Mixture Proportion Estimation via Kernel Embeddings of Distributions

Supplementary Material

A. Proof of Propositions 1, 2, 3 and 4

Proposition.

$$d(\lambda) = 0, \qquad \forall \lambda \in [0, \lambda^*],$$
$$\widehat{d}(\lambda) = 0, \qquad \forall \lambda \in [0, 1].$$

Proof. The second equality is obvious and follows from convexity of $\mathcal{C}_S$ and that both $\phi(\widehat{F})$ and $\phi(\widehat{H})$ are in $\mathcal{C}_S$.

The first statement is due to the following. Let $\lambda \in [0, \lambda^*]$, then we have that,

$$\begin{aligned} d(\lambda) &= \inf_{w \in \mathcal{C}} \|\lambda \phi(F) + (1-\lambda)\phi(H) - w\|_{\mathcal{H}} \\ &= \inf_{w \in \mathcal{C}} \left\| \frac{\lambda}{\lambda^*}(\lambda^* \phi(F) + (1-\lambda^*)\phi(H)) + \left(1 - \frac{\lambda}{\lambda^*}\right)\phi(H) - w \right\|_{\mathcal{H}} \\ &= \inf_{w \in \mathcal{C}} \left\| \frac{\lambda}{\lambda^*}(\phi(G)) + \left(1 - \frac{\lambda}{\lambda^*}\right)\phi(H) - w \right\|_{\mathcal{H}} \\ &= 0\,. \end{aligned}$$

□

Proposition. *$d(.)$ and $\widehat{d}(.)$ are non-decreasing convex functions.*

Proof. Let $0 < \lambda_1 < \lambda_2$. Let $\epsilon > 0$. Let $w_1, w_2 \in \mathcal{C}$ be such that

$$d(\lambda_1) \geq \|(\lambda_1)\phi(F) + (1-\lambda_1)\phi(H) - w_1\|_{\mathcal{H}} - \epsilon,$$
$$d(\lambda_2) \geq \|(\lambda_2)\phi(F) + (1-\lambda_2)\phi(H) - w_2\|_{\mathcal{H}} - \epsilon\,.$$

By definition of $d(.)$ such w_1, w_2 exist for all $\epsilon > 0$.

Let $\gamma \in [0, 1]$, $\lambda_\gamma = (1-\gamma)\lambda_1 + \gamma\lambda_2$ and $w_\gamma = (1-\gamma)w_1 + \gamma w_2$. We then have that

$$\begin{aligned} d(\lambda_\gamma) &\leq \|(\lambda_\gamma)\phi(F) + (1-\lambda_\gamma)\phi(H) - w_\gamma\|_{\mathcal{H}} \\ &= \|((1-\gamma)\lambda_1 + \gamma\lambda_2)\phi(F) + (1 - (1-\gamma)\lambda_1 - \gamma\lambda_2)\phi(H) - w_\gamma\|_{\mathcal{H}} \\ &= \|((1-\gamma)\lambda_1 + \gamma\lambda_2)\phi(F) + ((1-\gamma)(1-\lambda_1) + \gamma(1-\lambda_2))\phi(H) - w_\gamma\|_{\mathcal{H}} \\ &= \|(1-\gamma)\left(\lambda_1\phi(F) + (1-\lambda_1)\phi(H) - w_1\right) + \gamma\left(\lambda_2\phi(F) + (1-\lambda_2)\phi(H) - w_2\right)\| \\ &\leq (1-\gamma)\left\|(\lambda_1\phi(F) + (1-\lambda_1)\phi(H) - w_1)\right\| + \gamma\left\|(\lambda_2\phi(F) + (1-\lambda_2)\phi(H) - w_2)\right\| \\ &\leq (1-\gamma)(d(\lambda_1) + \epsilon) + \gamma(d(\lambda_2) + \epsilon) \\ &= (1-\gamma)d(\lambda_1) + \gamma d(\lambda_2) + \epsilon\,. \end{aligned}$$

As the above holds for all $\epsilon > 0$ and $d(\lambda_\gamma)$ is independent of ϵ, we have

$$d(\lambda_\gamma) = d((1-\gamma)\lambda_1 + \gamma\lambda_2) \leq (1-\gamma)d(\lambda_1) + \gamma d(\lambda_2).$$

Thus we have that $d(.)$ is convex.

As $\mathcal{C}$ is convex and $\phi(H), \phi(F) \in \mathcal{C}$, we have that $d(\lambda) = 0$ for $\lambda \in [0, \lambda^*]$, and hence $\nabla d(\lambda) = 0$ for $\lambda \in [0, \lambda^*]$. By convexity, we then have that for all $\lambda \geq 0$, all elements of the sub-differential $\partial d(\lambda)$ are non-negative and hence $d(.)$ is a non-decreasing function.

By very similar arguments, we can also show that $\widehat{d}(.)$ is convex and non-decreasing. □

Proposition. *For all* $\mu \geq 0$

$$d(\lambda^* + \mu) = \inf_{w \in \mathcal{C}} \|\phi(G) + \mu(\phi(F) - \phi(H)) - w\|_{\mathcal{H}}.$$

Proof.

$$\begin{aligned} d(\lambda^* + \mu) &= \inf_{w \in \mathcal{C}} \|(\lambda^* + \mu)\phi(F) + (1 - \lambda^* - \mu)\phi(H) - w\|_{\mathcal{H}} \\ &= \inf_{w \in \mathcal{C}} \|\lambda^* \phi(F) + (1 - \lambda^*)\phi(H) + \mu(\phi(F) - \phi(H)) - w\|_{\mathcal{H}} \\ &= \inf_{w \in \mathcal{C}} \|\phi(\lambda^* F + (1 - \lambda^*)H) + \mu(\phi(F) - \phi(H)) - w\|_{\mathcal{H}} \,. \end{aligned}$$

□

Proposition. *For all* $\lambda, \mu \geq 0$*,*

$$d(\lambda) \geq \lambda \|\phi(F) - \phi(H)\| - \sup_{w \in \mathcal{C}} \|\phi(H) - w\|, \tag{5}$$

$$d(\lambda^* + \mu) \leq \mu \|\phi(F) - \phi(H)\|, . \tag{6}$$

Proof. The proof of the first inequality above follows from applying triangle inequality to $d(.)$ from Equation (1).

