$$

(by the Cauchy − Schwarz inequality)

$$= \sqrt{\sum_{l=1}^{m} C_l d_l \|w_l\|^2}\sqrt{\sum_{l=1}^{m}\frac{d_l}{C_l}\left\|\sum_{i=1}^{n} y_i^l x_i^l\right\|_2^2} = \sqrt{\sum_{l=1}^{m} C_l d_l \|w_l\|^2}\sqrt{\sum_{l=1}^{m}\frac{d_l}{C_l}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i^l y_j^l \left(x_i^l\right)^T x_j^l}$$

$$= \sqrt{\sum_{l=1}^{m} C_l d_l \|w_l\|^2}\sqrt{\sum_{l=1}^{m}\frac{d_l}{C_l}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i^l y_j^l K_l(i,j)} \le \sqrt{\sum_{l=1}^{m} C_l d_l \|w_l\|^2}\sqrt{\sum_{l=1}^{m}\frac{d_l}{C_l}(\boldsymbol{y}^l)^T K_l \boldsymbol{y}^l} \quad (*)$$

$$\le \sqrt{2B}\sqrt{\sum_{l=1}^{m}\frac{d_l}{C_{min}} n\lambda_l} \quad \text{(by (2.c) and } \sup_{\|z\|^2\le 1} z^T K_l z = \lambda_l, \text{when } K_l \text{ is s.p.d..)}$$

$$\le \sqrt{2Bn\|\boldsymbol{d}\|_p \|(\lambda_l)_{l=1}^m\|_q / C_{min}} \quad \text{(by Hölder's inequality)}$$

$$\le \sqrt{2Bn\|(\lambda_l)_{l=1}^m\|_q / C_{min}} \quad \text{(by (2.a))} \tag{4}$$

$$\le n m^{1/(2q)}\sqrt{2B/C_{min}} \quad \text{(by } \lambda_l \le Tr(K_l) = n.)$$

∎

We are now ready to prove Lemma 2.

Proof. (For Lemma 2.) Consider any $(\boldsymbol{d}, \boldsymbol{w}, \boldsymbol{v}) \in \mathcal{H}_{1'}(B, X, \boldsymbol{C_l})$. It suffices to verify that $(\boldsymbol{d}, \hat{\boldsymbol{w}}, \boldsymbol{v})$ satisfies (3.c) when $\overline{E} = E_1$.

Since $0 \le d_l \le 1$, $C_l > 0$, $\hat{w}_l = d_l w_l$, and $d_l = 0$ implies $\hat{w}_l = \mathbf{0}$, we have

$$\frac{1}{2}\sum_{l=1}^{m} C_l d_l \|w_l\|^2 = \frac{1}{2}\sum_{l=1}^{m}\frac{C_l}{d_l}\|\hat{w}_l\|^2$$

$$\ge \frac{C_{min}}{2}\sum_{l=1}^{m}\frac{1}{d_l}\|\hat{w}_l\|^2 \ge \frac{C_{min}}{2}\sum_{l=1}^{m}\|\hat{w}_l\|^2$$

Thus, (2.c) implies that

$$\frac{\overline{D}}{2}\sum_{l=1}^{m}\|\hat{w}_l\|^2 \le \frac{B}{2C_{min}} \tag{5}$$

(Recall that $\overline{D} = 1/2$.)

By Lemma 3, we have

$$\begin{aligned}
&\frac{1}{2}\sum_{l=1}^{m}\sum_{i=1}^{n} d_l(w_l^T x_i^l - y_i^l)^2 = \frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{m} d_l(w_l^T x_i^l - y_i^l)^2 \\
&\geq \frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{m} d_l(w_l^T x_i^l)^2 - \Big|\sum_{l=1}^{m}\sum_{i=1}^{n} d_l y_i^l w_l^T x_i^l\Big| \\
&\geq \frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{m} \frac{1}{d_l}(\hat{w}_l^T x_i^l)^2 - nm^{1/(2q)}\sqrt{2B/C_{min}} \\
&\geq \frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{m} (\hat{w}_l^T x_i^l)^2 - nm^{1/(2q)}\sqrt{2B/C_{min}} \\
&\geq \frac{1}{2m}\sum_{i=1}^{n}\Big(\sum_{l=1}^{m} \hat{w}_l^T x_i^l\Big)^2 - nm^{1/(2q)}\sqrt{2B/C_{min}}
\end{aligned}$$

