Supplementary Material to "Sparse Feature Selection in Kernel Discriminant Analysis via Optimal Scoring"

Alexander F. Lapanowski **Irina Gaynanova**
Texas A&M University
{alapanow, irinag}@stat.tamu.edu

Abstract

This supplement contains the derivation of projection formula (6), proofs of Theorems 1 and 2, as well as proofs of supplementary Theorems and Lemmas.

S1 Derivation of projection formula (6)

Proof. Since $\widehat{f} = \sum_{i=1}^{n} \widehat{\alpha}_i[\Phi(x_i) - \overline{\Phi}]$,

$$
\begin{aligned}
\left\langle \Phi(x) - \overline{\Phi}, \widehat{f} \right\rangle_{\mathcal{H}} &= \left\langle \Phi(x) - \overline{\Phi}\,, \sum_{i=1}^{n} \widehat{\alpha}_i[\Phi(x_i) - \overline{\Phi}] \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{n} \widehat{\alpha}_i \left\langle \Phi(x) - \overline{\Phi}, \Phi(x_i) - \overline{\Phi} \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{n} \widehat{\alpha}_i \left\langle \Phi(x), \Phi(x_i) \right\rangle_{\mathcal{H}} - \sum_{i=1}^{n} \widehat{\alpha}_i \left\langle \Phi(x), \overline{\Phi} \right\rangle_{\mathcal{H}} - \sum_{i=1}^{n} \widehat{\alpha}_i \left\langle \overline{\Phi}, \Phi(x_i) \right\rangle_{\mathcal{H}} + \sum_{i=1}^{n} \widehat{\alpha}_i \left\langle \overline{\Phi}, \overline{\Phi} \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{n} \widehat{\alpha}_i k(x, x_i) - (\mathbf{1}^\top \widehat{\alpha}) \frac{1}{n} \sum_{i=1}^{n} k(x, x_i) - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \widehat{\alpha}_i k(x_j, x_i) + (\mathbf{1}^\top \widehat{\alpha}) \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(x_i, x_j).
\end{aligned}
$$

Let $K(X, x) := \begin{pmatrix} k(x_1, x) & \cdots & k(x_n, x) \end{pmatrix}^\top$. Then from the above display

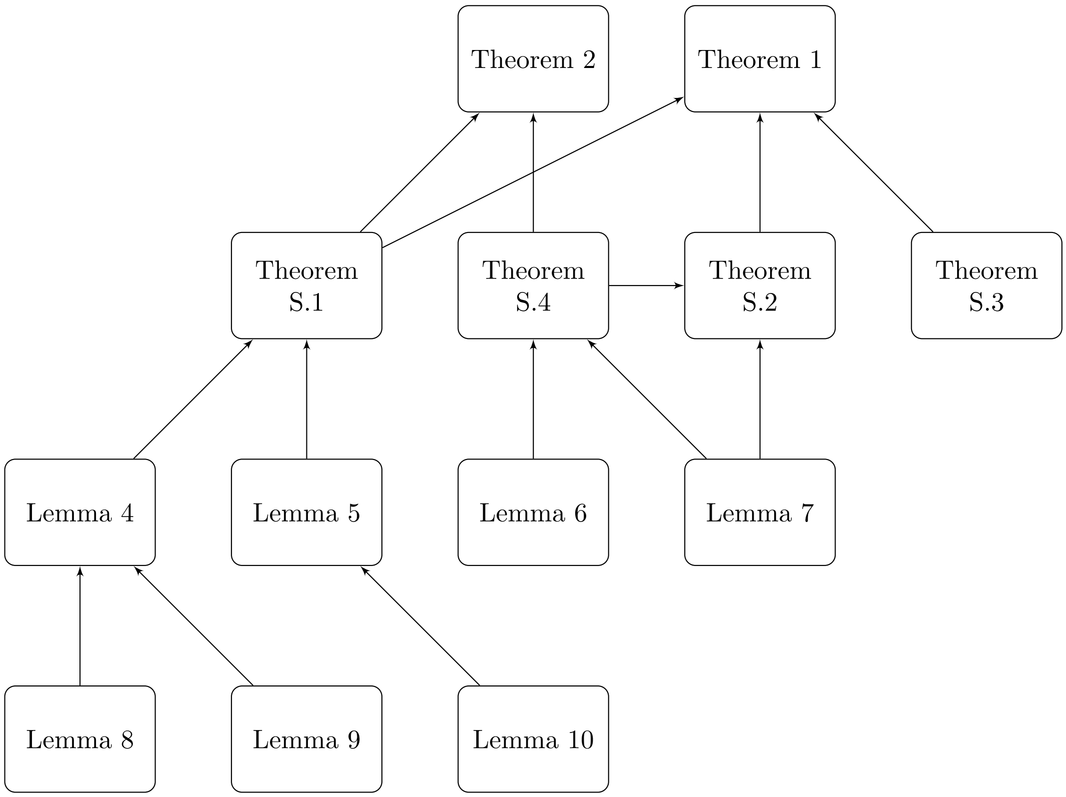
$$
\begin{aligned}
\left\langle \Phi(x) - \overline{\Phi}, \widehat{f} \right\rangle_{\mathcal{H}} &= K(X, x)^\top \widehat{\alpha} - n^{-1} K(X, x)^\top \mathbf{1}\mathbf{1}^\top \widehat{\alpha} - n^{-1} \mathbf{1}^\top K \widehat{\alpha} + \frac{1}{n^2} \mathbf{1}^\top K \mathbf{1} (\mathbf{1}^\top \widehat{\alpha}) \\
&= K(X, x)^\top C \widehat{\alpha} - \frac{1}{n} \mathbf{1}^\top K C \widehat{\alpha} \\
&= (K(X, x)^\top - \frac{1}{n} \mathbf{1}^\top K) C \widehat{\alpha},
\end{aligned}
$$

where $C = I - n^{-1}\mathbf{1}\mathbf{1}^\top$ is the centering matrix. □

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

S2 Technical Proofs

In this section we prove the results stated within the main text. We use C, C_1, C_2, ... to denote absolute positive constants that do not depend on the sample size n but which may depend on $\|\theta^*\|_\infty, \kappa$, or τ. Their values may change from line to line. The dependence between the main Theorems and supplementary results is depicted below.



S.2.1 Proofs of Theorems 1 and 2

Proof of Theorem 1. Consider

$$R(\widehat{f},\widehat{\beta}) - R(f^*,\beta^*) = \underbrace{R(\widehat{f},\widehat{\beta}) - \widetilde{R}_{\text{emp}}(\widehat{f},\widehat{\beta})}_{I_1} + \underbrace{\widetilde{R}_{\text{emp}}(\widehat{f},\widehat{\beta}) - \widetilde{R}_{\text{emp}}(\widetilde{f},\widetilde{\beta})}_{I_2} + \underbrace{\widetilde{R}_{\text{emp}}(\widetilde{f},\widetilde{\beta}) - R(f^*,\beta^*)}_{I_3}.$$

By the union bound and de Morgan's law,

$$\mathbb{P}\Big(R(\widehat{f},\widehat{\beta}) - R(f^*,\beta^*) > \varepsilon\Big) \le \mathbb{P}\Big(I_1 > \frac{\varepsilon}{3}\Big) + \mathbb{P}\Big(I_2 > \frac{\varepsilon}{3}\Big) + \mathbb{P}\Big(I_3 > \frac{\varepsilon}{3}\Big).$$

Applying Theorems S.1, S.2 and S.3 to I_1, I_2 and I_3 correspondingly, there exist constants $C, C_i > 0$ such that

$$\begin{aligned}
&\mathbb{P}\Big(R(\widehat{f},\widehat{\beta}) - R(f^*,\beta^*) > \varepsilon\Big) \\
&\qquad \le 2\mathcal{N}_\varepsilon \exp\Big(-\frac{n\varepsilon^2}{128(\|\theta^*\|_\infty + \kappa\tau)^4}\Big) + C_2\exp\Big(-\frac{C_3 n\varepsilon^2}{1+(\kappa\tau)^2}\Big) + 2\exp\Big(-\frac{n\varepsilon^2}{16(\|\theta^*\|_\infty + \kappa\tau)^4}\Big) \\
&\qquad \le C_4\mathcal{N}_\varepsilon \exp\Big(-\frac{C_5 n\varepsilon^2}{(\|\theta^*\|_\infty + \kappa\tau)^4}\Big),
\end{aligned}$$

where $\mathcal{N}_\varepsilon = \{1 + 2(\|\theta^*\|_\infty + \kappa\tau)/\varepsilon\}\exp(C\tau^2\varepsilon^{-2})$. This concludes the proof of Theorem 1.