The proof of the second inequality above follows from Proposition 3 by setting $h = \phi(G)$. □

B. Proof of Lemma 5

Lemma. *Let the kernel* k *be such that* $k(x, x) \leq 1$ *for all* $x \in \mathcal{X}$*. Let* $\delta \in (0, 1/4]$*. We have that, the following holds with probability* $1 - 4\delta$ *(over the sample* $x_1, \ldots, x_{n+m}$*) if* $n > 2(\lambda^*)^2 \log\left(\frac{1}{\delta}\right)$*.*

$$\begin{aligned} \|\phi(F) - \phi(\widehat{F})\|_{\mathcal{H}} &\leq \frac{3\sqrt{\log(1/\delta)}}{\sqrt{n}}, \\ \|\phi(H) - \phi(\widehat{H})\|_{\mathcal{H}} &\leq \frac{3\sqrt{\log(1/\delta)}}{\sqrt{m}}, \\ \|\phi(G) - \phi(\widehat{G})\|_{\mathcal{H}} &\leq \frac{3\sqrt{\log(1/\delta)}}{\sqrt{n/(2\lambda^*)}} \,. \end{aligned}$$

The proof for the first two statements is a direct application of Theorem 2 of Smola et al. (Smola et al., 2007), along with bounds on the Rademacher complexity. The proof of the third statement also uses Hoeffding's inequality to show that out of the n samples drawn from F, at least $n/(2\lambda^*)$ samples are drawn from G.

Lemma 14. *Let the kernel* k *be such that* $k(x, x) \leq 1$ *for all* $x \in \mathcal{X}$*. Then we have the following*

1. *For all* $h \in \mathcal{H}$ *such that* $\|h\|_{\mathcal{H}} \leq 1$ *we have that* $\sup_{x \in \mathcal{X}} |h(x)| \leq 1$*.*
2. *For all distributions* P *over* $\mathcal{X}$*, the Rademacher complexity of* $\mathcal{H}$ *is bounded above as follows:*

$$R_n(\mathcal{H}, P) = \frac{1}{n} \mathbf{E}_{x_1, \ldots, x_n \sim P} \mathbf{E}_{\sigma_1, \ldots, \sigma_n} \left[\sup_{h : \|h\|_{\mathcal{H}} \leq 1} \left| \sum_{i=1}^n \sigma_i h(x_i) \right| \right] \leq \frac{1}{\sqrt{n}} \,.$$

Proof. The first item simply follows from Cauchy-Schwarz and the reproducing property of $\mathcal{H}$

$$|h(x)| = |\langle h, k(x, .) \rangle| \leq \|h\|_{\mathcal{H}} \|k(x, .)\|_{\mathcal{H}} \leq 1 \,.$$

The second item is also a standard result and follows from the reproducing property and Jensen's inequality.

$$\frac{1}{n} \mathbf{E}_{\sigma_1, \ldots, \sigma_n} \left[\sup_{h : \|h\|_{\mathcal{H}} \leq 1} \left| \sum_{i=1}^n \sigma_i h(x_i) \right| \right] = \frac{1}{n} \mathbf{E}_{\sigma_1, \ldots, \sigma_n} \left[\sup_{h : \|h\|_{\mathcal{H}} \leq 1} \left| \langle \sum_{i=1}^n \sigma_i k(x_i, .), h \rangle \right| \right]$$

$$
\begin{aligned}
&= \frac{1}{n}\mathbf{E}_{\sigma_1,\ldots,\sigma_n}\left[\left\|\sum_{i=1}^{n}\sigma_i k(x_i,.)\right\|\right] \\
&\leq \frac{1}{n}\sqrt{\mathbf{E}_{\sigma_1,\ldots,\sigma_n}\left[\left\|\sum_{i=1}^{n}\sigma_i k(x_i,.)\right\|^2\right]} \\
&= \frac{1}{n}\sqrt{\mathbf{E}_{\sigma_1,\ldots,\sigma_n}\left[\sum_{i=1}^{n}k(x_i,x_i)\right]} \\
&\leq \frac{1}{\sqrt{n}}\,.
\end{aligned}
$$

□

Theorem 15. *(Smola et al., 2007) Let $\delta \in (0, 1/4]$. Let all $h \in \mathcal{H}$ with $\|h\|_{\mathcal{H}} \leq 1$ be such that $\sup_{x\in\mathcal{X}} |h(x)| \leq R$. Let $\widehat{P}$ be the empirical distribution induced by n i.i.d. samples from a distribution. Then with probability at least $1-\delta$*

$$\|\phi(P) - \phi(\widehat{P})\| \leq 2R_n(\mathcal{H}, P) + R\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{n}}\,.$$

Lemma 16. *Let $\delta \in (0, 1/4]$. Let $n > 2(\lambda^*)^2 \log\left(\frac{1}{\delta}\right)$. Then with at least probability $1-\delta$ the following holds. At least $\frac{n}{2\lambda^*}$ of the n samples $x_1, \ldots, x_n$ drawn from F (which is a mixture of G and H) are drawn from G.*

Proof. For all $1 \leq i \leq n$ let

$$z_i = \begin{cases} 1 & \text{if } x_i \text{ is drawn from } G \\ 0 & \text{otherwise} \end{cases}\,.$$