Hence, (2.d) implies that

$$\frac{D}{2n}\sum_{i=1}^{n}\Big(\sum_{l=1}^{m} \hat{w}_l^T x_i^l\Big)^2 \leq \frac{m}{2n}\Big(B + nm^{1/(2q)}\sqrt{2B/C_{min}}\Big) \tag{6}$$

(Recall that $D = 1/2$.)

(5) and (6) together imply (3.c). Therefore, $(\boldsymbol{d}, \hat{\boldsymbol{w}}, \boldsymbol{v}) \in \mathcal{H}_2(B, E_1, X)$. ∎

We can see that (3.c) has the same form as the hypothesis class considered by Shivaswamy and Jebara (2010) (i.e., $\sum$−SVM) because we can view the classifier in (3.c) as a vector concatenating $\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_m$ and each datum as a vector concatenating $x^1, x^2, \ldots, x^m$. Hence, we can use the results from Shivaswamy and Jebara (2010). Here we fix D and $\overline{D}$ to be 1/2. The definition of D and $\overline{D}$ by Shivaswamy and Jebara (2010) is more general than the definition here. The reason of using D and $\overline{D}$ here instead of just using the values 1/2 is to explicitly show the similarity between (3.c) and the hypothesis class of $\sum$−SVM (Shivaswamy and Jebara 2010).

Step 2: Introducing the landmark set.

In $\mathcal{H}_2$, (3.c) is dependent on the training data and this creates a difficulty to derive a bound on the true risk by the empirical Rademacher complexity. Shivaswamy and Jebara (2010) developed a method by using the so-called *landmark set* to overcome this difficulty. For our specific problem, the key idea is to eliminate the dependence of (3.c) on the training data by replacing the training data appearing in (3.c) with the so-called landmark variables, which are i.i.d. variables drawn from the same distribution $\mathcal{P}$ as the training data. After the introduction of the landmark variables, we can obtain a hypothesis class independent of the training data, which can be considered as fixed before observing the training data, and the usual method to derive error bounds on the true risk using the empirical Rademacher complexity can be applied.

Let's use $U_i = \Big(\big(u_i^l\big)_{l=1}^{m+1}, \big(\bar{y}_i^l\big)_{l=1}^{m+1}\Big)$, $1 \leq i \leq n$ to denote the i-th landmark variable corresponding to $X_i = \Big(\big(x_i^l\big)_{l=1}^{m+1}, \big(y_i^l\big)_{l=1}^{m+1}\Big)$, $1 \leq i \leq n$, which is the i-th training data. Each of the n landmark variables is drawn i.i.d. from the distribution $\mathcal{P}$, which is the same as the training set. U denotes a set of landmark variables, i.e., the *landmark set*, and X represents the training set. To replace the training data in (3.c) with the landmark

variables, we define $\mathcal{H}_3$ as: for any positive B, $\overline{E}$, and a landmark set U of n i.i.d. data

$$\mathcal{H}_3(B,\overline{E},U)=\Bigg\{\big(x^l\big)_{l=1}^{m+1}\to\sum_{l=1}^{m+1}\big(\hat{w}_l+d_lv_l\big)^Tx^l\Big| \tag{7}$$
$$\hat{w}_l,v_l\in\mathcal{X}_l,1\le l\le m+1;\ \hat{w}_{m+1}=\mathbf{0},$$
$$(7.a)\ \|\boldsymbol{d}\|_p^p\le 1,\ \boldsymbol{d}\ge 0;$$
$$(7.b)\ \frac{1}{2}\sum_{l=1}^{m+1}d_l\|v_l\|^2\le B;$$
$$(7.c)\ \frac{\overline{D}}{2}\sum_{l=1}^{m}\|\hat{w}_l\|^2+\frac{D}{2n}\sum_{i=1}^{n}\big(\sum_{l=1}^{m}\hat{w}_l^Tu_i^l\big)^2\le\overline{E}\Bigg\}$$

