Mixture Density Estimation via Hilbert Space Embedding of Measures

Bharath K. Sriperumbudur
Gatsby Computational Neuroscience Unit
University College London
Alexandra House, 17 Queen Square
London WC1N 3AR
bharath@gatsby.ucl.ac.uk

***Abstract*—In this paper, we consider the problem of estimating a density using a finite combination of densities from a given class, $\mathcal{C}$. Unlike previous works, where Kullback-Leibler (KL) divergence is used as a notion of distance, in this paper, we consider a distance measure based on the embedding of densities into a reproducing kernel Hilbert space (RKHS). We analyze the estimation and approximation errors for an M-estimator and show the estimation error rate to be *better* than that obtained with KL divergence while achieving the *same* approximation error rate. Another advantage of the Hilbert space embedding approach is that these results are achieved without making any assumptions on $\mathcal{C}$, in contrast to the KL divergence approach, where the densities in $\mathcal{C}$ are assumed to be bounded (and away from zero) with $\mathcal{C}$ having a finite Dudley entropy integral.**

I. Introduction

The problem of density estimation deals with estimating an unknown density, f, given an i.i.d. sample, $S := \{X_j\}_{j=1}^n$ drawn from it. One popular approach is parametric estimation, where a particular parametric form is assumed for f and the parameters are then estimated using the sample, S. In this paper, we deal with mixture density estimation, which is elaborated below.

Consider a parametric family of probability density functions,

$$\mathcal{C} := \left\{\phi_\theta(x) : \theta \in \Theta \subset \mathbb{R}^d\right\}$$

over a measurable space, $(\mathcal{X}, \mathscr{A})$ with a base σ-finite measure μ defined on $\mathscr{A}$ (whenever we mention that a probability measure on $\mathscr{A}$ has a density, we mean it has a Radon-Nikodym derivative with respect to μ). The class of k-component mixtures g_k is defined as

$$\mathcal{G}_k := \left\{ g_k(x) = \sum_{j=1}^k \lambda_j \phi_{\theta_j}(x),\, (\lambda)_k \in \Delta_k,\, (\theta)_k \subset \Theta \right\},$$

where

$$\Delta_k := \left\{ (\lambda_1, \ldots, \lambda_k) : \sum_{j=1}^k \lambda_j = 1,\, \lambda_j \geq 0 \right\},$$

$(\lambda)_k := (\lambda_1, \ldots, \lambda_k)$ and $(\theta)_k := (\theta_1, \ldots, \theta_k)$. Now given S, the goal is to estimate $(\lambda)_k$ and $(\theta)_k$ so that g_k is a good approximation to f. Define the KL divergence between g_k and f as

$$D(f\|g_k) := \int_{\mathcal{X}} f(x) \log \frac{f(x)}{g_k(x)}\, d\mu(x).$$

The popular maximum likelihood estimator (MLE) is obtained by minimizing the KL divergence between g_k and S, given by

$$D(S\|g_k) := \frac{1}{n}\sum_{i=1}^n \log \frac{f(X_i)}{g_k(X_i)},$$

i.e., $g_k^{\text{mle}} = \arg\min_{g_k \in \mathcal{G}_k} D(S\|g_k) = \sum_{j=1}^k \lambda_j^* \phi_{\theta_j^*}(x)$, where

$$((\lambda^*)_k, (\theta^*)_k) = \arg \max_{(\lambda)_k \in \Delta_k, (\theta)_k \subset \Theta} \sum_{i=1}^n \log\Big(\sum_{j=1}^k \lambda_j \phi_{\theta_j}(X_i)\Big).$$