From the definition of F, we have that z_i are i.i.d. Bernoulli random variables with a bias of $\frac{1}{\lambda^*}$. Therefore by Hoeffding's inequality we have that,

$$
\begin{aligned}
Pr\left(\sum_{i=1}^{n} z_i > \frac{n}{2\lambda^*}\right) &= Pr\left(\frac{1}{n}\sum_{i=1}^{n} z_i - \frac{1}{\lambda^*} > \frac{-1}{2\lambda^*}\right) \\
&= 1 - Pr\left(\frac{1}{n}\sum_{i=1}^{n} z_i - \frac{1}{\lambda^*} \leq \frac{-1}{2\lambda^*}\right) \\
&\geq 1 - e^{-\frac{2n}{(2\lambda^*)^2}} \geq 1-\delta\,.
\end{aligned}
$$

□

Proof. (*Proof of Lemma 5*) From Theorem 15 and Lemma 14, we have that with probability $1-\delta$

$$\|\phi(F) - \phi(\widehat{F})\|_{\mathcal{H}} \leq 2\frac{1}{\sqrt{n}} + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{n}}\,.$$

We also have that with probability $1-\delta$

$$\|\phi(H) - \phi(\widehat{H})\|_{\mathcal{H}} \leq 2\frac{1}{\sqrt{m}} + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{m}}\,.$$

Let n' be the number of samples in $x_1, \ldots, x_n$ drawn from G. From Lemma 16, we have that with probability $1-\delta$ the $n' \geq \frac{n}{2\lambda^*}$.

We also have that with probability $1-\delta$

$$\|\phi(G)-\phi(\widehat{G})\|_{\mathcal{H}} \leq 2\frac{1}{\sqrt{n'}} + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{n'}}\,.$$

Putting the above four $1-\delta$ probability events together completes the proof. □

C. Proofs of Lemmas 6 and 7

Lemma. *Let $k(x,x)\leq 1$ for all $x\in\mathcal{X}$. Assume E_δ. For all $\lambda\in[1,\lambda^*]$ we have that*

$$\widehat{d}(\lambda) \leq \left(2-\frac{1}{\lambda^*}+\frac{\sqrt{2}}{\sqrt{\lambda^*}}\right)\lambda\cdot\frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}$$

Proof. For any $\lambda\in[1,\lambda^*]$, let $w_\lambda = \frac{\lambda}{\lambda^*}\phi(\widehat{G}) + (1-\frac{\lambda}{\lambda^*})\phi(\widehat{H}) \in \mathcal{C}_S$.

$$\begin{aligned}
\widehat{d}(\lambda) &= \inf_{w\in\mathcal{C}_S}\|\lambda\phi(\widehat{F})+(1-\lambda)\phi(\widehat{H})-w\|_{\mathcal{H}}\\
&\leq \inf_{w\in\mathcal{C}_S}\|\lambda\phi(F)+(1-\lambda)\phi(H)-w\|_{\mathcal{H}} + (2\lambda-1)\cdot\frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}\\
&= \inf_{w\in\mathcal{C}_S}\|\phi(G)+(\lambda-\lambda^*)(\phi(F)-\phi(H))-w\|_{\mathcal{H}} + (2\lambda-1)\cdot\frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}\\
&= \inf_{w\in\mathcal{C}_S}\left\|\phi(G)+\frac{\lambda-\lambda^*}{\lambda^*}(\phi(G)-\phi(H))-w\right\|_{\mathcal{H}} + (2\lambda-1)\cdot\frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}\\
&\leq \left\|\phi(G)+\frac{\lambda-\lambda^*}{\lambda^*}(\phi(G)-\phi(H))-w_\lambda\right\|_{\mathcal{H}} + (2\lambda-1)\cdot\frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}\\
&= \left\|\frac{\lambda}{\lambda^*}(\phi(G)-\phi(\widehat{G}))+\left(1-\frac{\lambda}{\lambda^*}\right)(\phi(H)-\phi(\widehat{H}))\right\|_{\mathcal{H}} + (2\lambda-1)\cdot\frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}\\
&\leq \frac{\lambda}{\lambda^*}\|(\phi(G)-\phi(\widehat{G}))\|_{\mathcal{H}}+\left(1-\frac{\lambda}{\lambda^*}\right)\|(\phi(H)-\phi(\widehat{H}))\|_{\mathcal{H}} + (2\lambda-1)\cdot\frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}\\
&\leq \frac{\lambda}{\lambda^*}\frac{3\sqrt{\log(1/\delta)}}{\sqrt{n/(2\lambda^*)}}+\left(1-\frac{\lambda}{\lambda^*}\right)\frac{3\sqrt{\log(1/\delta)}}{\sqrt{m}} + (2\lambda-1)\cdot\frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}\\
&\leq \frac{\lambda}{\lambda^*}\sqrt{2\lambda^*}\frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}+\left(1-\frac{\lambda}{\lambda^*}\right)\frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}} + (2\lambda-1)\cdot\frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}\\
&= \left(\frac{\sqrt{2}}{\sqrt{\lambda^*}}\lambda+1-\frac{\lambda}{\lambda^*}+2\lambda-1\right)\frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}\\
&= \left(2-\frac{1}{\lambda^*}+\frac{\sqrt{2}}{\sqrt{\lambda^*}}\right)\lambda\cdot\frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}\,.
\end{aligned}$$

□

Lemma. *Let $k(x,x)\leq 1$ for all $x\in\mathcal{X}$. Assume E_δ. For all $\lambda\geq 1$, we have*

$$\widehat{d}(\lambda) \geq d(\lambda) - (2\lambda-1)\cdot\frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}\,.$$

Proof.