□

Proof of Theorem 2. Consider

$$R(\widehat{f},\widehat{\beta}) - R_{\text{emp}}(\widehat{f}) = \underbrace{R(\widehat{f},\widehat{\beta}) - \widetilde{R}_{\text{emp}}(\widehat{f},\widehat{\beta})}_{I_1} + \underbrace{\widetilde{R}_{\text{emp}}(\widehat{f},\widehat{\beta}) - R_{\text{emp}}(\widehat{f})}_{I_2}.$$

By the union bound and de Morgan's law,

$$\mathbb{P}\Big(R(\widehat{f},\widehat{\beta}) - R_{\text{emp}}(\widehat{f}) > \varepsilon\Big) \leq \mathbb{P}\Big(I_1 > \frac{\varepsilon}{2}\Big) + \mathbb{P}\Big(I_2 > \frac{\varepsilon}{2}\Big).$$

Applying Theorem S.1 for I_1 and Theorem S.4 for I_2, the exist constants $C_i > 0$ such that

$$\begin{aligned}\mathbb{P}\Big(R(\widehat{f},\widehat{\beta}) - R_{\text{emp}}(\widehat{f}) > \varepsilon\Big) &\leq 2\mathcal{N}_\varepsilon \exp\Big(-\frac{n\varepsilon^2}{128(\|\theta^*\|_\infty + \kappa\tau)^4}\Big) + C_3 \exp\Big(-\frac{C_4 n\varepsilon^2}{1+(\kappa\tau)^2}\Big)\\ &\leq C_5\mathcal{N}_\varepsilon \exp\Big(-\frac{C_6 n\varepsilon^2}{(\|\theta^*\|_\infty + \kappa\tau)^4}\Big),\end{aligned}$$

where $\mathcal{N}_\varepsilon = \{1 + 2(\|\theta^*\|_\infty + \kappa\tau)/\varepsilon\}\exp(C_1\tau^2\varepsilon^{-2})$. This concludes the proof of Theorem 2. □

S.2.2 Supplementary Theorems

Theorem S.1. *Under Assumptions 1-3, there exists a constant $C_2 > 0$ such that for all $\varepsilon > 0$,*

$$\mathbb{P}\Big(\sup_{f\in\mathcal{H}_\tau,\,\beta\in I_\tau}\{R(f,\beta) - \widetilde{R}_{emp}(f,\beta)\} > \varepsilon\Big) \leq 2\mathcal{N}_\varepsilon \exp\Big(-\frac{n\varepsilon^2}{128(\|\theta^*\|_\infty + \kappa\tau)^4}\Big),$$

where $\mathcal{N}_\varepsilon = \{1 + 2(\|\theta^*\|_\infty + \kappa\tau)/\varepsilon\}\exp(C_2\tau^2\varepsilon^{-2})$.

Theorem S.2. *Let $\widehat{\beta} = -\Big\langle \overline{\Phi}, \widehat{f}\Big\rangle_{\mathcal{H}}$. Under Assumptions 1 and 2, there exist constants $C_1, C_2 > 0$ such that for all $\varepsilon > 0$,*

$$\mathbb{P}\Big(\Big|\widetilde{R}_{emp}(\widehat{f},\widehat{\beta}) - \widetilde{R}_{emp}(\widetilde{f},\widetilde{\beta})\Big| > \varepsilon\Big) \leq C_1 \exp\Big(-\frac{C_2 n\varepsilon^2}{1+(\kappa\tau)^2}\Big).$$

Theorem S.3. *Under Assumptions 1 and 2, for all $\varepsilon > 0$*

$$\mathbb{P}\Big(\widetilde{R}_{emp}(\widetilde{f},\widetilde{\beta}) - R(f^*,\beta^*) > \varepsilon\Big) \leq 2\exp\Big(-\frac{n\varepsilon^2}{16(\|\theta^*\|_\infty + \kappa\tau)^4}\Big).$$

Theorem S.4. *Let Assumptions 1 and 2 be true, and let $\beta(f) := n^{-1}\sum_{i=1}^n y_i^\top\theta^* - \langle\overline{\Phi}, f\rangle_{\mathcal{H}} = \overline{Y\theta^*} - \langle\overline{\Phi}, f\rangle_{\mathcal{H}}$ be the minimizing $\beta \in I_\tau$ for fixed $f \in \mathcal{H}_\tau$ in the modified empirical risk. There exists constants $C_1, C_2 > 0$ such that for all $\varepsilon > 0$*

$$\mathbb{P}\Big(\sup_{f\in\mathcal{H}_\tau}|R_{emp}(f) - \widetilde{R}_{emp}(f,\beta(f))| > \varepsilon\Big) \leq C_1\exp\Big(-\frac{C_2 n\varepsilon^2}{1+(\kappa\tau)^2}\Big).$$

Definition 1. *The* empirical measure *T_x with respect to $\{x_i\}_{i=1}^n$ is defined as $T_x := n^{-1}\sum_{i=1}^n \delta(x_i)$, where $\delta(x_i)$ is the point mass at x_i. The space $L^2(T_x)$ is the set $\mathcal{H}_\tau$ equipped with the semi-norm*

$$\|f\|_{L^2(T_x)} := \sqrt{\frac{1}{n}\sum_{i=1}^n |f(x_i)|^2} = \sqrt{\frac{1}{n}\sum_{i=1}^n |\langle\Phi(x_i), f\rangle_{\mathcal{H}}|^2}.$$

Definition 2. *Let (X,d) be a pseudometric space. An ε-net is any subset $\widetilde{X} \subset X$ such that for any $x \in X$, there exists a $\widetilde{x} \in \widetilde{X}$ satisfying $d(x,\widetilde{x}) < \varepsilon$. The* ε-covering number *of (X,d) is the minimum size of an ε-net for X.*

Remark 1. *Distances in $\mathcal{H}_\tau$ are given by the semi-norm generated by $L^2(T_x)$. Distances in I_τ are given by the Euclidean distance $d(\beta_1,\beta_2) = |\beta_1 - \beta_2|$.*