Let us consider the convex hull of $\mathcal{C}$ defined as

$$\mathcal{G} = \left\{ g(x) = \int_\Theta \phi_\theta(x)\, d\mathbb{P}(\theta),\, \mathbb{P} \in M_+^1(\Theta) \right\},$$

which is the class of continuous convex combinations of densities in $\mathcal{C}$ and $M_+^1(\Theta)$ is the set of probability measures on Θ. Under the assumption that $\sup_{\theta,\theta',x} \log \frac{\phi_\theta(x)}{\phi_{\theta'}(x)} < \infty$, which is satisfied if $0 < a \leq \phi_\theta(x) \leq b < \infty$, $\forall \theta \in \Theta, x \in \mathcal{X}$, Li and Barron [1], [2] showed that for any f, there exists $g_k \in \mathcal{G}_k$ such that

$$D(f\|g_k) \leq \inf_{g \in \mathcal{G}} D(f\|g) + O\left(\frac{1}{k}\right), \tag{1}$$

where the constants (which depend only on a and b) are absorbed in the order notation. In particular, they showed that $g_k \in \mathcal{G}_k$ satisfying (1) can be constructed by the following greedy procedure: Initialize $g_1 = \phi_\theta$ to minimize $D(f\|g_1)$ and at step k construct g_k from g_{k-1} by finding α and θ such that

$$D(f\|g_k) \leq \min_{\alpha,\theta} D(f\|(1-\alpha)g_{k-1}(x) + \alpha\phi_\theta(x)).$$

Based on the above greedy algorithm, one can estimate f greedily from S as g_k^{gre} by choosing ϕ_θ at step k so that

$$\sum_{i=1}^n \log g_k^{\text{gre}}(X_i) \geq \max_{\alpha,\theta} \sum_{i=1}^n \log((1-\alpha)g_k^{\text{gre}}(X_i) + \alpha\phi_\theta(X_i)).$$

Note that the approximation bound in (1) also holds for the maximum likelihood estimator. However, the advantage with the greedy approach over MLE is that only two parameters are optimized at a time instead of $2k$ parameters, therefore reducing the complexity of the optimization problem.

Given g_k^{mle} and g_k^{gre}, Rakhlin et al. [3] improved on the results of [1] and [2] by showing that for any $g_k \in \mathcal{G}_k$ and for any f,

$$D(f\|\hat{g}_k) - D(f\|g_k) \leq \frac{c_1}{k} + \frac{c_2}{\sqrt{n}} + c_3 \int_0^b \sqrt{\frac{\log \mathcal{N}(\mathcal{C}, \epsilon, d_n)}{n}}\, d\epsilon,$$

where c_1, c_2 and c_3 are some constants, $\hat{g}_k$ is either g_k^{mle} or g_k^{gre}, $d_n^2(\phi_1, \phi_2) := n^{-1} \sum_{j=1}^n (\phi_1(X_j) - \phi_2(X_j))^2$ and $\mathcal{N}(\mathcal{C}, \epsilon, d_n)$ represents the ϵ-covering number of $\mathcal{C}$. Therefore, if the above integral is finite and is of the order of $n^{-1/2}$, we have

$$D(f\|\hat{g}_k) - D(f\|g_k) \leq O\left(\frac{1}{k}\right) + O_f\left(\frac{1}{\sqrt{n}}\right). \tag{2}$$

Combining (1), which characterizes the approximation error, and (2), which characterizes the estimation error, we have

$$D(f\|\hat{g}_k) - \inf_{g \in \mathcal{G}} D(f\|g) \leq O\left(\frac{1}{k}\right) + O_f\left(\frac{1}{\sqrt{n}}\right), \tag{3}$$

which shows that as $n \to \infty$ and $k \to \infty$, the approximation to f in terms of g_k^{mle} or g_k^{gre} approaches the best possible approximation to f in $\mathcal{G}$. Overall, though (3) is a nice result, it assumes that f and ϕ_θ are bounded (and away from zero) and the class $\mathcal{C}$ is not large enough so that the entropy integral (shown above) is finite. To avoid problems near zero (which is why the densities are assumed to be bounded away from zero), [4] considered Hellinger distance to analyze the rates of convergence for MLE in mixture models. However, the bounds depend on the entropy integral of $\mathcal{G}$, which is undesirable as $\mathcal{G}$ is a much larger class than $\mathcal{C}$ and therefore the entropy integral associated with $\mathcal{G}$ need not exist even if the one associated with $\mathcal{C}$ exists.