$$\widehat{d}(\lambda) = \inf_{w\in\mathcal{C}_S}\|\lambda\phi(\widehat{F})+(1-\lambda)\phi(\widehat{H})-w\|_{\mathcal{H}}$$

$$
\begin{aligned}
&\geq \inf_{w \in \mathcal{C}_S} \|\lambda\phi(F) + (1-\lambda)\phi(H) - w\|_{\mathcal{H}} - \lambda\|\phi(\widehat{F}) - \phi(F)\|_{\mathcal{H}} - (\lambda - 1)\|\phi(H) - \phi(\widehat{H})\|_{\mathcal{H}} \\
&\geq d(\lambda) - \lambda \cdot \frac{3\sqrt{\log(1/\delta)}}{\sqrt{n}} - (\lambda - 1) \cdot \frac{3\sqrt{\log(1/\delta)}}{\sqrt{m}} \\
&\geq d(\lambda) - (2\lambda - 1) \cdot \frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}} \,.
\end{aligned}
$$

□

D. Proof of Theorem 10

Theorem. *Let the kernel k, and distributions G, H satisfy the separability condition with margin $\alpha > 0$ and tolerance β. Then $\forall \mu > 0$*

$$d(\lambda^* + \mu) \geq \frac{\alpha\mu}{\lambda^*} - \beta \,.$$

Proof. Let $g \in \mathcal{H}$ be the witness to the separability condition – (i.e.) $\|g\|_{\mathcal{H}} \leq 1$ and $\mathbf{E}_{X\sim G} g(X) \leq \inf_x g(x) + \beta \leq \mathbf{E}_{X\sim H} g(X) + \alpha$. Let $\Delta_{\mathcal{X}}$ denote the set of all probability distributions over $\mathcal{X}$. One can show that

$$
\begin{aligned}
d(\lambda^* + \mu) &= \inf_{w \in \mathcal{C}} \|\phi(G) + \mu(\phi(F) - \phi(H)) - w\|_{\mathcal{H}} \\
&= \inf_{P \in \Delta_{\mathcal{X}}} \|\phi(G) + \frac{\mu}{\lambda^*}(\phi(G) - \phi(H)) - \phi(P)\|_{\mathcal{H}} \\
&= \inf_{P \in \Delta_{\mathcal{X}}} \sup_{h \in \mathcal{H}: \|h\| \leq 1} \left\langle \phi(P) + \frac{\mu}{\lambda^*}(\phi(H) - \phi(G)) - \phi(G), h \right\rangle \\
&= \inf_{P \in \Delta_{\mathcal{X}}} \sup_{h \in \mathcal{H}: \|h\| \leq 1} \mathbf{E}_P[h(X)] - \mathbf{E}_G[h(X)] + \frac{\mu}{\lambda^*}(\mathbf{E}_H[h(X)] - \mathbf{E}_G[h(X)]) \\
&\geq \inf_{P \in \Delta_{\mathcal{X}}} \mathbf{E}_P[g(X)] + \frac{\mu}{\lambda^*}\mathbf{E}_H[g(X)] - \left(1 + \frac{\mu}{\lambda^*}\right) \mathbf{E}_G[g(X)]) \\
&\geq \inf_x g(x) + \frac{\mu}{\lambda^*}(\alpha) - (\inf_x g(x) + \beta) \\
&= \frac{\alpha\mu}{\lambda^*} - \beta \,.
\end{aligned}
$$

□

E. Proof of Theorem 11

Theorem. *Let the kernel $k : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ be universal. Let the distributions G, H be such that they satisfy the anchor set condition with margin $\gamma > 0$ for some family of subsets of $\mathcal{X}$. Then, for all $\theta > 0$, there exists a $\beta > 0$ such that the kernel k, and distributions G, H satisfy the separability condition with margin $\beta\theta$ and tolerance β, i.e.*

$$\mathbf{E}_{X\sim G} h(X) \leq \inf_x h(x) + \beta \leq \mathbf{E}_{X\sim H} h(X) - \beta\theta$$

Proof. Fix some $\theta > 0$. Let $A \subseteq \mathcal{X}$ be the witness to the anchor set condition, i.e., A is a compact set such that $A \subseteq \text{supp}(H) \setminus \text{supp}(G)$ and $H(A) \geq \gamma$. A is a compact (and hence closed) set that is disjoint from $\text{supp}(G)$ (which is a closed, compact set), hence there exists a continuous function $f : \mathcal{X} \to \mathbb{R}$ such that,

$$
\begin{aligned}
&f(x) \geq 0, \forall x \in \mathcal{X}, \\
&f(x) = 0, \forall x \in \text{supp}(G), \\
&f(x) \geq 1, \forall x \in A.
\end{aligned}
$$

By universality of the kernel k, we have that

$$\forall \epsilon > 0, \exists h_\epsilon \in \mathcal{H}, \text{ s.t. } \sup_{x \in \mathcal{X}} |f(x) - h_\epsilon(x)| \leq \epsilon.$$

We then have the following:

$$\mathbf{E}_G h_\epsilon(X) \le \epsilon, \tag{7}$$

$$\inf_{x\in\mathcal{X}} h_\epsilon(x) \le \epsilon, \tag{8}$$

$$\inf_{x\in\mathcal{X}} h_\epsilon(x) \ge -\epsilon, \tag{9}$$

$$\inf_{x\in A} h_\epsilon(x) \ge 1-\epsilon, \tag{10}$$

$$\begin{aligned}\mathbf{E}_H h_\epsilon(X) &\ge (-\epsilon)(1-H(A)) + (1-\epsilon)H(A)\\ &\ge \gamma - \epsilon.\end{aligned} \tag{11}$$

From Equations (7), (8), (9) and (11), we have that

$$\mathbf{E}_G h_\epsilon(X) \le \epsilon \le \inf_x h_\epsilon(x) + 2\epsilon \le 3\epsilon \le \mathbf{E}_H h_\epsilon(X) - (\gamma - 4\epsilon).$$

Let $\overline{h}_\epsilon = h_\epsilon/\|h_\epsilon\|_\mathcal{H}$ be the normalized version of h_ϵ. We then have that

$$\mathbf{E}_G \overline{h}_\epsilon(X) \le \inf_x \overline{h}_\epsilon(x) + \frac{2\epsilon}{\|h_\epsilon\|_\mathcal{H}} \le \mathbf{E}_H \overline{h}_\epsilon(X) - \frac{\gamma - 4\epsilon}{\|h_\epsilon\|_\mathcal{H}}.$$