$\mathcal{H}_3$ is completely independent of the training data X and therefore it can be considered as fixed before observing the training data, and standard arguments to derive true risk bounds based on the Rademacher complexity (or its empirical version) can be applied to $\mathcal{H}_3$.

Step 3: Obtaining a true risk bound for $\mathcal{H}_3$.

Let $\widehat{\mathcal{R}}_n(\mathcal{H})$ denotes the empirical Rademacher complexity (Bartlett and Mendelson, 2002) of a function class $\mathcal{H}$. We restate a previous result stated by Bartlett and Mendelson (2002) and Shivaswamy and Jebara (2010), and we adapt the theorem to our setting in the following way:

Theorem 2 [**(Bartlett and Mendelson, 2002) and (Shivaswamy and Jebara, 2010)**] *Fix $\gamma>0$. Let $\mathcal{F}$ be the class of functions from $\mathcal{X}\times\{\pm1\}\to\mathbb{R}$ given by $f\big((x^l)_{l=1}^{m+1},y^{m+1}\big)=-y^{m+1}h\big((x^l)_{l=1}^{m+1}\big)$ for any $h\in\mathcal{H}$. Let $X_i=\Big((x_i^l)_{l=1}^{m+1},(y_i^l)_{l=1}^{m+1}\Big)$, $1\le i\le n$, be drawn i.i.d. from a probability distribution $\mathcal{P}$. Then, with probability at least $1-\delta$ over the samples of size n, the following bound holds:*

$$\Pr_{\mathcal{P}}\Big[y^{m+1}\ne\operatorname{sign}\Big(h\big((x^l)_{l=1}^{m+1}\big)\Big)\Big]\le\frac{1}{n\gamma}\sum_{i=1}^{n}\xi_i^{m+1}+\frac{2}{\gamma}\widehat{\mathcal{R}}_n(\mathcal{F})+3\sqrt{\frac{\ln(2/\delta)}{2n}} \tag{8}$$

where $\xi_i^{m+1}=\max\Big(0,\gamma-y_i^{m+1}h\big((x_i^l)_{l=1}^{m+1}\big)\Big)$ are the so-called slack variables.

Notice that $\widehat{\mathcal{R}}_n(\mathcal{F})=\widehat{\mathcal{R}}_n(\mathcal{H})$ for $y^{m+1}\in\{\pm1\}$.

By Theorem 2, to obtain a bound on the true risk for any function from $\mathcal{H}_3$, we just need to get a bound on $\widehat{\mathcal{R}}_n(\mathcal{H}_3)$. To bound $\widehat{\mathcal{R}}_n(\mathcal{H}_3)$, we can use Theorem 11 and Theorem 15(i) proposed in the paper by Shivaswamy and Jebara (2010). We first restate these two results by adapting them to our setting. Before introducing these two theorems by Shivaswamy and Jebara (2010), we need some definitions:

$$\mathcal{G}_{\overline{E},D}^{U}:=\Bigg\{\big(x^l\big)_{l=1}^{m+1}\to\sum_{l=1}^{m}\hat{w}_l^Tx^l\Big|\hat{w}_l\in\mathcal{X}_l,1\le l\le m; \tag{9}$$
$$(9.a)\ \frac{\overline{D}}{2}\sum_{l=1}^{m}\|\hat{w}_l\|^2+\frac{D}{2n}\sum_{i=1}^{n}\big(\sum_{l=1}^{m}\hat{w}_l^Tu_i^l\big)^2\le\overline{E}\Bigg\}$$

We also define a matrix K' with its (i,j) element $K'(i,j):=\langle\big((x_i^l)_{l=1}^m\big),(x_j^l)_{l=1}^m\big)\rangle$. We define R as an upper bound on the norm of $\big((x_i^l)_{l=1}^m\big)$. The superscripts here are m and this is because in the definition of $\mathcal{G}_{\overline{E},D}^{U}$ (Eq. 9), x^{m+1} does not have any effect.

