Ultrahigh Dimensional Feature Screening via RKHS Embeddings - Supplementary material

This document contains the statement and proof of Lemma 1 which is used to prove Theorem 4.1.

A Statement and Proof of Lemma 1

Lemma 1. *Let $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ be measurable kernels satisfying assumptions* **A1** *and* **A2**. *Then for any $1 \leq r \leq p_n$, with probability at least $1-\delta$ over the choice of samples, $\{(x_r^{(i)}, y^{(i)})\}$,*

$$|\widehat{\omega}_r - \omega_r| \leq \sqrt{\frac{8U_n(\mathcal{K};\{(x_r^{(i)}, y^{(i)})\})}{n}} + \sqrt{\frac{8AU_n(\mathcal{K}_{\mathcal{X}};\{x_r^{(i)}\})}{n}} + \sqrt{\frac{8AU_n(\mathcal{K}_{\mathcal{Y}};\{y^{(i)}\})}{n}} + \sqrt{\frac{162A^2}{n}\log\frac{6}{\delta}} + \frac{6A}{\sqrt{n}}.$$

Proof. The proof technique is similar to that of Theorem 7 in (Sriperumbudur et al., 2009). Consider $|\widehat{\omega}_r - \omega_r| = |\widehat{\gamma}_r(\mathbb{P}^{X_rY}, \mathbb{P}^{X_r}\mathbb{P}^{Y}) - \gamma_r(\mathbb{P}^{X_rY}, \mathbb{P}^{X_r}\mathbb{P}^{Y})| \leq \sup_{k\in\mathcal{K}} \|\mathbb{P}_n^{X_rY}k - \mathbb{P}^{X_rY}k\|_{\mathcal{H}} + \sup_{k\in\mathcal{K}} \|\mathbb{P}_n^{X_r}\mathbb{P}_n^{Y}k - \mathbb{P}^{X_r}\mathbb{P}^{Y}k\|_{\mathcal{H}}$. We now bound the terms $\theta := \sup_{k\in\mathcal{K}} \|\mathbb{P}_n^{X_rY}k - \mathbb{P}^{X_rY}k\|_{\mathcal{H}}$ and $\phi := \sup_{k\in\mathcal{K}} \|\mathbb{P}_n^{X_r}\mathbb{P}_n^{Y}k - \mathbb{P}^{X_r}\mathbb{P}^{Y}k\|_{\mathcal{H}}$. Since θ satisfies the bounded difference property, using McDiarmid's inequality gives that with probability at least $1-\frac{\delta}{6}$ over the choice of $\{(x_r^{(i)}, y^{(i)})\}_{i=1}^n$, we have

$$\theta \leq \mathsf{E}\sup_{k\in\mathcal{K}} \|\mathbb{P}_n^{X_rY}k - \mathbb{P}^{X_rY}k\|_{\mathcal{H}} + \sqrt{\frac{2A^2}{n}\log\frac{6}{\delta}}. \tag{1}$$

By invoking symmetrization for $\mathsf{E}\sup_{k\in\mathcal{K}} \|\mathbb{P}_n^{X_rY}k - \mathbb{P}^{X_rY}k\|_{\mathcal{H}}$, we have

$$\mathsf{E}\,\theta \leq 2\mathsf{E}\,\mathsf{E}_{\rho}\sup_{k\in\mathcal{K}} \left\| \frac{1}{n}\sum_{i=1}^n \rho_i k(., (x_r^{(i)}, y^{(i)})) \right\|_{\mathcal{H}}, \tag{2}$$

where $\{\rho_i\}_{i=1}^n$ represent i.i.d. Rademacher random variables and E_{ρ} represents the expectation w.r.t. $\{\rho_i\}$ conditioned on $\{(x_r^{(i)}, y^{(i)})\}$. Since $\mathsf{E}_{\rho}\sup_{k\in\mathcal{K}} \left\| \frac{1}{n}\sum_{i=1}^n \rho_i k(., (x_r^{(i)}, y^{(i)})) \right\|_{\mathcal{H}}$ satisfies the bounded difference property, by McDiarmid's inequality, with probability at least $1-\frac{\delta}{6}$ over the choice of the random samples of size n, we have

$$\mathsf{E}\,\mathsf{E}_{\rho}\sup_{k\in\mathcal{K}} \left\| \frac{1}{n}\sum_{i=1}^n \rho_i k(., (x_r^{(i)}, y^{(i)})) \right\|_{\mathcal{H}} \leq \sqrt{\frac{2A^2}{n}\log\frac{6}{\delta}} + \mathsf{E}_{\rho}\sup_{k\in\mathcal{K}} \left\| \frac{1}{n}\sum_{i=1}^n \rho_i k(., (x_r^{(i)}, y^{(i)})) \right\|_{\mathcal{H}}. \tag{3}$$