Setting $\epsilon = \frac{\gamma}{2\theta+4}$ and $\beta = \frac{2\gamma}{(2\theta+4)\|h_{\gamma/(2\theta+4)}\|_\mathcal{H}}$ we get that there exists $h \in \mathcal{H}$ such that $\|h\|_\mathcal{H} \le 1$ and

$$\mathbf{E}_G h(X) \le \inf_x h(x) + \beta \le \mathbf{E}_H h(X) - \beta\theta.$$

□

F. Proof of Theorem 12

Theorem. *Let $\delta \in (0, \frac{1}{4}]$. Let $k(x,x) \le 1$ for all $x \in \mathcal{X}$. Let the kernel k, and distributions G, H satisfy the separability condition with tolerance β and margin $\alpha > 0$. Let the number of samples be large enough such that $\min(m,n) > \frac{(12\cdot\lambda^*)^2 \log(1/\delta)}{\alpha^2}$. Let the threshold τ be such that $\frac{3\lambda^*\sqrt{\log(1/\delta)}(2-1/\lambda^*+\sqrt{2/\lambda^*})}{\sqrt{\min(m,n)}} \le \tau \le \frac{6\lambda^*\sqrt{\log(1/\delta)}(2-1/\lambda^*+\sqrt{2/\lambda^*})}{\sqrt{\min(m,n)}}$. We then have with probability $1-4\delta$*

$$\begin{aligned}\lambda^* - \widehat{\lambda}^V_\tau &\le 0,\\ \widehat{\lambda}^V_\tau - \lambda^* &\le \frac{\beta\lambda^*}{\alpha} + c\cdot\sqrt{\log(1/\delta)}\cdot(\min(m,n))^{-1/2},\end{aligned}$$

for constant $c = \left(\frac{6\alpha(\lambda^*)^2(2-1/\lambda^*+\sqrt{2/\lambda^*})+2\lambda^*(3\alpha+6\lambda^*(2+\alpha+\beta))}{\alpha^2}\right)$.

Lemma 17. *Let $k(x,x) \le 1$ for all $x \in \mathcal{X}$. Let the kernel k, and distributions G, H satisfy the separability condition with margin α and tolerance β. Assume E_δ. Then*

$$\begin{aligned}\widehat{d}(\lambda) &\le \left(2 - \frac{1}{\lambda^*} + \frac{\sqrt{2}}{\sqrt{\lambda^*}}\right)\lambda \cdot \frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}, \qquad \forall\lambda\in[1,\lambda^*],\\ \widehat{d}(\lambda) &\ge \frac{(\lambda-\lambda^*)\alpha}{\lambda^*} - \beta - (2\lambda-1)\cdot\frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}, \qquad \forall\lambda\in[\lambda^*,\infty).\end{aligned}$$

Proof. The proof follows from Lemmas 7, 6 and Theorem 10. The upper bound forms the line $(\lambda, U(\lambda))$ and the lower bound forms the line $(\lambda, L(\lambda))$ in Figure 1a. □

Lemma 18. *Let $k(x,x) \le 1$ for all $x \in \mathcal{X}$. Let the kernel k, and distributions G, H satisfy the separability condition with margin α and tolerance β. Assume E_δ. We then have*

$$\widehat{\lambda}^V_\tau \ge \min\left(\lambda^*, \frac{\tau\sqrt{\min(m,n)}}{3\sqrt{\log(1/\delta)}(2-1/\lambda^*+\sqrt{2/\lambda^*})}\right), \tag{12}$$

$$\widehat{\lambda}_\tau^V \le \lambda^* \cdot \frac{(\tau+\beta+\alpha)\sqrt{\min(m,n)} + 3\sqrt{\log(1/\delta)}}{\alpha\sqrt{\min(m,n)} - 6\lambda^*\sqrt{\log(1/\delta)}} \,. \tag{13}$$

Proof. As $\widehat{d}$ is a continuous function, we have that $\widehat{d}(\widehat{\lambda}_\tau^V) = \tau$. If $\widehat{\lambda}_\tau^V \le \lambda^*$, we have from Lemma 17 that

$$\tau \le \left(2 - \frac{1}{\lambda^*} + \frac{\sqrt{2}}{\sqrt{\lambda^*}}\right)\widehat{\lambda}_\tau^V \cdot \frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}},$$

and hence

$$\widehat{\lambda}_\tau^V \ge \min\left(\lambda^*, \frac{\tau\sqrt{\min(m,n)}}{3\sqrt{\log(1/\delta)}(2 - 1/\lambda^* + \sqrt{2/\lambda^*})}\right) .$$

If $\widehat{\lambda}_\tau^V \ge \lambda^*$, we have

$$\begin{aligned}\tau &\ge \frac{(\widehat{\lambda}_\tau^V - \lambda^*)\alpha}{\lambda^*} - \beta - (2\widehat{\lambda}_\tau^V - 1)\cdot\frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}, \\ &= \widehat{\lambda}_\tau^V\left(\frac{\alpha}{\lambda^*} - \frac{6\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}\right) - \alpha - \beta - \frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}} .\end{aligned}$$

Rearranging terms, we have that if $\widehat{\lambda}_\tau^V \ge \lambda^*$, then

$$\widehat{\lambda}_\tau^V \le \frac{\tau + \alpha + \beta + \frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}}{\frac{\alpha}{\lambda^*} - \frac{6\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}} .$$

And thus

$$\begin{aligned}\widehat{\lambda}_\tau^V &\le \max\left(\lambda^*, \lambda^*\cdot\frac{(\tau+\beta+\alpha)\sqrt{\min(m,n)} + 3\sqrt{\log(1/\delta)}}{\alpha\sqrt{\min(m,n)} - 6\lambda^*\sqrt{\log(1/\delta)}}\right) \\ &= \lambda^*\cdot\frac{(\tau+\beta+\alpha)\sqrt{\min(m,n)} + 3\sqrt{\log(1/\delta)}}{\alpha\sqrt{\min(m,n)} - 6\lambda^*\sqrt{\log(1/\delta)}} .\end{aligned}$$

□

Proof. (Proof of Theorem 12)

As $\min(m,n) > \frac{(12\cdot\lambda^*)^2\log(1/\delta)}{\alpha^2} > 2(\lambda^*)^2\log(1/\delta)$, we have that E_δ is $1-4\delta$ probability event. Assume E_δ.