To address these issues, in this paper, we propose to use a distance metric (we refer to it as the *kernel distance*) based on the notion of embedding probability measures into a reproducing kernel Hilbert space (RKHS) [5]–[8]. Using this metric, we consider an M-estimator for which we show in Section III that without requiring any assumptions on f and $\mathcal{C}$, a *faster* rate of $O_f(n^{-1/2})$ for the estimation error and a *similar* rate of $O(k^{-1/2})$ for the approximation error can be achieved (see footnote 1 for an explanation about the comparison of these rates). Before presenting our results, we provide a brief review of RKHS and the notion of embedding measures into RKHS in Section II.

II. RKHS Embedding of Probability Measures

In this section, we provide a brief overview of RKHS and the notion of embedding probability measures into an RKHS. First, we start with the definition of an RKHS, which we quote from [6].

Definition 1 (Reproducing kernel Hilbert space): A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $(x, y) \mapsto K(x, y)$ is a *reproducing kernel* of the Hilbert space $\mathcal{H}$ if and only if the following hold:

(i) $\forall\, y \in \mathcal{X}$, $K(\cdot, y) \in \mathcal{H}$
(ii) $\forall\, y \in \mathcal{X}$, $\forall\, f \in \mathcal{H}$, $\langle f, K(\cdot, y)\rangle_{\mathcal{H}} = f(y)$.

$\mathcal{H}$ is called a *reproducing kernel Hilbert space*.

It can be shown that every reproducing kernel (r.k.), K is symmetric and positive definite. Conversely, Moore-Aronszajn theorem states that for every positive definite kernel, K, there exists a unique RKHS, $\mathcal{H}$ for which K is the r.k. Examples of K include $\exp(-\sigma\|x-y\|_2^2)$, $\exp(-\sigma\|x-y\|_1)$, $\langle x, y\rangle_2$, $x, y \in \mathbb{R}^d$, $\sigma > 0$, etc.

Given a measurable and bounded kernel, K, any $\mathbb{P} \in M_+^1(\mathcal{X})$ can be embedded into $\mathcal{H}$ [6], [7] as

$$\mathbb{P} \mapsto \int_{\mathcal{X}} K(\cdot, x)\, d\mathbb{P}(x). \tag{4}$$

Given the embedding in (4), we can define a pseudo-metric on $M_+^1(\mathcal{X})$ as

$$\gamma_K(\mathbb{P}, \mathbb{Q}) = \left\| \int_{\mathcal{X}} K(\cdot, x)\, d\mathbb{P}(x) - \int_{\mathcal{X}} K(\cdot, x)\, d\mathbb{Q}(x) \right\|_{\mathcal{H}}, \tag{5}$$

which is the distance between the embeddings of $\mathbb{P}$ and $\mathbb{Q}$ in $\mathcal{H}$. We refer to γ_K as the kernel distance. Note that γ_K is in general not a metric. The choice of K determines whether γ_K is a metric or not. Suppose $K(x, y) = \langle x, y\rangle_2$, $x, y \in \mathbb{R}^d$. It is easy to check that γ_K is the Euclidean distance between the means of $\mathbb{P}$ and $\mathbb{Q}$ and therefore is not a metric on $M_+^1(\mathbb{R}^d)$. Also note that choosing $k(x, y) = \exp(-\sqrt{-1}\langle x, y\rangle_2)$ in (5) yields the L_2 distance between the characteristic functions of $\mathbb{P}$ and $\mathbb{Q}$. Therefore, the embedding in (4) can be seen as a generalization of the characteristic function of $\mathbb{P}$. The question of when is γ_K a metric on $M_+^1(\mathcal{X})$ is addressed in [8]–[11]. Examples of kernels for which γ_K is a metric include $\exp(-\sigma\|x - y\|_2^2)$, $\exp(-\sigma\|x - y\|_1)$, $x, y \in \mathbb{R}^d$, $\sigma > 0$, etc. We would like to mention that γ_K is weaker than KL divergence [8], i.e., $\gamma_K(\mathbb{P}, \mathbb{Q}) \leq C\sqrt{2D(\mathbb{P}\|\mathbb{Q})}$, where $\mathbb{P}$ is absolutely continuous w.r.t. $\mathbb{Q}$ and $C := \sup\{\sqrt{K(x, x)} : x \in \mathcal{X}\}$.[1] This means there can be two distinct $\mathbb{P}$ and $\mathbb{Q}$ which need not be distinguished by γ_K while it is distinguished by KL divergence.