S.2.3 Proofs of Supplementary Theorems

Proof of Theorem S.1. Let $\{(x_j, y_j)\}_{j=n+1}^{2n}$ be independent from $\{(x_i, y_i)\}_{i=1}^{n}$ and identically distributed set of n pairs, and let T_x be the empirical measure on $\{(x_i, y_i)\}_{i=1}^{2n}$. Let $\widetilde{R}_{\text{emp}}(f, \beta)$ be the modified empirical risk on $\{(x_i, y_i)\}_{i=1}^{n}$, and $\widetilde{R}'_{\text{emp}}(f, \beta)$ on $\{(x_j, y_j)\}_{i=n+1}^{2n}$. By symmetrization lemma (see, for example, Lemma 2 in [1]), for $n\varepsilon^2 \geq 2$

$$\mathbb{P}\Big(\sup_{f\in\mathcal{H}_\tau,\, \beta\in I_\tau}\{R(f,\beta) - \widetilde{R}_{\text{emp}}(f,\beta)\} > \varepsilon\Big) \leq 2\mathbb{P}\Big(\sup_{f\in\mathcal{H}_\tau,\, \beta\in I_\tau}\{\widetilde{R}'_{\text{emp}}(f,\beta) - \widetilde{R}_{\text{emp}}(f,\beta)\} > \frac{\varepsilon}{2}\Big).$$

Let $c = 64(\|\theta^*\|_\infty + \kappa\tau)$, and let $\{f_1, \ldots, f_M\}$ be the smallest $L^2(T_x)$ $\varepsilon/\sqrt{2}c$-net of $\mathcal{H}_\tau$ and $\{\beta_1, \ldots, \beta_K\}$ an ε/c-net of I_τ. Applying Lemma 4 to the above display

$$\mathbb{P}\Big(\sup_{f\in\mathcal{H}_\tau,\, \beta\in I_\tau}\{R(f,\beta) - \widetilde{R}_{\text{emp}}(f,\beta)\} > \varepsilon\Big) \leq 2\mathbb{P}\Big(\underset{\substack{f\in\{f_1,\ldots,f_M\}\\ \beta\in\{\beta_1,\ldots,\beta_K\}}}{\text{maximize}}\{\widetilde{R}'_{\text{emp}}(f,\beta) - \widetilde{R}_{\text{emp}}(f,\beta)\} > \frac{\varepsilon}{4}\Big).$$

Applying Lemma 5 to the right-hand expression gives the final inequality

$$\mathbb{P}\Big(\sup_{f\in\mathcal{H}_\tau,\, \beta\in I_\tau}\{R(f,\beta) - \widetilde{R}_{\text{emp}}(f,\beta)\} > \varepsilon\Big) \leq 2\{1 + 2(\|\theta^*\|_\infty + \kappa\tau)/\varepsilon\}\exp\Big(\frac{C_1\tau^2}{\varepsilon^2}\Big)\exp\Big(-\frac{n\varepsilon^2}{128(\|\theta^*\|_\infty + \kappa\tau)^4}\Big).$$

This completes the proof of Theorem S.1. □

Proof of Theorem S.2. Let $\beta(f) = \overline{Y\theta^*} - \langle\overline{\Phi}, f\rangle_{\mathcal{H}}$. By definition of $\widetilde{f}$, $\widetilde{\beta} = \beta(\widetilde{f})$, $\widetilde{R}_{\text{emp}}(\widehat{f}, \widehat{\beta}) \geq \widetilde{R}_{\text{emp}}(\widetilde{f}, \widetilde{\beta})$. On the other hand, since $R_{\text{emp}}(\widehat{f}) \leq R_{\text{emp}}(\widetilde{f})$,

$$\begin{aligned}
\widetilde{R}_{\text{emp}}(\widehat{f},\widehat{\beta}) - \widetilde{R}_{\text{emp}}(\widetilde{f},\widetilde{\beta}) &= \widetilde{R}_{\text{emp}}(\widehat{f},\widehat{\beta}) - R_{\text{emp}}(\widehat{f}) + R_{\text{emp}}(\widehat{f}) - R_{\text{emp}}(\widetilde{f}) + R_{\text{emp}}(\widetilde{f}) - \widetilde{R}_{\text{emp}}(\widetilde{f},\widetilde{\beta})\\
&\leq \widetilde{R}_{\text{emp}}(\widehat{f},\widehat{\beta}) - R_{\text{emp}}(\widehat{f}) + R_{\text{emp}}(\widetilde{f}) - \widetilde{R}_{\text{emp}}(\widetilde{f},\widetilde{\beta})\\
&\leq \widetilde{R}_{\text{emp}}(\widehat{f},\widehat{\beta}) - \widetilde{R}_{\text{emp}}(\widehat{f},\beta(\widehat{f})) + \widetilde{R}_{\text{emp}}(\widehat{f},\beta(\widehat{f})) - R_{\text{emp}}(\widehat{f}) + R_{\text{emp}}(\widetilde{f}) - \widetilde{R}_{\text{emp}}(\widetilde{f},\widetilde{\beta})\\
&\leq \underbrace{\Big|\widetilde{R}_{\text{emp}}(\widehat{f},\widehat{\beta}) - \widetilde{R}_{\text{emp}}(\widehat{f},\beta(\widehat{f}))\Big|}_{I_1} + \underbrace{2\sup_{f\in\mathcal{H}_\tau}\Big|R_{\text{emp}}(f) - \widetilde{R}_{\text{emp}}(f,\beta(f))\Big|}_{I_2}.
\end{aligned}$$

The union bound and de Morgan's law proves

$$\mathbb{P}\Big(\widetilde{R}_{\text{emp}}(\widehat{f},\widehat{\beta}) - \widetilde{R}_{\text{emp}}(\widetilde{f},\widetilde{\beta}) > \varepsilon\Big) \leq \mathbb{P}\Big(I_1 > \frac{\varepsilon}{2}\Big) + \mathbb{P}\Big(I_2 > \frac{\varepsilon}{2}\Big).$$

Consider I_1

$$\begin{aligned}
&\Big|\widetilde{R}_{\text{emp}}(\widehat{f},\widehat{\beta}) - \widetilde{R}_{\text{emp}}(\widehat{f},\beta(\widehat{f}))\Big|\\
&= \Big|\frac{1}{n}\sum_{i=1}^{n}\Big(y_i^\top\theta^* - \Big\langle\Phi(x_i) - \overline{\Phi}\,,\,\widehat{f}\Big\rangle_{\mathcal{H}}\Big)^2 - \frac{1}{n}\sum_{i=1}^{n}\Big(y_i^\top\theta^* - \overline{Y\theta^*} - \Big\langle\Phi(x_i) - \overline{\Phi}\,,\,\widehat{f}\Big\rangle_{\mathcal{H}}\Big)^2\Big|\\
&= \Big|2\frac{1}{n}\sum_{i=1}^{n}\overline{Y\theta^*}\Big(y_i^\top\theta^* - \Big\langle\Phi(x_i) - \overline{\Phi}, \widehat{f}\Big\rangle_{\mathcal{H}}\Big) - \frac{1}{n}\sum_{i=1}^{n}(\overline{Y\theta^*})^2\Big|\\
&= \Big|(\overline{Y\theta^*})^2 - 2(\overline{Y\theta^*})\frac{1}{n}\sum_{i=1}^{n}\Big\langle\Phi(x_i) - \overline{\Phi}, \widehat{f}\Big\rangle_{\mathcal{H}}\Big|\\
&= |\overline{Y\theta^*}|^2.
\end{aligned}$$

By Lemma 7, there exists $C_1 > 0$ such that $\mathbb{P}(I_1 > \varepsilon/2) \leq 2\exp(-C_1 n\varepsilon)$ for all $\varepsilon > 0$. By Theorem S.4, there exists constants $C_2, C_3 > 0$ such that $\mathbb{P}(I_2 > \varepsilon/2) \leq C_2\exp[-C_3(n\varepsilon^2)/\{1 + (\kappa\tau)^2\}]$. Combining the bounds for