As $\tau \ge \frac{3\lambda^*\sqrt{\log(1/\delta)}(2-1/\lambda^*+\sqrt{2/\lambda^*})}{\sqrt{\min(m,n)}}$, we have from Equation (12)

$$\widehat{\lambda}_\tau^V \ge \lambda^* .$$

From Equation (13), we have

$$\begin{aligned}\widehat{\lambda}_\tau^V &\le \lambda^*\cdot\frac{(\tau+\alpha+\beta)\sqrt{\min(m,n)} + 3\sqrt{\log(1/\delta)}}{\alpha\sqrt{\min(m,n)} - 6\lambda^*\sqrt{\log(1/\delta)}} \\ &= \lambda^*\left(\frac{\tau+\beta+\alpha}{\alpha} + \frac{(3 + \frac{6\lambda^*(\tau+\alpha+\beta)}{\alpha})\sqrt{\log(1/\delta)}}{\alpha\sqrt{\min(m,n)} - 6\lambda^*\sqrt{\log(1/\delta)}}\right) \\ &\le \lambda^*\left(1 + \frac{\beta}{\alpha}\right) + \frac{\tau\lambda^*}{\alpha} + \frac{2\lambda^*(3 + \frac{6\lambda^*(\tau+\alpha+\beta)}{\alpha})\sqrt{\log(1/\delta)}}{\alpha\sqrt{\min(m,n)}}\end{aligned}$$

$$\leq \lambda^* \left(1 + \frac{\beta}{\alpha}\right) + \frac{6(\lambda^*)^2\sqrt{\log(1/\delta)}(2 - 1/\lambda^* + \sqrt{2/\lambda^*})}{\alpha\sqrt{\min(m,n)}} + \frac{2\lambda^*(3\alpha + 6\lambda^*(\tau+\alpha+\beta))\sqrt{\log(1/\delta)}}{\alpha^2\sqrt{\min(m,n)}}$$
$$\leq \lambda^* \left(1 + \frac{\beta}{\alpha}\right) + \left(\frac{6\alpha(\lambda^*)^2(2 - 1/\lambda^* + \sqrt{2/\lambda^*}) + 2\lambda^*(3\alpha + 6\lambda^*(2+\alpha+\beta))}{\alpha^2}\right) \cdot \sqrt{\log(1/\delta)} \cdot (\min(m,n))^{-1/2}.$$

The third line above follows, because $\min(m,n) > \frac{(12\cdot\lambda^*)^2\log(1/\delta)}{\alpha^2}$. The last two lines follow, because $\tau \leq \frac{6\lambda^*\sqrt{\log(1/\delta)}(2-1/\lambda^*+\sqrt{2/\lambda^*})}{\sqrt{\min(m,n)}}$, which in turn is upper bounded by 2 under the conditions on $\min(m,n)$.

□

G. Proof of Theorem 13

Theorem. *Let $k(x,x) \leq 1$ for all $x \in \mathcal{X}$. Let the kernel k, and distributions G, H satisfy the separability condition with tolerance β and margin $\alpha > 0$. Let $\nu \in [\frac{\alpha}{4\lambda^*}, \frac{3\alpha}{4\lambda^*}]$ and $\sqrt{\min(m,n)} \geq \frac{36\sqrt{\log(1/\delta)}}{\frac{\alpha}{\lambda^*} - \nu}$. We then have with probability $1 - 4\delta$,*

$$\lambda^* - \widehat{\lambda}^G_\nu \leq c \cdot \sqrt{\log(1/\delta)} \cdot (\min(m,n))^{-1/2},$$
$$\widehat{\lambda}^G_\nu - \lambda^* \leq \frac{4\beta\lambda^*}{\alpha} + c' \cdot \sqrt{\log(1/\delta)} \cdot (\min(m,n))^{-1/2},$$

for constants $c = (2\lambda^* - 1 + \sqrt{2\lambda^*}) \cdot \frac{12\lambda^*}{\alpha}$ and $c' = \frac{144(\lambda^*)^2(\alpha+4\beta)}{\alpha^2}$.

Lemma 19. *Let $k(x,x) \leq 1$ for all $x \in \mathcal{X}$. Let the kernel k, and distributions G, H satisfy the separability condition with margin α and tolerance β. Assume E_δ. We then have*

$$\sup\{g \in \partial\widehat{d}(\lambda)\} \leq \frac{1}{\lambda^* - \lambda} \cdot (2\lambda^* - 1 + \sqrt{2\lambda^*}) \cdot \frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}, \qquad \forall \lambda \in [1, \lambda^*],$$
$$\inf\{g \in \partial\widehat{d}(\lambda)\} \geq \left(\frac{\alpha}{\lambda^*} - \frac{\beta}{\lambda - \lambda^*} - \frac{6\lambda}{\lambda - \lambda^*} \cdot \frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}\right), \qquad \forall \lambda \in [\lambda^*, \infty).$$

Proof. As $\widehat{d}(.)$ is convex, we have that for all $\lambda \in [1, \lambda^*]$, and all $g \in \partial\widehat{d}(\lambda)$

$$g \leq \frac{\widehat{d}(\lambda^*) - \widehat{d}(\lambda)}{\lambda^* - \lambda}$$
$$\leq \frac{\widehat{d}(\lambda^*)}{\lambda^* - \lambda}.$$