Suppose we are given random samples, S drawn i.i.d. from $\mathbb{P}$. Define $\mathbb{P}_n := \frac{1}{n}\sum_{j=1}^n \delta_{X_j}$, where δ_x represents the Dirac measure at $x \in \mathcal{X}$. Then, we have the following result which will be useful to prove our main result in Theorem 3.

Theorem 2: Let $\psi : \mathcal{X} \to \mathcal{H}$ be a measurable $\mathcal{H}$-valued function such that $\sup_{x \in \mathcal{X}} \|\psi(x)\|_{\mathcal{H}} \leq C < \infty$, where $\mathcal{X}$ is a measurable space and $\mathcal{H}$ is a separable Hilbert space. Then with probability at least $1 - \delta$ over the choice of $\{X_j\}_{j=1}^n$

[1] Since γ_K possibly behaves like $\sqrt{D}$, the approximation error rate of $O(1/k)$ in KL sense should probably be compared to $O(k^{-1/2})$ rate in γ_K sense. Similar is the case with the estimation error where the rate of $O_f(n^{-1/2})$ in KL sense should be compared to $O_f(n^{-1/4})$ rate in γ_K sense. From this perspective, we can say that γ_K provides a better estimation error rate than the KL while maintaining the same approximation error rate.

drawn i.i.d. from $\mathbb{P}$, the following holds:

$$\left\|\int_{\mathcal{X}} \psi(x)\, d(\mathbb{P}_n - \mathbb{P})(x)\right\|_{\mathcal{H}} \leq \frac{2C}{\sqrt{n}} + \sqrt{\frac{2C^2}{n}\log\frac{1}{\delta}}.$$

Proof: Define

$$\gamma_\psi(\mathbb{P}_n, \mathbb{P}) := \left\|\frac{1}{n}\sum_{j=1}^{n}\psi(X_j) - \int_{\mathcal{X}}\psi(x)\, d\mathbb{P}(x)\right\|_{\mathcal{H}}.$$

Suppose we replace X_j by X_j' which is also drawn from $\mathbb{P}$. We denote the associated $\gamma_\psi(\mathbb{P}_n, \mathbb{P})$ as $\gamma_\psi^{\setminus j}(\mathbb{P}_n, \mathbb{P})$. Then it is easy to check that $|\gamma_\psi(\mathbb{P}_n, \mathbb{P}) - \gamma_\psi^{\setminus j}(\mathbb{P}_n, \mathbb{P})| \leq \frac{1}{n}\left\|\psi(X_j) - \psi(X_j')\right\|_{\mathcal{H}} \leq \frac{2C}{n}$. Therefore $\gamma_\psi(\mathbb{P}_n, \mathbb{P})$ satisfies the bounded difference inequality. Hence, by invoking McDiarmid's inequality, we have that with probability at least $1-\delta$ over S,

$$\gamma_\psi(\mathbb{P}_n, \mathbb{P}) \leq \mathbb{E}\gamma_\psi(\mathbb{P}_n, \mathbb{P}) + \sqrt{\frac{2C^2}{n}\log\frac{1}{\delta}}.$$

By using the symmetrization argument [12], it can be shown that $\mathbb{E}\gamma_\psi(\mathbb{P}_n, \mathbb{P}) \leq 2\mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n}\rho_j\psi(X_j)\right\|_{\mathcal{H}} = 2\mathbb{E}\mathbb{E}_\rho\left\|\frac{1}{n}\sum_{j=1}^{n}\rho_j\psi(X_j)\right\|_{\mathcal{H}}$, where $(\rho)_k$ are the i.i.d. Rademacher random variables, i.e., $\Pr(\rho_j = 1) = \Pr(\rho_j = -1) = \frac{1}{2}, \forall j$, and $\mathbb{E}_\rho\left\|\frac{1}{n}\sum_{j=1}^{n}\rho_j\psi(X_j)\right\|_{\mathcal{H}} := \mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\rho_j\psi(X_j)\right\|_{\mathcal{H}}|\{X_j\}_{j=1}^n\right]$. Now consider