Let's restate Theorem 11 in the paper of Shivaswamy and Jebara (2010) in our setting:

Theorem 3 [**Shivaswamy and Jebara, 2010**] $\widehat{\mathcal{R}}_n(\mathcal{G}_{\overline{E},D}^{U})\le T_1(U,X)$, *where for any training set $\mathcal{B}$ and landmark set $\mathcal{A}$,*

$$T_1(\mathcal{A},\mathcal{B}):=\frac{2\sqrt{2\overline{E}}}{|\mathcal{B}|}\Bigg(\sum_{(x^l)_{l=1}^{m+1}\in\mathcal{B}}\big((x^l)_{l=1}^m\big)^T\Big(\overline{D}\boldsymbol{I}+\frac{D}{|\mathcal{A}|}\sum_{(u^l)_{l=1}^{m+1}\in\mathcal{A}}\big((u^l)_{l=1}^m\big)\big((u^l)_{l=1}^m\big)^T\Big)^{-1}\big((x^l)_{l=1}^m\big)\Bigg)^{\frac{1}{2}}$$

and Theorem 15(i) in their paper:

Theorem 4 [**Shivaswamy and Jebara, 2010**] *With probability at least* $1-\delta$,

$$T_1(U,S) \leq \boldsymbol{E}_U[T_1(U,S)] + \frac{2R^4\sqrt{\ln(1/\delta)\overline{E}}}{\overline{D}\sqrt{n}\sqrt{tr(K')}} \tag{10}$$

Recall that we assume that $\langle x, x\rangle = 1$, for any $x \in \mathcal{X}_l$ and for all $1 \leq l \leq m+1$. Thus, $tr(K') = mn$ and $R^4 = m^2$.

Before bounding $\widehat{\mathcal{R}}_n(\mathcal{H}_3)$, we need one more lemma as follows: Let $\sup_{\mathcal{H}}$ represent the supremum over all classifiers from $\mathcal{H}$.

Lemma 4 *Given training data* X, $\boldsymbol{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathcal{H}_3}\left|\sum_i^n \sigma_i \sum_l^{m+1} d_l v_l^T x_i^l\right|\right] \leq \sqrt{2Bn\mathcal{C}(\mathcal{K}^+)}$, *where* $\boldsymbol{\sigma}$ *is the Rademacher random variable.*

Proof. We can bound the LHS (left hand side) following the proof of Lemma 3. We just need to modify the proof of Lemma 3 in the following way: (a). every C_l equals 1; (b). replacing m by $m+1$ and w by v; (c). $y_i^l = \sigma_i$, and $\boldsymbol{y}^l = (\sigma_i)_{i=1}^n = \boldsymbol{\sigma}$. Thus, inequality (4) becomes

$$\left|\sum_i^n \sigma_i \sum_l^{m+1} d_l v_l^T x_i^l\right| \leq \sqrt{2Bn\|(\lambda_l)_{l=1}^{m+1}\|_q} \tag{11}$$