I_1 and I_2 gives

$$\mathbb{P}\Big(\widetilde{R}_{\text{emp}}(\widehat{f},\widehat{\beta}) - \widetilde{R}_{\text{emp}}(\widetilde{f},\widetilde{\beta}) > \varepsilon\Big) \leq 2\exp(-C_1 n\varepsilon) + C_2 \exp\Big(-\frac{C_3 n\varepsilon^2}{1+(\kappa\tau)^2}\Big) \leq C_4 \exp\Big(-\frac{C_5 n\varepsilon^2}{1+(\kappa\tau)^2}\Big)$$

for some constants $C_i > 0$. This completes the proof of Theorem S.2. □

Proof of Theorem S.3. Consider

$$\begin{aligned}\widetilde{R}_{\text{emp}}(\widetilde{f},\widetilde{\beta}) - R(f^*,\beta^*) &= \widetilde{R}_{\text{emp}}(\widetilde{f},\widetilde{\beta}) - \widetilde{R}_{emp}(f^*,\beta^*) + \widetilde{R}_{emp}(f^*,\beta^*) - R(f^*,\beta^*)\\ &\leq \widetilde{R}_{emp}(f^*,\beta^*) - R(f^*,\beta^*),\end{aligned}$$

where the last inequality follows since $\widetilde{R}_{\text{emp}}(\widetilde{f},\widetilde{\beta}) \leq \widetilde{R}_{emp}(f^*,\beta^*)$ by the definition of $\widetilde{f}$, $\widetilde{\beta}$.

Let $z_i := |y_i^\top\theta^* - \beta^* - \langle\Phi(x_i), f^*\rangle_{\mathcal{H}}|^2$, then $\widetilde{R}_{\text{emp}}(f^*,\beta^*) = n^{-1}\sum_{i=1}^n z_i$ is the average of i.i.d. random variables with $\mathbb{E}z_i = R(f^*,\beta^*)$ by definition of expected risk. Since $|z_i| \leq 4(\|\theta^*\|_\infty + \kappa\tau)^2$, by Hoeffding's inequality

$$\mathbb{P}(|\widetilde{R}_{\text{emp}}(f^*,\beta^*) - R(f^*,\beta^*)| > \varepsilon) = \mathbb{P}\Big(\Big|n^{-1}\sum_{i=1}^n (z_i - \mathbb{E}z_i)\Big| > \varepsilon\Big) \leq 2\exp\Big(-\frac{n\varepsilon^2}{16(\|\theta^*\|_\infty + \kappa\tau)^4}\Big).$$

□

Proof of Theorem S.4. By definition of $R_{\text{emp}}(f)$ and $\widetilde{R}_{\text{emp}}(f,\beta(f))$,

$$\begin{aligned}R_{\text{emp}}(f) - \widetilde{R}_{\text{emp}}(f,\beta(f)) &= \frac{1}{n}\sum_{i=1}^n |y_i^\top\widehat{\theta} - \big\langle\Phi(x_i) - \overline{\Phi}, f\big\rangle_{\mathcal{H}}|^2 - \frac{1}{n}\sum_{i=1}^n |y_i^\top\theta^* - \beta(f) - \langle\Phi(x_i), f\rangle_{\mathcal{H}}|^2\\ &= \frac{1}{n}\sum_{i=1}^n |y_i^\top\widehat{\theta} - \big\langle\Phi(x_i) - \overline{\Phi}, f\big\rangle_{\mathcal{H}}|^2 - \frac{1}{n}\sum_{i=1}^n |y_i^\top\theta^* - \overline{Y\theta^*} - \big\langle\Phi(x_i) - \overline{\Phi}, f\big\rangle_{\mathcal{H}}|^2.\end{aligned}$$

Expanding the squares and cancelling equal terms yields

$$\begin{aligned}&R_{\text{emp}}(f) - \widetilde{R}_{\text{emp}}(f,\beta(f))\\ &= \frac{1}{n}\sum_{i=1}^n\Big\{(y_i^\top\widehat{\theta})^2 - (y_i^\top\theta^*)^2 - 2y_i^\top(\widehat{\theta}-\theta^*)\big\langle\Phi(x_i)-\overline{\Phi}, f\big\rangle_{\mathcal{H}} - 2\overline{Y\theta^*}\big\langle\Phi(x_i)-\overline{\Phi}, f\big\rangle_{\mathcal{H}} + 2y_i^\top\theta^*\overline{Y\theta^*} - (\overline{Y\theta^*})^2\Big\}\\ &= \frac{1}{n}\sum_{i=1}^n\Big\{(y_i^\top\widehat{\theta})^2 - (y_i^\top\theta^*)^2\Big\} - \frac{1}{n}\sum_{i=1}^n\Big\{2y_i^\top(\widehat{\theta}-\theta^*)\big\langle\Phi(x_i)-\overline{\Phi}, f\big\rangle_{\mathcal{H}}\Big\} + (\overline{Y\theta^*})^2\\ &= I_1 + I_2(f) + I_3,\end{aligned}$$

where I_1 and I_3 are independent of f. By the union bound and de Morgan's law,

$$\mathbb{P}\Big(\sup_{f\in\mathcal{H}_\tau}|R_{\text{emp}}(f) - \widetilde{R}_{\text{emp}}(f,\beta(f))| > \varepsilon\Big) \leq \mathbb{P}\Big(|I_1| > \frac{\varepsilon}{3}\Big) + \mathbb{P}\Big(\sup_{f\in\mathcal{H}_\tau}|I_2(f)| > \frac{\varepsilon}{3}\Big) + \mathbb{P}\Big(|I_3| > \frac{\varepsilon}{3}\Big).$$

We bound each probability separately. Since $y_i \in \mathbb{R}^2$ is an indicator vector of class membership for sample i, using the definition of $\widehat{\theta}$ and θ^*

$$|I_1| = \Big|\frac{1}{n}\sum\Big\{(y_i^\top\widehat{\theta})^2 - (y_i^\top\theta^*)^2\Big\}\Big| \leq \max_i |(y_i^\top\widehat{\theta})^2 - (y_i^\top\theta^*)^2| = \max\Big(|n_1/n_2 - \pi_1/\pi_2|, |n_2/n_1 - \pi_2/\pi_1|\Big).$$

By Lemma 6, there exist $C_1, C_2 > 0$ such that $\mathbb{P}(|I_1| > \varepsilon/3) \leq C_1\exp(-C_2 n\,\varepsilon^2)$.

By Hólder's and Cauchy-Schwarz inequalities

$$\begin{aligned}
|I_2(f)| &= \left|\frac{1}{n}\sum_{i=1}^n 2y_i^\top(\widehat{\theta}-\theta^*)\left\langle\Phi(x_i)-\overline{\Phi},f\right\rangle_{\mathcal{H}}\right| \\
&\le \frac{1}{n}\sum_{i=1}^n 2|y_i^\top(\widehat{\theta}-\theta^*)|\cdot|\left\langle\Phi(x_i)-\overline{\Phi},f\right\rangle_{\mathcal{H}}| \\
&\le 2\|\widehat{\theta}-\theta^*\|_\infty \max_i |\left\langle\Phi(x_i)-\overline{\Phi},f\right\rangle_{\mathcal{H}}| \\
&\le 2\max\Big(|\sqrt{n_1/n_2}-\sqrt{\pi_1/\pi_2}|,|\sqrt{n_2/n_1}-\sqrt{\pi_2/\pi_1}|\Big)\max_i\|\Phi(x_i)-\overline{\Phi}\|_{\mathcal{H}}\,\|f\|_{\mathcal{H}} \\
&\le 4\max\Big(|\sqrt{n_1/n_2}-\sqrt{\pi_1/\pi_2}|,|\sqrt{n_2/n_1}-\sqrt{\pi_2/\pi_1}|\Big)\kappa\tau,
\end{aligned}$$

where we used Assumption 2 in the last inequality. Since the upper bound does not depend on f, the same bound holds for $\sup_{f\in\mathcal{H}_\tau}|I_2(f)|$. Combining the bound with Lemma 6 gives for some $C_3, C_4 > 0$

$$\mathbb{P}\Big(\sup_{f\in\mathcal{H}_\tau}|I_2(f)|>\varepsilon\Big)\le\mathbb{P}\Bigg(\max\Big(|\sqrt{n_1/n_2}-\sqrt{\pi_1/\pi_2}|,|\sqrt{n_2/n_1}-\sqrt{\pi_2/\pi_1}|>\frac{\varepsilon}{4\kappa\tau}\Big)\le C_3\exp(-C_4\frac{n\varepsilon^2}{(\kappa\tau)^2}).$$

By Lemma 7, there exists $C_5 > 0$ such that $\mathbb{P}(|I_3| > \varepsilon/3) \le 2\exp(-C_5 n\,\varepsilon)$.