$$\begin{aligned}\mathbb{E}_\rho\left\|\frac{1}{n}\sum_{j=1}^{n}\rho_j\psi(X_j)\right\|_{\mathcal{H}} &= \frac{1}{n}\mathbb{E}_\rho\sqrt{\sum_{i,j=1}^{n}\rho_i\rho_j\langle\psi(X_i),\psi(X_j)\rangle_{\mathcal{H}}} \\ &\leq \frac{1}{n}\mathbb{E}_\rho\sqrt{\sum_{i=1}^{n}\rho_i^2\langle\psi(X_i),\psi(X_i)\rangle_{\mathcal{H}}} \\ &\quad + \frac{1}{n}\mathbb{E}_\rho\sqrt{\sum_{i\neq j}\rho_i\rho_j\langle\psi(X_i),\psi(X_j)\rangle_{\mathcal{H}}}.\end{aligned}$$

Clearly the first summand is bounded above by $C/\sqrt{n}$. Note that by invoking Jensen's inequality in the second summand, we obtain $\mathbb{E}_\rho\sqrt{\sum_{i\neq j}\rho_i\rho_j\langle\psi(X_i),\psi(X_j)\rangle_{\mathcal{H}}} \leq \sqrt{\mathbb{E}_\rho\sum_{i\neq j}\rho_i\rho_j\langle\psi(X_i),\psi(X_j)\rangle_{\mathcal{H}}} = 0$, thereby proving the result. ∎

III. Bounds for Mixture Density Estimation

In this section, we present our main result of establishing the estimation and approximation errors for mixture density estimation using γ_K as the measure of goodness of fit. We show in Theorem 3 that an estimation error rate of $O_f(n^{-1/2})$—which is a *faster* rate than that obtained with KL divergence—can be achieved with γ_K while the *same* rate of $O(k^{-1/2})$ is obtained for the approximation error (see footnote 1). However, we would like to mention that unlike the results in [1]–[3], we do not make any assumptions on f and $\mathcal{C}$.

Define

$$\gamma_K(f, g) := \left\|\int_{\mathcal{X}} K(\cdot, x)(f(x) - g(x))\, d\mu(x)\right\|_{\mathcal{H}},$$

$$\gamma_K(S, g) := \left\|\frac{1}{n}\sum_{i=1}^{n}K(\cdot, X_i) - \int K(\cdot, x)g(x)\, d\mu(x)\right\|_{\mathcal{H}},$$

and

$$g_{\text{emp}} := \arg\min_{g\in\mathcal{G}_k}\gamma_K(S, g).$$

Note that g_{emp} is an M-estimator.

Theorem 3: Let $C := \sup_{x\in\mathcal{X}}\sqrt{K(x,x)}$, where K is a continuous kernel on a separable topological space, $\mathcal{X}$. Then with probability at least $1-\delta$ over the choice of samples $\{X_j\}_{j=1}^n$ drawn i.i.d. from f, the following holds:

$$\gamma_K(f, g_{\text{emp}}) - \inf_{g\in\mathcal{G}}\gamma_K(f, g) \leq \frac{4C}{\sqrt{n}} + \sqrt{\frac{8C^2}{n}\log\frac{2}{\delta}} + \frac{2C}{\sqrt{k}}. \quad (6)$$

In addition,

$$\begin{aligned}-\frac{2C}{\sqrt{n}} - \sqrt{\frac{2C^2}{n}\log\frac{1}{\delta}} &\leq \gamma_K(S, g_{\text{emp}}) - \inf_{g\in\mathcal{G}}\gamma_K(f, g) \\ &\leq \frac{2C}{\sqrt{n}} + \sqrt{\frac{2C^2}{n}\log\frac{1}{\delta}} + \frac{2C}{\sqrt{k}}. \quad (7)\end{aligned}$$