Applying Lemma 6 to $\widehat{d}(\lambda^*)$, we get $\forall \lambda \in [1, \lambda^*]$

$$\sup\{g \in \partial\widehat{d}(\lambda)\} \leq \frac{1}{\lambda^* - \lambda} \cdot (2\lambda^* - 1 + \sqrt{2\lambda^*}) \cdot \frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}.$$

Once again by convexity of $\widehat{d}(.)$, we have that for all $\lambda \geq \lambda^*$ and all $g \in \partial\widehat{d}(\lambda)$

$$g \geq \frac{\widehat{d}(\lambda) - \widehat{d}(\lambda^*)}{\lambda - \lambda^*}.$$

Applying Lemma 7 and Theorem 10 to $\widehat{d}(\lambda)$ and Lemma 6 to $\widehat{d}(\lambda^*)$, we get $\forall \lambda \in [\lambda^*, \infty)$

$$\inf\{g \in \partial\widehat{d}(\lambda)\} \geq \left(\frac{\alpha}{\lambda^*} - \frac{\beta}{\lambda - \lambda^*} - \frac{2\lambda + 2\lambda^* - 2 + \sqrt{2\lambda^*}}{\lambda - \lambda^*} \cdot \frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}\right).$$

□

Lemma 20. *Let $k(x,x) \leq 1$ for all $x \in \mathcal{X}$. Let the kernel k, and distributions G, H satisfy the separability condition with margin α and tolerance β. Assume E_δ. We then have*

$$\widehat{\lambda}_\nu^G \geq \lambda^* - (2\lambda^* - 1 + \sqrt{2\lambda^*}) \cdot \frac{3\sqrt{\log(1/\delta)}}{\nu\sqrt{\min(m,n)}}, \tag{14}$$

$$\widehat{\lambda}_\nu^G \leq \lambda^* \cdot \frac{\frac{\alpha+\beta}{\lambda^*} - \nu}{\frac{\alpha}{\lambda^*} - \nu - \frac{18\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}}. \tag{15}$$

Proof. By definition of the gradient thresholding estimator $\widehat{\lambda}_\nu^G$ we have

$$\inf\{g \in \partial\widehat{d}(\widehat{\lambda}_\nu^G)\} \leq \nu \leq \sup\{g \in \partial\widehat{d}(\widehat{\lambda}_\nu^G)\}.$$

Firstly, note that $\widehat{\lambda}_\nu^G \geq 1$, because $\nu \geq \frac{\alpha}{4\lambda^*} > 0$. By Lemma 19 we have that if $\widehat{\lambda}_\nu^G \in [1, \lambda^*]$ then

$$\nu \leq \sup\{g \in \partial\widehat{d}(\widehat{\lambda}_\nu^G)\} \leq \frac{1}{\lambda^* - \widehat{\lambda}_\nu^G} \cdot (2\lambda^* - 1 + \sqrt{2\lambda^*}) \cdot \frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}. \tag{16}$$

Once again by Lemma 19, we have that if $\widehat{\lambda}_\nu^G > \lambda^*$ then

$$\nu \geq \inf\{g \in \partial\widehat{d}(\widehat{\lambda}_\nu^G)\} \geq \left(\frac{\alpha}{\lambda^*} - \frac{\beta}{\widehat{\lambda}_\nu^G - \lambda^*} - \frac{6\widehat{\lambda}_\nu^G}{\widehat{\lambda}_\nu^G - \lambda^*} \cdot \frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}\right). \tag{17}$$

Rearranging Equation (16), we get that if $\widehat{\lambda}_\nu^G \in [1, \lambda^*]$ then

$$\widehat{\lambda}_\nu^G \geq \lambda^* - (2\lambda^* - 1 + \sqrt{2\lambda^*}) \cdot \frac{3\sqrt{\log(1/\delta)}}{\nu\sqrt{\min(m,n)}}.$$

Hence

$$\widehat{\lambda}_\nu^G \geq \min\left(\lambda^*, \lambda^* - (2\lambda^* - 1 + \sqrt{2\lambda^*}) \cdot \frac{3\sqrt{\log(1/\delta)}}{\nu\sqrt{\min(m,n)}}\right) = \lambda^* - (2\lambda^* - 1 + \sqrt{2\lambda^*}) \cdot \frac{3\sqrt{\log(1/\delta)}}{\nu\sqrt{\min(m,n)}}.$$

Rearranging Equation (17), we get that if $\widehat{\lambda}_\nu^G > \lambda^*$, then

$$\begin{aligned}
\frac{\alpha}{\lambda^*} - \nu &\leq \left(\frac{6\widehat{\lambda}_\nu^G}{\widehat{\lambda}_\nu^G - \lambda^*} \cdot \frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}} + \frac{\beta}{\widehat{\lambda}_\nu^G - \lambda^*}\right) \\
\left(\widehat{\lambda}_\nu^G - \lambda^*\right)\left(\frac{\alpha}{\lambda^*} - \nu\right) &\leq \left(6\widehat{\lambda}_\nu^G\right) \cdot \frac{3\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}} + \beta \\
\left(\widehat{\lambda}_\nu^G\right)\left(\frac{\alpha}{\lambda^*} - \nu - \frac{18\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}\right) &\leq \lambda^*\left(\frac{\alpha}{\lambda^*} - \nu\right) + \beta \\
\widehat{\lambda}_\nu^G &\leq \frac{\lambda^*\left(\frac{\alpha+\beta}{\lambda^*} - \nu\right)}{\frac{\alpha}{\lambda^*} - \nu - \frac{18\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}}.
\end{aligned}$$

Thus we have

$$\widehat{\lambda}_\nu^G \leq \max\left(\lambda^*, \frac{\lambda^*\left(\frac{\alpha+\beta}{\lambda^*} - \nu\right)}{\frac{\alpha}{\lambda^*} - \nu - \frac{18\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}}\right) = \frac{\lambda^*\left(\frac{\alpha+\beta}{\lambda^*} - \nu\right)}{\frac{\alpha}{\lambda^*} - \nu - \frac{18\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}}.$$

□

Proof. (Proof of Theorem 13)

As $(\frac{\alpha}{\lambda^*} - \nu)\sqrt{\min(m,n)} \geq 36\sqrt{\log(1/\delta)}$, we have that

$$\min(m,n) \geq \frac{(36\lambda^*)^2 \log(1/\delta)}{\alpha^2} \geq 2(\lambda^*)^2 \log(1/\delta),$$

and hence E_δ is a $1-4\delta$ probability event. Assume E_δ.