Combining the bounds for I_1, I_2 and I_3 gives

$$\begin{aligned}
\mathbb{P}\Big(\sup_{f\in\mathcal{H}_\tau}|R_{\text{emp}}(f)-\widetilde{R}_{\text{emp}}(f,\beta(f))|>\varepsilon\Big) &\le C_1\exp(-C_2n\varepsilon^2)+C_3\exp(-C_4\frac{n\varepsilon^2}{(\kappa\tau)^2})+2\exp(-C_5n\,\varepsilon) \\
&\le C_6\exp\Big(-C_7\frac{n\varepsilon^2}{1+(\kappa\tau)^2}\Big)
\end{aligned}$$

for some $C_6, C_7 > 0$. This completes the proof of Theorem S.4. □

S3 Supplementary Lemmas

Lemma 1. *Consider minimizing $f(w) = 2^{-1}w^\top Qw - \beta^T w + 2^{-1}\lambda\|w\|_1$ with respect to $w \in \mathbb{R}^p$ with $w_i \in [-1, 1]$, where Q is positive semi-definite and $\lambda \ge 0$. If $\lambda \ge 2\|\beta\|_\infty$, then the minimizing w is the zero vector.*

Proof. Consider $2^{-1}\lambda\|w\|_1 - \beta^\top w = \sum_{i=1}^p(\lambda/2|w_i| - \beta_i w_i)$. If $\lambda \ge 2\|\beta\|_\infty$, this expression is non-negative for all $w \in \mathbb{R}^p$ and a minimum occurs at $w = 0$. Since Q is positive semi-definite, $w^\top \frac{1}{2}Qw$ is always non-negative with a minimum at $w = 0$. It follows that for $\lambda \ge 2\|\beta\|_\infty$ the sum of these terms attains minimum at $w = 0$. □

Lemma 2. *Let $M = [(C\mathbf{K}C)^2 + n\gamma(C\mathbf{K}C)]^- C\mathbf{K}C$, then $\|M\|_{op} \le (n\gamma)^{-1}$.*

Proof of Lemma 2. The kernel matrix K is positive semi-definite since by the reproducing property for any $\alpha \in \mathbb{R}^n$

$$\alpha^\top\mathbf{K}\alpha=\left\langle\sum_{i=1}^n\alpha_i\Phi(x_i),\sum_{i=1}^n\alpha_i\Phi(x_i)\right\rangle_{\mathcal{H}}=\Big\|\sum_{i=1}^n\alpha_i\Phi(x_i)\Big\|_{\mathcal{H}}^2\ge 0.$$

It follows that $C\mathbf{K}C$ is also positive semi-definite. Let $\{\lambda_i\}_{i=1}^k$ be the set of non-zero eigenvalues of $C\mathbf{K}C$, then $\{\lambda_i/(\lambda_i^2 + n\gamma\lambda_i)\}_{i=1}^k$ are the non-zero eigenvalues of $M = [(C\mathbf{K}C)^2 + n\gamma(C\mathbf{K}C)]^- C\mathbf{K}C$. The function $t \mapsto t/(t^2 + n\gamma t)$ is bounded above by $(n\gamma)^{-1}$ for $t > 0$, hence $\|M\|_{\text{op}} \le (n\gamma)^{-1}$. □

Lemma 3. *Let $\gamma > 0$. The minimizer $\widehat{f}$ in (4) satisfies $\|\widehat{f}\|_{\mathcal{H}} \le 1/\sqrt{\gamma}$. Additionally, if Assumption 2 holds for $\kappa > 0$, then $\|\widehat{f}\|_{\mathcal{H}} \le 2\kappa/\gamma$.*

Proof of Lemma 3. Comparing the value of objective function in (4) at $f = \widehat{f}$ with the value at $f = 0$ gives

$$\gamma\|\widehat{f}\|_{\mathcal{H}}^2 \leq \frac{1}{n}\sum_{i=1}^{n}\Big|y_i^\top\widehat{\theta} - \Big\langle \Phi(x_i) - \overline{\Phi}, \widehat{f}\,\Big\rangle_{\mathcal{H}}\Big|^2 + \gamma\|\widehat{f}\|_{\mathcal{H}}^2 \leq \frac{1}{n}\sum_{i=1}^{n}|y_i^\top\widehat{\theta}|^2 = 1.,$$

where the last equality follows since $n^{-1}\widehat{\theta}Y^\top Y\widehat{\theta} = 1$. It follows that $\|\widehat{f}\|_{\mathcal{H}} \leq 1/\sqrt{\gamma}$.

On the other hand, since $\widehat{f} = \sum_{i=1}^{n}\alpha_i(\Phi(x_i) - \overline{\Phi})$, by the triangle inequality and Assumption 2

$$\|\widehat{f}\|_{\mathcal{H}} = \Big\|\sum_{i=1}^{n}\alpha_i(\Phi(x_i) - \overline{\Phi})\Big\|_{\mathcal{H}} \leq \sum_{i=1}^{n}|\alpha_i|\|\Phi(x_i) - \overline{\Phi}\|_{\mathcal{H}} \leq \max_i\|\Phi(x_i) - \overline{\Phi}\|_{\mathcal{H}}\|\alpha\|_1 \leq 2\kappa\|\alpha\|_1 \leq 2\kappa\sqrt{n}\|\alpha\|_2.$$

Since $\alpha = \{(C\mathbf{K}C)^2 + \gamma nC\mathbf{K}C\}^{-}C\mathbf{K}CY\widehat{\theta}$, applying Lemma 2 and using $\|Y\widehat{\theta}\|_2 = \sqrt{\widehat{\theta}Y^\top Y\widehat{\theta}} = \sqrt{n}$ gives

$$\|\alpha\|_2 \leq \|\{(C\mathbf{K}C)^2 + \gamma nC\mathbf{K}C\}^{-}C\mathbf{K}C\|_{\text{op}}\|Y\widehat{\theta}\|_2 \leq \frac{\|Y\widehat{\theta}\|_2}{n\gamma} \leq \frac{1}{\sqrt{n}\gamma}.$$

Combining the above two displays gives $\|\widehat{f}\|_{\mathcal{H}} \leq 2\kappa/\gamma$. □