By writing

$$\left\| \frac{1}{n}\sum_{i=1}^n \rho_i k(., (x_r^{(i)}, y^{(i)})) \right\|_{\mathcal{H}} \leq \frac{A}{\sqrt{n}} + \frac{\sqrt{2}}{n}\sqrt{\left|\sum_{i<j}^n \rho_i\rho_j k((x_r^{(i)}, y^{(i)}), (x_r^{(j)}, y^{(j)}))\right|} \tag{4}$$

we have with probability at least $1-\frac{\delta}{6}$, the following holds:

$$\mathsf{E}\,\mathsf{E}_{\rho}\sup_{k\in\mathcal{K}} \left\| \frac{1}{n}\sum_{i=1}^n \rho_i k(., (x_r^{(i)}, y^{(i)})) \right\|_{\mathcal{H}} \leq \sqrt{\frac{2A^2}{n}\log\frac{6}{\delta}} + \frac{A}{\sqrt{n}} + \sqrt{\frac{2U_n(\mathcal{K};\{(x_r^{(i)}, y^{(i)})\})}{n}}. \tag{5}$$

Tying (1)-(5), we have that w.p. at least $1-\frac{\delta}{3}$ over the choice of $\{(x_r^{(i)}, y^{(i)})\}$, the following holds:

$$\theta \leq \sqrt{\frac{8U_n(\mathcal{K};\{(x_r^{(i)}, y^{(i)})\})}{n}} + \frac{2A}{\sqrt{n}} + \sqrt{\frac{18A^2}{n}\log\frac{6}{\delta}}. \tag{6}$$

Now we consider bounding ϕ

$$
\begin{aligned}
\phi &\stackrel{\text{def}}{=} \sup_{k \in \mathcal{K}} \|\mathbb{P}_n^{X_r}\mathbb{P}_n^Y k - \mathbb{P}^{X_r}\mathbb{P}^Y k\|_{\mathcal{H}} \\
&= \sup_{k \in \mathcal{K}} \left\| \int k_{\mathcal{X}}(\cdot, x) k_{\mathcal{Y}}(\cdot, y)\, d[(\mathbb{P}^{X_r} \times \mathbb{P}^Y) - (\mathbb{P}_n^{X_r} \times \mathbb{P}_n^Y)](x, y) \right\|_{\mathcal{H}} \\
&= \sup_{k \in \mathcal{K}} \left\| \int k_{\mathcal{X}}(\cdot, x)\, d\mathbb{P}^{X_r}(x) \int k_{\mathcal{Y}}(\cdot, y)\, d\mathbb{P}^Y(y) - \int k_{\mathcal{X}}(\cdot, x)\, d\mathbb{P}_n^{X_r}(x) \int k_{\mathcal{Y}}(\cdot, y)\, d\mathbb{P}_n^Y(y) \right\|_{\mathcal{H}} \\
&\le \sup_{k \in \mathcal{K}} \left\| \int k_{\mathcal{X}}(\cdot, x)\, d(\mathbb{P}^{X_r} - \mathbb{P}_n^{X_r})(x) \int k_{\mathcal{Y}}(\cdot, y)\, d\mathbb{P}^Y(y) \right\|_{\mathcal{H}} \\
&\qquad + \sup_{k \in \mathcal{K}} \left\| \int k_{\mathcal{X}}(\cdot, x)\, d\mathbb{P}_n^{X_r}(x) \int k_{\mathcal{Y}}(\cdot, y)\, d(\mathbb{P}^Y - \mathbb{P}_n^Y)(y) \right\|_{\mathcal{H}} \\
&= \sup_{k \in \mathcal{K}} \left\| \int k_{\mathcal{X}}(\cdot, x)\, d(\mathbb{P}^{X_r} - \mathbb{P}_n^{X_r})(x) \right\|_{\mathcal{H}_{\mathcal{X}}} \left\| \int k_{\mathcal{Y}}(\cdot, y)\, d\mathbb{P}^Y(y) \right\|_{\mathcal{H}_{\mathcal{Y}}} \\
&\qquad + \sup_{k \in \mathcal{K}} \left\| \int k_{\mathcal{X}}(\cdot, x)\, d\mathbb{P}_n^{X_r}(x) \right\|_{\mathcal{H}_{\mathcal{X}}} \left\| \int k_{\mathcal{Y}}(\cdot, y)\, d(\mathbb{P}^Y - \mathbb{P}_n^Y)(y) \right\|_{\mathcal{H}_{\mathcal{Y}}} \\
&= \sup_{k_{\mathcal{X}} \in \mathcal{K}_{\mathcal{X}}} \left\| \int k_{\mathcal{X}}(\cdot, x)\, d(\mathbb{P}^{X_r} - \mathbb{P}_n^{X_r})(x) \right\|_{\mathcal{H}_{\mathcal{X}}} \sup_{k_{\mathcal{Y}} \in \mathcal{K}_{\mathcal{Y}}} \left\| \int k_{\mathcal{Y}}(\cdot, y)\, d\mathbb{P}^Y(y) \right\|_{\mathcal{H}_{\mathcal{Y}}} \\
&\qquad + \sup_{k_{\mathcal{Y}} \in \mathcal{K}_{\mathcal{Y}}} \left\| \int k_{\mathcal{Y}}(\cdot, y)\, d(\mathbb{P}^Y - \mathbb{P}_n^Y)(y) \right\|_{\mathcal{H}_{\mathcal{Y}}} \sup_{k_{\mathcal{X}} \in \mathcal{K}_{\mathcal{X}}} \left\| \int k_{\mathcal{X}}(\cdot, x)\, d\mathbb{P}_n^{X_r}(x) \right\|_{\mathcal{H}_{\mathcal{X}}} \\
&\le \sqrt{A} \sup_{k_{\mathcal{X}} \in \mathcal{K}_{\mathcal{X}}} \left\| \int k_{\mathcal{X}}(\cdot, x)\, d(\mathbb{P}^{X_r} - \mathbb{P}_n^{X_r})(x) \right\|_{\mathcal{H}_{\mathcal{X}}} + \sqrt{A} \sup_{k_{\mathcal{Y}} \in \mathcal{K}_{\mathcal{Y}}} \left\| \int k_{\mathcal{Y}}(\cdot, y)\, d(\mathbb{P}^Y - \mathbb{P}_n^Y)(y) \right\|_{\mathcal{H}_{\mathcal{Y}}}.
\end{aligned}
$$