We can also bound the LHS following the method proposed by Lanckriet et al. (2004). We still use the proof of Lemma 3 as above until the inequality (*) in the proof of Lemma 3. Then, we obtain another bound for the LHS by

$$\begin{aligned}
&\boldsymbol{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathcal{H}_3}\left|\sum_i^n \sigma_i \sum_l^{m+1} d_l v_l^T x_i^l\right|\right] \\
&\leq \sqrt{2B}\boldsymbol{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathcal{H}_3}\sqrt{\sum_{l=1}^{m+1} d_l \boldsymbol{\sigma}^T \boldsymbol{K}_l \boldsymbol{\sigma}}\right] \quad \text{(by inequality } (*) \text{ in the proof of Lemma 3)} \\
&\leq \sqrt{2B}\boldsymbol{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathcal{H}_3}\sqrt{\sum_{l=1}^{m+1} \boldsymbol{\sigma}^T \boldsymbol{K}_l \boldsymbol{\sigma}}\right] = \sqrt{2B}\boldsymbol{E}_{\boldsymbol{\sigma}}\left[\sqrt{\sum_{l=1}^{m+1} \boldsymbol{\sigma}^T \boldsymbol{K}_l \boldsymbol{\sigma}}\right] \\
&\leq \sqrt{2B}\left(\sum_{l=1}^{m+1} \boldsymbol{E}_{\boldsymbol{\sigma}}\left[\boldsymbol{\sigma}^T \boldsymbol{K}_l \boldsymbol{\sigma}\right]\right)^{1/2} = \sqrt{2B}\left(\sum_{l=1}^{m+1} \boldsymbol{E}_{\boldsymbol{\sigma}}\Big[\sum_{i=1}^n\sum_{j=1}^n \sigma_i\sigma_j K_l(i,j)\Big]\right)^{1/2} \\
&= \sqrt{2B}\left(\sum_{l=1}^{m+1} \boldsymbol{E}_{\boldsymbol{\sigma}}\Big[\sum_{i=1}^n \sigma_i^2 K_l(i,i)\Big]\right)^{1/2} = \sqrt{2B(m+1)n} \quad \text{(recall that } K_l(i,i)=1)
\end{aligned} \tag{12}$$

Combining inequalities (11) and (12) yields the bound stated in the lemma. ∎

We are now ready to derive a bound for $\widehat{\mathcal{R}}_n(\mathcal{H}_3)$. With probability at least $1-\delta$, for any positive $\overline{E}$ and B, we

have

$$\begin{aligned}
\widehat{\mathcal{R}}_n(\mathcal{H}_3(B,\overline{E},U)) =& \boldsymbol{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathcal{H}_3}\left|\frac{2}{n}\sum_i^n \sigma_i \sum_{l=1}^{m+1}\left(\hat{w}_l + d_l v_l\right)^T x_i^l\right|\right] \\
\leq& \frac{2}{n}\boldsymbol{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathcal{H}_3}\left|\sum_i^n \sigma_i \sum_l^m \hat{w}_l^T x_i^l\right|\right] \\
&+ \frac{2}{n}\boldsymbol{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathcal{H}_3}\left|\sum_i^n \sigma_i \sum_l^{m+1} d_l v_l^T x_i^l\right|\right] \\
\leq& \left(\boldsymbol{E}_U[T_1(U,X)] + \frac{2m^{3/2}\sqrt{\ln(1/\delta)\overline{E}}}{\overline{D}n}\right) \text{ (by Theorem 2 and Theorem 3)} \\
&+ \frac{\sqrt{2B\mathcal{C}(\mathcal{K}^+)}}{\sqrt{n}} \quad \text{(by Lemma 4)} \\
\leq& \left(\frac{2\sqrt{2\overline{E}}}{n}\boldsymbol{E}_U[T(U,X)] + \frac{2m^{3/2}\sqrt{\ln(1/\delta)\overline{E}}}{\overline{D}n}\right) + \frac{\sqrt{2B\mathcal{C}(\mathcal{K}^+)}}{\sqrt{n}} \qquad (13)
\end{aligned}$$

The last inequality comes from the relation between the definitions of $T(U,X)$ in Theorem 1(i) and $T_1(U,X)$ in Theorem 3. Plugging inequality (13) into Theorem 2, we obtain a bound on the true risk when learning with the hypothesis class $\mathcal{H}_3$. But, note that $\mathcal{H}_3$ is defined on the landmark variables. Our original problem is not learning with $\mathcal{H}_3$, and we will eliminate the landmark variables in Step 4.