Lemma 4. *Under Assumptions 1 and 2, let $\{(x_i, y_i)\}_{i=1}^{n}$ and $\{(x_j, y_j)\}_{j=n+1}^{2n}$ be two independent copies of i.i.d. data, and let T_x be the empirical measure on their union. Let $\widetilde{R}_{emp}(f,\beta)$ be the modified empirical risk on $\{(x_i, y_i)\}_{i=1}^{n}$, and $\widetilde{R}'_{emp}(f,\beta)$ on $\{(x_j, y_j)\}_{i=n+1}^{2n}$. Let $c = 64(\|\theta^*\|_\infty + \tau\kappa)$, and let $\{f_1, \dots, f_M\}$ be the smallest $L^2(T_x)$ $\varepsilon/\sqrt{2}c$-net of $\mathcal{H}_\tau$, and let $\{\beta_1, \dots, \beta_K\}$ be an ε/c-net of I_τ. Then*

$$\mathbb{P}\Bigg(\sup_{\substack{f\in H_\tau\\ \beta\in I_\tau}}\{\widetilde{R}_{emp}(f,\beta) - \widetilde{R}'_{emp}(f,\beta)\} > \frac{\varepsilon}{2}\Bigg) \leq \mathbb{P}\Bigg(\operatorname*{maximize}_{\substack{f\in\{f_1,\dots,f_M\}\\ \beta\in\{\beta_1,\dots,\beta_K\}}}\{\widetilde{R}_{emp}(f,\beta) - \widetilde{R}'_{emp}(f,\beta)\} > \frac{\varepsilon}{4}\Bigg).$$

Proof of Lemma 4. Let $f \in \mathcal{H}_\tau$, $\beta \in I_\tau$ be such that $\widetilde{R}_{\text{emp}}(f,\beta) - \widetilde{R}'_{\text{emp}}(f,\beta) > \varepsilon/2$. There exists $f_j \in \{f_1, \dots, f_M\}$ and $\beta_\ell \in \{\beta_1, \dots, \beta_K\}$ such that $\|f_j - f\|_{L^2(T_x)} < \varepsilon/\sqrt{2}c$ and $|\beta - \beta_\ell| < \varepsilon/c$. Applying Lemma 9 gives

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}|f(x_i) - f_j(x_i)|^2} < \frac{\varepsilon}{c} \quad \text{and} \quad \sqrt{\frac{1}{n}\sum_{i=n+1}^{2n}|f(x_i) - f_j(x_i)|^2} < \frac{\varepsilon}{c}.$$

Applying Lemma 8 yields

$$|\widetilde{R}_{\text{emp}}(f,\beta) - \widetilde{R}_{\text{emp}}(f_j,\beta_\ell)| < 8\frac{\varepsilon}{c}(\|\theta^*\|_\infty + \kappa\tau) = \frac{\varepsilon}{8},$$

and similarly $|\widetilde{R}'_{\text{emp}}(f,\beta) - \widetilde{R}'_{\text{emp}}(f_j,\beta_\ell)| < \varepsilon/8$. Therefore, $\widetilde{R}'_{\text{emp}}(f,\beta) - \widetilde{R}_{\text{emp}}(f,\beta) > \varepsilon/2$ for some $f \in \mathcal{H}_\tau$, $\beta \in I_\tau$ implies $\widetilde{R}'_{\text{emp}}(f_j,\beta_\ell) - \widetilde{R}_{\text{emp}}(f_j,\beta_\ell) > \varepsilon/4$ for some f_j and β_ℓ. Therefore,

$$\mathbb{P}\Bigg(\sup_{f\in\mathcal{H}_\tau,\,\beta\in I_\tau}\{\widetilde{R}'_{\text{emp}}(f,\beta) - \widetilde{R}_{\text{emp}}(f,\beta)\} > \frac{\varepsilon}{2}\Bigg) \leq \mathbb{P}\Bigg(\operatorname*{maximize}_{\substack{f\in\{f_1,\dots,f_M\}\\ \beta\in\{\beta_1,\dots,\beta_K\}}}\{\widetilde{R}'_{\text{emp}}(f_j,\beta_\ell) - \widetilde{R}_{\text{emp}}(f_j,\beta_\ell)\} > \frac{\varepsilon}{4}\Bigg).$$

□

Lemma 5. *Under Assumptions 1-3, let $\{f_1, \dots, f_M\}$ and $\{\beta_1, \dots, \beta_K\}$ be as in Lemma 4. There exist a constant $C_1 > 0$ such that for all $\varepsilon > 0$,*

$$\mathbb{P}\Bigg(\operatorname*{maximize}_{\substack{f\in\{f_1,\dots,f_M\}\\ \beta\in\{\beta_1,\dots,\beta_K\}}}\{\widetilde{R}_{emp}(f,\beta) - \widetilde{R}'_{emp}(f,\beta)\} > \frac{\varepsilon}{4}\Bigg) \leq \mathcal{N}_\varepsilon\,\exp\Big(-\frac{n\varepsilon^2}{128(\|\theta^*\|_\infty + \kappa\tau)^4}\Big),$$

where $\mathcal{N}_\varepsilon = \{1 + 2(\|\theta^*\|_\infty + \kappa\tau)/\,\varepsilon\}\exp(C_1\tau^2\varepsilon^{-2})$.

Proof of Lemma 5. Let $\sigma = \{\sigma_i\}_{i=1}^n$ be *i.i.d.* Radamacher random variables, $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 1/2$. Let

$$\widetilde{R}^{\sigma}_{\text{emp}} = \frac{1}{n}\sum_{i=1}^{n}\sigma_i|y_i^\top\theta^* - \beta - \langle\Phi(x_i), f\rangle_{\mathcal{H}}|^2, \quad \widetilde{R}'^{\sigma}_{\text{emp}} = \frac{1}{n}\sum_{i=n+1}^{2n}\sigma_i|y_i^\top\theta^* - \beta - \langle\Phi(x_i), f\rangle_{\mathcal{H}}|^2.$$

Since (y_i, x_i) and (y_{n+i}, x_{n+i}) are independent, and have the same distribution, the distribution of $\xi_i := (|y_i^\top\theta^* - \beta - \langle\Phi(x_i), f\rangle_{\mathcal{H}}|^2 - |y_{n+i}^\top\theta^* - \beta - \langle\Phi(x_{n+i}), f\rangle_{\mathcal{H}}|^2)$ is the same as distribution of $\sigma_i\xi_i$. Let $Z = \{(x_i, y_i)\}_{i=1}^{2n}$, then

$$\mathbb{P}_Z\Bigg(\max_{\substack{f\in\{f_1,\dots,f_M\}\\ \beta\in\{\beta_1,\dots,\beta_K\}}}\{\widetilde{R}_{\text{emp}}(f,\beta) - \widetilde{R}'_{\text{emp}}(f,\beta)\} > \frac{\varepsilon}{4}\Bigg) = \mathbb{P}_{Z,\sigma}\Bigg(\max_{\substack{f\in\{f_1,\dots,f_M\}\\ \beta\in\{\beta_1,\dots,\beta_K\}}}\{\widetilde{R}^{\sigma}_{\text{emp}}(f,\beta) - \widetilde{R}'^{\sigma}_{\text{emp}}(f,\beta)\} > \frac{\varepsilon}{4}\Bigg).$$

Let $\mathcal{A}_{m,k}$ be the event $\mathcal{A}_{m,k} = \{\widetilde{R}^{\sigma}_{\text{emp}}(f_m,\beta_k) - \widetilde{R}'^{\sigma}_{\text{emp}}(f_m,\beta_k) > \varepsilon/4\}$ for $m = 1,\dots,M(Z)$; $k = 1,\dots,K$; where $M(Z)$ emphasizes the dependence of M on Z. Using properties of conditional expectation and union bound

$$\begin{aligned}\mathbb{P}_{Z,\sigma}\Bigg(\max_{\substack{f\in\{f_1,\dots,f_M\}\\ \beta\in\{\beta_1,\dots,\beta_K\}}}\{\widetilde{R}^{\sigma}_{\text{emp}}(f,\beta) - \widetilde{R}'^{\sigma}_{\text{emp}}(f,\beta)\} > \frac{\varepsilon}{4}\Bigg) &= \mathbb{P}_{Z,\sigma}(\cup_{m=1}^{M(Z)}\cup_{k=1}^{K}\mathcal{A}_{m,k})\\ &= \mathbb{E}_Z\left\{\mathbb{P}_{\sigma}(\cup_{m=1}^{M(Z)}\cup_{k=1}^{K}\mathcal{A}_{m,k}|Z)\right\}\\ &\le \mathbb{E}_Z\left\{M(Z)K\mathbb{P}_{\sigma}(\mathcal{A}_{m,k}|Z)\right\}.\end{aligned}$$