Now, $\phi_{\mathcal{X}} \stackrel{\text{def}}{=} \sup_{k_{\mathcal{X}} \in \mathcal{K}_{\mathcal{X}}} \| \int k_{\mathcal{X}}(\cdot, x)\, d(\mathbb{P}^{X_r} - \mathbb{P}_n^{X_r})(x)\|_{\mathcal{H}_{\mathcal{X}}}$ and $\phi_{\mathcal{Y}} \stackrel{\text{def}}{=} \sup_{k_{\mathcal{Y}} \in \mathcal{K}_{\mathcal{Y}}} \| \int k_{\mathcal{Y}}(\cdot, y)\, d(\mathbb{P}^Y - \mathbb{P}_n^Y)(y)\|_{\mathcal{H}_{\mathcal{Y}}}$ can be bounded by using Theorem 7 of (Sriperumbudur et al., 2009), which yields that probability at least $1 - \frac{\delta}{3}$

$$
\phi_{\mathcal{X}} \le \sqrt{\frac{8U_n(\mathcal{K}_{\mathcal{X}}; \{x_r^{(i)}\})}{n}} + \frac{2\sqrt{A}}{\sqrt{n}} + \sqrt{\frac{18A}{n} \log \frac{6}{\delta}} \tag{7}
$$

and

$$
\phi_{\mathcal{Y}} \le \sqrt{\frac{8U_n(\mathcal{K}_{\mathcal{Y}}; \{y^{(i)}\})}{n}} + \frac{2\sqrt{A}}{\sqrt{n}} + \sqrt{\frac{18A}{n} \log \frac{6}{\delta}}. \tag{8}
$$

Using (7) and (8), with probability at least $1 - \frac{2\delta}{3}$ over the choice of $\{x_r^{(i)}\}$ and $\{y^{(i)}\}$, we have

$$
\phi \le \sqrt{\frac{8AU_n(\mathcal{K}_{\mathcal{Y}}; \{y^{(i)}\})}{n}} + \frac{4A}{\sqrt{n}} + \sqrt{\frac{72A^2}{n} \log \frac{6}{\delta}} + \sqrt{\frac{8AU_n(\mathcal{K}_{\mathcal{X}}; \{x_r^{(i)}\})}{n}}. \tag{9}
$$

Combining (6) and (9) provides the result. □

References

Sriperumbudur, B., Fukumizu, K., Gretton, A., Lanckriet, G., and Schölkopf, B. (2009). Kernel choice and classifiability for RKHS embeddings of probability distributions. In *Advances in Neural Information Processing Systems 22*, pages 1750–1758. MIT Press.