Step 4: Associating $\mathcal{H}_2$ with $\mathcal{H}_3$.

Let

$$E_1'(\overline{E},c) := \overline{E} + \left(4R\sqrt{\frac{2\overline{E}}{nD}} + 6\sqrt{\frac{\ln(c/\delta)}{2n}}\right)\left(\frac{\overline{E}}{2} + \frac{D\overline{E}R^2}{2\overline{D}}\right) \qquad (14)$$

Recall that $R^2 = m$ and $D = 1/2$. So $E_1'(\overline{E},c) = E'(\overline{E},c)$. Again, the reason we use E_1' here is to explicitly show the connection to the function class considered by Shivaswamy and Jebara (2010).

Let

$$\begin{aligned}
\mathcal{G}_{\overline{E},D}^X := \Bigg\{ & \left(x^l\right)_{l=1}^{m+1} \to \sum_{l=1}^m \hat{w}_l^T x^l \Big| \hat{w}_l \in \mathcal{X}_l, 1 \leq l \leq m; \qquad (15) \\
& (15.a)\ \frac{\overline{D}}{2}\sum_{l=1}^m \|\hat{w}_l\|^2 + \frac{D}{2n}\sum_{i=1}^n \left(\sum_{l=1}^m \hat{w}_l^T x_i^l\right)^2 \leq \overline{E} \Bigg\}
\end{aligned}$$

where X is the training data. So, $\mathcal{G}_{\overline{E},D}^X$ is dependent on the training data while in $\mathcal{G}_{\overline{E},D}^U$ (see Eq. 9) the training data is replaced by the landmark variables.

We need to restate Theorem 16 from Shivaswamy and Jebara (2010) as follows:

Theorem 5 [**Shivaswamy and Jebara 2010**] *With probability at least* $1-2\delta$*, for any* $\overline{E} > 0$*,* $\mathcal{G}_{\overline{E},D}^X \subseteq \mathcal{G}_{E_1'(\overline{E},2),D}^U$.

We obtain a corollary of Theorem 5

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

Proof. Consider any $(\boldsymbol{d},\hat{\boldsymbol{w}},\boldsymbol{v}) \in \mathcal{H}_2(B,E_1,X)$. $\boldsymbol{d}$ and $\boldsymbol{v}$ satisfy (7.a) and (7.b). Thus, to determine whether $(\boldsymbol{d},\hat{\boldsymbol{w}},\boldsymbol{v}) \in \mathcal{H}_3(B,E'(E_1,2),X)$ holds, it is equivalent to check whether $\hat{\boldsymbol{w}}$ satisfies (7.c). By Theorem 5, we know that the probability that $\hat{\boldsymbol{w}}$ satisfies (7.c) is at least $1-2\delta$ and this concludes the proof. ∎

Step 5: Obtaining the true risk bound for $\mathcal{H}_1$. By Lemma 1, Lemma 2, and Corollary 1 we can easily obtain the following corollary:

Corollary 2 *If $B > 0$, $C_l > 0$, $1 \leq l \leq m$, X is a random draw of n i.i.d. training data from the distribution $\mathcal{P}$, and U is a landmark set of n i.i.d. data drawn from the same distribution $\mathcal{P}$, with probability at least $1 - 2\delta$, $\mathcal{H}_1(B, X, \boldsymbol{C_l}) \subseteq \mathcal{H}_3(B, E'(E_1, 2), U)$.*

By the set-inclusion result in Corollary 2, we can derive the final risk bound on the data-dependent function class $\mathcal{H}_1$ by the true risk bound of the data-independent function class $\mathcal{H}_3$.

Using a union bound to combine Theorem 2, inequality (13), and Corollary 2, we finally obtain the bound in Theorem 1(i). Notice that we replace δ with $\delta/4$ and this is because the bound in Theorem 2 holds with probability at least $1 - \delta$, inequality (13) holds with probability at least $1 - \delta$, and the set-inclusion in Corollary 2 holds with probability at least $1 - 2\delta$.