For fixed f_m, β_k and conditionally on Z, the terms $\psi_i := \sigma_i(|y_i^\top\theta^* - \beta_k - \langle\Phi(x_i), f_m\rangle_{\mathcal{H}}|^2 - |y_{n+i}^\top\theta^* - \beta_k - \langle\Phi(x_{n+i}), f_m\rangle_{\mathcal{H}}|^2)$, $i = 1,\dots,n$, are independent, mean-zero random variables with $|\psi_i| \le 4(\|\theta^*\|_\infty + \kappa\tau)^2$. Applying Hoeffding's inequality gives

$$\mathbb{P}_{\sigma}(\mathcal{A}_{m,k}|Z) = \mathbb{P}_{\sigma}\Big(\frac{1}{n}\sum_{i=1}^{n}\psi_i > \varepsilon/4\,\Big|\,Z\Big) \le \exp\Big(-\frac{n\varepsilon^2}{128(\|\theta^*\|_\infty + \kappa\tau)^4}\Big).$$

On the other hand, since I_τ is a one-dimensional sphere of radius $\|\theta^*\| + \kappa\tau$, K is independent of the data and $K \le 1 + 2(\|\theta^*\|_\infty + \kappa\tau)/\varepsilon$. Combining this with the above two displays gives

$$\begin{aligned}&\mathbb{P}_{Z,\sigma}\Bigg(\max_{\substack{f\in\{f_1,\dots,f_M\}\\ \beta\in\{\beta_1,\dots,\beta_K\}}}\{\widetilde{R}^{\sigma}_{\text{emp}}(f,\beta) - \widetilde{R}'^{\sigma}_{\text{emp}}(f,\beta)\} > \frac{\varepsilon}{4}\Bigg)\\ &\le \{1 + 2(\|\theta^*\|_\infty + \kappa\tau)/\varepsilon\}\,\mathbb{E}_Z\{M(Z)\}\exp\Big(-\frac{n\varepsilon^2}{128(\|\theta^*\|_\infty + \kappa\tau)^4}\Big).\end{aligned}$$

Recall that $\{f_1,\dots,f_M\}$ is the smallest $L^2(T_x)$ $\varepsilon/\sqrt{2}c$-net of $\mathcal{H}_\tau$, with $c = 64(\|\theta^*\|_\infty + \tau\kappa)$. By Lemma 10

$$\mathbb{E}_Z\{M(Z)\} \le \sup_{Z=\{(x_i,y_i)\}_{i=1}^{2n}} M(Z) \le \exp\left(\frac{C_1\tau^2}{\varepsilon^2}\right) \tag{S3.1}$$

for some constant $C_1 > 0$. Setting $\mathcal{N}_\varepsilon = \{1 + 2(\|\theta^*\|_\infty + \kappa\tau)/\varepsilon\}\exp(C_1\tau^2\varepsilon^{-2})$ completes the proof of Lemma 5. □

Lemma 6. *Under Assumption 1 there exist constants $C_1, C_2 > 0$ such that for all $\varepsilon > 0$,*

$$\mathbb{P}\bigg(\max\Big(|n_1/n_2 - \pi_1/\pi_2|, |n_2/n_1 - \pi_2/\pi_1|\Big) > \varepsilon\bigg) \le C_1\exp\Big(-C_2n\varepsilon^2\Big),$$

$$\mathbb{P}\bigg(\max\Big(|\sqrt{n_1/n_2} - \sqrt{\pi_1/\pi_2}|, |\sqrt{n_2/n_1} - \sqrt{\pi_2/\pi_1}|\Big) > \varepsilon\bigg) \le C_1\exp\Big(-C_2n\varepsilon^2\Big).$$

Proof of Lemma 6. We provide the proof for n_1/n_2, the proof for n_2/n_1 is analogous. The first inequality is equivalent to Lemma 1 in [2]. For the second inequality, by Taylor expansion of the square root function centered at π_1/π_2

$$\sqrt{n_1/n_2} - \sqrt{\pi_1/\pi_2} = 2^{-1}\sqrt{\pi_2/\pi_1}(n_1/n_2 - \pi_1/\pi_2) + o(n_1/n_2 - \pi_1/\pi_2).$$

Since $|n_1/n_2 - \pi_1/\pi_2| = O_p(n^{-1/2})$ by the first inequality, it follows that there exist a constant $C_3 > 0$ such that $|\sqrt{n_1/n_2} - \sqrt{\pi_1/\pi_2}| \leq C_2\{\log(\eta^{-1})/n\}^{1/2}$ with probability at least $1-\eta$. Setting $\varepsilon = C_3\{\log(\eta^{-1})/n\}^{1/2}$ and solving for η completes the proof. □

Lemma 7. *Let Assumption 1 be true. For all $\varepsilon > 0$, we have $\mathbb{P}\big((\overline{Y\theta^*})^2 > \varepsilon\big) \leq 2\exp(-n\varepsilon/\|\theta^*\|_\infty)$.*

Proof of Lemma 7. Let $z_i = y_i^\top\theta^*$, then z_i are independent,

$$\mathbb{E}(z_i) = \mathbb{E}(y_i)^\top\theta^* = \pi_1\sqrt{\frac{\pi_2}{\pi_1}} - \pi_2\sqrt{\frac{\pi_1}{\pi_2}} = \sqrt{\pi_1\pi_2} - \sqrt{\pi_1\pi_2} = 0$$

and

$$(\overline{Y\theta^*})^2 = (n^{-1}\sum_{i=1}^n y_i^\top\theta^*)^2 = (n^{-1}\sum_{i=1}^n z_i)^2.$$

Since $|z_i| \leq \|\theta^*\|_\infty = \sqrt{\pi_{\max}/\pi_{\min}}$, by Hoeffding's inequality for $\varepsilon > 0$

$$\mathbb{P}\Big(\Big|n^{-1}\sum_{i=1}^n z_i\Big|^2 > \varepsilon\Big) = \mathbb{P}\Big(\Big|n^{-1}\sum_{i=1}^n z_i\Big| > \sqrt{\varepsilon}\Big) \leq 2\exp(-n\varepsilon/\|\theta^*\|_\infty).$$

□

Lemma 8. *Let Assumptions 1 and 2 be true, and suppose that $\{f_1,\dots,f_M\}$ is an $L^2(T_x)$ ε-net of $\mathcal{H}_\tau$ and that $\{\beta_1,\dots,\beta_K\}$ be an ε-net of I_τ. Then for any admissible f and β, let f_j and β_ℓ be members of the ε-nets so that $\|f-f_j\|_{L^2(T_x)} < \varepsilon$ and $|\beta-\beta_\ell| < \varepsilon$. Then*

$$\Big|\widetilde{R}_{emp}(f,\beta) - \widetilde{R}_{emp}(f_j,\beta_l)\Big| \leq 8\varepsilon\Big(\|\theta^*\|_\infty + \kappa\tau\Big). \tag{S3.2}$$