Equation (14) immediately gives

$$\begin{aligned}
\lambda^* - \widehat{\lambda}_\nu^G &\leq (2\lambda^* - 1 + \sqrt{2\lambda^*}) \cdot \frac{3}{\nu} \cdot \sqrt{\log(1/\delta)} \cdot (\min(m,n))^{-1/2} \\
&\leq (2\lambda^* - 1 + \sqrt{2\lambda^*}) \cdot \frac{12\lambda^*}{\alpha} \cdot \sqrt{\log(1/\delta)} \cdot (\min(m,n)^{-1/2}.
\end{aligned}$$

The second inequality above is due to $\nu \geq \frac{\alpha}{4\lambda^*}$.

Let $\omega = \frac{\alpha+\beta-\nu\lambda^*}{\alpha-\nu\lambda^*} \leq 1 + \frac{4\beta}{\alpha}$. Equation (15) gives

$$\begin{aligned}
\widehat{\lambda}_\nu^G &\leq \lambda^* \cdot \frac{\frac{\alpha+\beta}{\lambda^*} - \nu}{\frac{\alpha}{\lambda^*} - \nu - \frac{18\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}} \\
&= \lambda^* \cdot \frac{\omega\left(\frac{\alpha}{\lambda^*} - \nu - \frac{18\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}\right) + \omega\left(\frac{18\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}\right)}{\frac{\alpha}{\lambda^*} - \nu - \frac{18\sqrt{\log(1/\delta)}}{\sqrt{\min(m,n)}}} \\
&= \omega\lambda^* + \frac{18\omega\lambda^*\sqrt{\log(1/\delta)}}{(\frac{\alpha}{\lambda^*} - \nu)\sqrt{\min(m,n)} - 18\sqrt{\log(1/\delta)}} \\
&\leq \omega\lambda^* + \frac{36\omega\lambda^*\sqrt{\log(1/\delta)}}{(\frac{\alpha}{\lambda^*} - \nu)\sqrt{\min(m,n)}} \\
&\leq \lambda^* + \frac{4\beta\lambda^*}{\alpha} + \frac{36(1+\frac{4\beta}{\alpha})\lambda^*\sqrt{\log(1/\delta)}}{(\frac{\alpha}{4\lambda^*})\sqrt{\min(m,n)}} \\
&\leq \lambda^* + \frac{4\beta\lambda^*}{\alpha} + \frac{144(\lambda^*)^2(\alpha+4\beta)}{\alpha^2} \cdot \sqrt{\log(1/\delta)} \cdot (\min(m,n))^{-1/2}.
\end{aligned}$$

The second inequality above is due to $(\frac{\alpha}{\lambda^*} - \nu)\sqrt{\min(m,n)} \geq 36\sqrt{\log(1/\delta)}$. The third inequality above is due to $\nu \leq \frac{3\alpha}{4\lambda^*}$. □

H. Experimental Results in Table Format

	KM1	KM2	alphamax	ROC	EN
waveform(400)	0.042	**0.032**	0.089*	0.117*	0.127*
waveform(800)	0.034	**0.027**	0.06*	0.072 *	0.112*
waveform(1600)	0.021	**0.017**	0.048*	0.051*	0.115*
waveform(3200)	0.015	**0.012**	0.079*	0.045*	0.102*
mushroom(400)	0.193*	0.123*	**0.084**	0.148*	0.125
mushroom(800)	0.096*	0.129*	**0.041**	0.074*	0.066*
mushroom(1600)	0.042	0.096*	**0.039**	0.053*	0.055*
mushroom(3200)	0.039*	0.067*	**0.023**	0.024	0.035*
pageblocks(400)	0.098	0.16	0.218	0.193*	**0.078**
pageblocks(800)	**0.038**	0.088*	0.203*	0.139*	0.081*
pageblocks(1600)	**0.034**	0.056*	0.083*	0.091*	0.055*
pageblocks(3200)	**0.02**	0.033*	0.166*	0.084*	0.047*
shuttle(400)	0.072	0.129	0.122	0.107*	**0.062**
shuttle(800)	0.065	0.091	0.054	0.057	**0.046**
shuttle(1600)	0.035	0.03	0.03	0.049*	**0.027**
shuttle(3200)	0.023*	**0.014**	0.02	0.041*	0.025*
spambase(400)	**0.086**	0.111	0.097	0.229*	0.186*
spambase(800)	0.079	**0.067**	0.096*	0.166*	0.171*
spambase(1600)	0.059	**0.043**	0.07*	0.092*	0.139*
spambase(3200)	0.032	**0.028**	0.063*	0.067*	0.129*
digits(400)	0.24*	**0.091**	0.115	0.186*	0.136
digits(800)	0.127*	**0.071**	0.073	0.113*	0.114*
digits(1600)	0.083*	0.034	**0.03**	0.071*	0.111*
digits(3200)	0.055*	**0.025**	0.031	0.046*	0.085*

Table 2. Average absolute error incurred in predicting the mixture proportion κ^*. The first column gives the dataset and the total number of samples used (mixture and component) in parantheses. The best performing algorithm for each dataset and sample size is highlighted in bold. Algorithms whose performances have been identified as significantly inferior to the best algorithm, by the Wilcoxon signed rank test (at significance level $p = 0.05$), are marked with a star.