3 A Complete Proof of Theorem 1(ii)

The goal here is to bound the RHS of the bound in Theorem 1(i) with data-independent terms. To achieve this, we need two inequalities.

Firstly, we can easily see that

$$\mathcal{C}(K^+) \leq m + 1 \tag{16}$$

Secondly let's bound the term $\boldsymbol{E}_U[T(U, X)]$ in the bound of Theorem 1(i). Letting the dimension of $\left(x_i^l\right)_{l=1}^m$ be $\bar{d}$, we have $\bar{x}^T(\boldsymbol{I} + (1/n)\sum_{i=1}^n \bar{u}_i\bar{u}_i^T)^{-1}\bar{x} \leq m\bar{\lambda}$, for any $\bar{x}, \bar{u}_i \in \mathbb{R}^{\bar{d}}$, $\bar{x}^T\bar{x} = m$, $\bar{u}_i^T\bar{u}_i = m$, where $\bar{\lambda}$ denotes the largest eigenvalue of the s.p.d. matrix $(\boldsymbol{I} + (1/n)\sum_{i=1}^n \bar{u}_i\bar{u}_i^T)^{-1}$. Letting z be the smallest eigenvalue of the symmetric positive semidefinite matrix $(1/n)\sum_{i=1}^n \bar{u}_i\bar{u}_i^T$, we have $\bar{\lambda} = 1/(1+z)$, and $z \geq 0$. Therefore $\bar{\lambda} \leq 1$. Thus,

$$\begin{aligned}
&\boldsymbol{E}_U\left[\sum_{i=1}^n \left(\left(x_i^l\right)_{l=1}^m\right)^T \left(\frac{1}{2}\boldsymbol{I} + \frac{1}{2n}\sum_{j=1}^n \left(\left(u_j^l\right)_{l=1}^m\right)\left(\left(u_j^l\right)_{l=1}^m\right)^T\right)^{-1} \left(\left(x_i^l\right)_{l=1}^m\right)\right]^{\frac{1}{2}} \\
&\leq \left(\sum_{i=1}^n \boldsymbol{E}_U\left[\left(\left(x_i^l\right)_{l=1}^m\right)^T \left(\frac{1}{2}\boldsymbol{I} + \frac{1}{2n}\sum_{j=1}^n \left(\left(u_j^l\right)_{l=1}^m\right)\left(\left(u_j^l\right)_{l=1}^m\right)^T\right)^{-1} \left(\left(x_i^l\right)_{l=1}^m\right)\right]\right)^{\frac{1}{2}} \\
&\text{(by convexity)} \\
&\leq \left(\sum_{i=1}^n \frac{1}{2}m\bar{\lambda}\right)^{\frac{1}{2}} \leq \sqrt{\frac{1}{2}mn}
\end{aligned} \tag{17}$$

Combining inequalities (16) and (17), we obtain the bound in Theorem 1(ii), which is independent of the training data X.

We finally point out that the techniques used by Srebro and Ben-david (2006) and Cortes et al. (2010) for error bound analysis of MKL can be applied to handle the terms (1.b) and (1.c) of $\mathcal{H}_1$ and to improve the fourth term in the bounds of Theorem 1(i),(ii). However, the major difficulty here is how to handle the data-dependent term (1.d) of $\mathcal{H}_1$. In addition, the approach by Srebro and Ben-david (2006) is based on the covering number and it is not very convenient to incorporate that into the proof here.

References

P. Bartlett, and S. Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 3:463–482, 2002.

C. Cortes, M. Mohri, and A. Rostamizadeh. Generalization Bounds for Learning Kernels. In *Proceedings of the 27th International Conference on Machine Learning*, ACM, 2010.

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.

N. Srebro and S. Ben-david. Learning bounds for support vector machines with learned kernels. In *Annual Conference On Learning Theory (COLT)*, pp. 169–183, 2006.