Proof of Lemma 8. By the reproducing property of $\mathcal{H}$, $\langle\Phi(x_i), f\rangle_{\mathcal{H}} = f(x_i)$, and

$$\begin{aligned}
\Big|\widetilde{R}_{emp}(f,\beta) - \widetilde{R}_{emp}(f_j,\beta_l)\Big| &= \Big|\frac{1}{n}\sum_{i=1}^n |y_i^\top\theta^* - \beta - \langle\Phi(x_i), f\rangle_{\mathcal{H}}|^2 - \frac{1}{n}\sum_{i=1}^n |y_i^\top\theta^* - \beta_\ell - \langle\Phi(x_i), f_j\rangle_{\mathcal{H}}|^2\Big| \\
&= \Big|\frac{1}{n}\sum_{i=1}^n |y_i^\top\theta^* - \beta - f(x_i)|^2 - \frac{1}{n}\sum_{i=1}^n |y_i^\top\theta^* - \beta_\ell - f_j(x_l)|^2\Big| \\
&= \Big| - 2\frac{1}{n}\sum_{i=1}^n y_i^\top\theta^*\{\beta + f(x_i) - \beta_\ell - f_j(x_i)\} + \frac{1}{n}\sum_{i=1}^n[\{\beta + f(x_i)\}^2 - \{\beta_\ell + f_j(x_i)\}^2]\Big| \\
&\leq \underbrace{2\|\theta^*\|_\infty\Big|\beta - \beta_l + \frac{1}{n}\sum_{i=1}^n\{f(x_i) - f_j(x_i)\}\Big|}_{I_1} + \underbrace{\Big|\frac{1}{n}\sum_{i=1}^n[\{\beta + f(x_i)\}^2 - \{\beta_\ell + f_j(x_i)\}^2]\Big|}_{I_2}.
\end{aligned}$$

Consider

$$\begin{aligned}
I_1 = 2\|\theta^*\|_\infty\Big|\beta - \beta_l + \frac{1}{n}\sum_{i=1}^n\{f(x_i) - f_j(x_i)\}\Big| &\leq 2\|\theta^*\|_\infty\left\{|\beta - \beta_l| + \frac{1}{n}\sum_{i=1}^n|f(x_i) - f_j(x_i)|\right\} \\
&\leq 2\|\theta^*\|_\infty\left\{\varepsilon + \Big[\frac{1}{n}\sum_{i=1}^n|f(x_i) - f_j(x_i)|^2\Big]^{1/2}\right\} \\
&\leq 4\|\theta^*\|_\infty\varepsilon,
\end{aligned}$$

where we used $n^{-1}\sum_{i=1}^n [|f(x_i)-f_j(x_i)|^2]^{1/2} \le [n^{-1}\sum_{i=1}^n |f(x_i)-f_j(x_i)|^2]^{1/2}$ due to Jensen's inequality, and that $\|f-f_j\|_{L^2(T_x)} < \varepsilon$ and $|\beta-\beta_\ell| < \varepsilon$.

Consider I_2. Using $a^2-b^2=(a+b)(a-b)$, the Cauchy-Schwarz inequalty, and Jensen's inequality,

$$\begin{aligned} I_2 &= \frac{1}{n}\Big|\sum_{i=1}^n \{\beta+f(x_i)+\beta_\ell+f_j(x_i)\}\{\beta-\beta_\ell+f(x_i)-f_j(x_i)\}\Big| \\ &\le 2(\sup_{\beta\in I_\tau}|\beta| + \sup_{x,f\in\mathcal{H}_\tau}|f(x)|)\frac{1}{n}\sum_{i=1}^n(|\beta-\beta_j|+|f(x_i)-f_j(x_i)|) \\ &\le 2(\|\theta^*\|_\infty + \kappa\tau + \sup_{x,f\in\mathcal{H}_\tau}|\langle\Phi(x),f\rangle_{\mathcal{H}}|)(\varepsilon + \frac{1}{n}\sum_{i=1}^n|f(x_i)-f_j(x_i)|) \\ &\le 2\Big(\|\theta^*\|_\infty+\kappa\tau+\kappa\tau\Big)\Big(\varepsilon+\sqrt{\frac{1}{n}\sum_{i=1}^n|f(x_i)-f_j(x_i)|^2}\Big) \\ &= 4\varepsilon\Big(\|\theta^*\|_\infty+2\kappa\tau\Big). \end{aligned}$$

Combining the bounds for I_1 and I_2 completes the proof of Lemma 8. □

Lemma 9. *Let $\{(x_i,y_i)\}_{i=1}^{2n}$ be the data, and consider an $L^2(T_x)$ ε-net $\{f_1,\dots,f_M\}$ of $\mathcal{H}_\tau$. Then $\{f_1,\dots,f_M\}$ is an $\sqrt{2}\varepsilon$-net with respect to the empirical measure on half of the data $\{(x_i,y_i)\}_{i=1}^n$.*

Proof of Lemma 9. Since $\{f_1,\dots,f_M\}$ is ε-net with respect to $\{(x_i,y_i)\}_{i=1}^{2n}$, for any $f\in\mathcal{H}_\tau$, there exists f_j such that

$$\sqrt{\frac{1}{2n}\sum_{i=1}^{2n}|f(x_i)-f_j(x_i)|^2} < \varepsilon.$$

If $\frac{1}{2n}\sum_{i=1}^{2n}|f(x_i)-f_j(x_i)|^2=0$, then $\frac{1}{n}\sum_{i=1}^n|f(x_i)-f_j(x_i)|^2=0$. Otherwise

$$\begin{aligned} \sqrt{\frac{1}{n}\sum_{i=1}^n|f(x_i)-f_j(x_i)|^2} &= \sqrt{\frac{2n}{2n}\frac{1}{n}\sum_{i=1}^n|f(x_i)-f_j(x_i)|^2\frac{\sum_{i=1}^{2n}|f(x_i)-f_j(x_i)|^2}{\sum_{i=1}^{2n}|f(x_i)-f_j(x_i)|^2}} \\ &= \sqrt{\frac{2n}{n}\frac{\sum_{i=1}^n|f(x_i)-f_j(x_i)|^2}{\sum_{i=1}^{2n}|f(x_i)-f_j(x_i)|^2}}\sqrt{\frac{1}{2n}\sum_{i=1}^{2n}|f(x_i)-f_j(x_i)|^2} < \sqrt{2}\varepsilon, \end{aligned}$$

hence $\{f_1,\dots,f_M\}$ is $\sqrt{2}\varepsilon$-net with respect to $\{(x_i,y_i)\}_{i=1}^n$. □

Lemma 10 (Theorem 2.1 of [3])**.** *Let Assumption 3 be true, and Let $M(Z)$ be the size of an $L^2(T_x)$ ε-covering number of $\mathcal{H}_\tau$ with data $Z=\{(x_i,y_i)\}_{i=1}^n$. There exists a $C>0$ independent of n, such that*

$$\sup_{Z=\{(x_i,y_i)\}_{i=1}^n} M(Z) \le \exp\left(\frac{C\tau^2}{\varepsilon^2}\right). \tag{S3.3}$$

Remark 2. *[4] notes that "Theorem 2.1 of [3] considered only the Gaussian RKHS, however the proof of the entropy bound for $p=2$ in their notation only requires that the RKHS is separable." It is this case which is presented in Lemma 10.*

References

[1] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Advanced lectures on machine learning*, pages 169–207. Springer, 2004.

[2] Irina Gaynanova and Tianying Wang. Sparse quadratic classification rules via linear dimension reduction. *arXiv preprint arXiv:1711.04817*, 2017.

[3] Ingo Steinwart and Clint Scovel. Fast rates for support vector machines using gaussian kernels. *The Annals of Statistics*, pages 575–607, 2007.

[4] Chong Zhang, Yufeng Liu, and Yichao Wu. On quantile regression in reproducing kernel hilbert spaces with data sparsity constraint. *Journal of Machine Learning Research*, 17(40):1–45, 2016.