Proof: First we show that for any $g \in \mathcal{G}$, there exists $\tilde{g}_k \in \mathcal{G}_k$ such that $\gamma_K(\tilde{g}_k, g) = O(k^{-1/2})$. Consider

$$\begin{aligned}\int_{\mathcal{X}} K(\cdot, x)g(x)\, d\mu(x) &= \int_{\mathcal{X}} K(\cdot, x)\int_\Theta \phi_\theta(x)\, d\mathbb{P}(\theta)\, d\mu(x) \\ &\overset{(\star)}{=} \int_\Theta \overbrace{\int_{\mathcal{X}} K(\cdot, x)\phi_\theta(x)\, d\mu(x)}^{\tilde{K}(\cdot,\theta)}\, d\mathbb{P}(\theta),\end{aligned}$$

where we have invoked Fubini's theorem in $(\star)$. Let us draw i.i.d. samples, $\{\tilde{\theta}_j\}_{j=1}^k$ from $\mathbb{P}$. Define

$$\tilde{g}_k(x) := \frac{1}{k}\sum_{j=1}^{k}\phi_{\tilde{\theta}_j}(x) \in \mathcal{G}_k.$$

Then

$$\int_{\mathcal{X}} K(\cdot, x)\tilde{g}_k(x)\, d\mu(x) = \frac{1}{k}\sum_{j=1}^{k}\tilde{K}(\cdot, \tilde{\theta}_j).$$

Therefore,

$$\gamma_K(\tilde{g}_k, g) = \gamma_{\tilde{K}}(\mathbb{P}_k, \mathbb{P}),$$

which from Theorem 2 (since k is a continuous kernel defined on a separable topological space, $\mathcal{X}$, by Lemma 4.33 of [13], $\mathcal{H}$ is separable and therefore Theorem 2 can be invoked) means that with probability at least $1-\delta$ over $\{\tilde{\theta}_j\}_{j=1}^k$, we have

$$\gamma_K(\tilde{g}_k, g) \leq \frac{2C}{\sqrt{k}} + \sqrt{\frac{2C^2}{k}\log\frac{1}{\delta}}.$$

By letting $\delta \to 1$, we conclude that for any $g \in \mathcal{G}$, there exists $g_k \in \mathcal{G}_k$ such that

$$\gamma_K(\tilde{g}_k, g) \leq \frac{2C}{\sqrt{k}}. \quad (8)$$

Let us fix an $\varepsilon > 0$ and a function $g_\varepsilon \in \mathcal{G}$ such that

$$\gamma_K(f, g_\varepsilon) \leq \inf_{g \in \mathcal{G}} \gamma_K(f, g) + \varepsilon.$$

Consider

$$\begin{aligned}
\gamma_K(f, g_{\text{emp}}) - \inf_{g \in \mathcal{G}} \gamma_K(f, g) &= \overbrace{\gamma_K(f, g_{\text{emp}}) - \gamma_K(S, g_{\text{emp}})}^{A_1} \\
&\quad + \overbrace{\gamma_K(S, g_{\text{emp}}) - \gamma_K(S, \tilde{g}_k)}^{A_2} \\
&\quad + \overbrace{\gamma_K(S, \tilde{g}_k) - \gamma_K(f, \tilde{g}_k)}^{A_3} \\
&\quad + \gamma_K(f, \tilde{g}_k) - \inf_{g \in \mathcal{G}} \gamma_K(f, g) \\
&\leq A_1 + A_2 + A_3 \\
&\quad + \overbrace{\gamma_K(f, \tilde{g}_k) - \gamma_K(f, g_\varepsilon)}^{A_4} + \varepsilon.
\end{aligned}$$

Note that

$$\begin{aligned}
A_1 &\overset{(*)}{\leq} \gamma_K(S, f), \\
A_2 &\overset{(\star\star)}{\leq} 0, \\
A_3 &\overset{(*)}{\leq} \gamma_K(S, f), \\
A_4 &\overset{(*)}{\leq} \gamma_K(\tilde{g}_k, g_\varepsilon) \overset{(8)}{\leq} \frac{2C}{\sqrt{k}},
\end{aligned}$$

where $(*)$ follows from the reverse triangle inequality and $(\star\star)$ follows from the fact that g_{emp} is the minimizer of $\gamma_K(S, g)$ over $g \in \mathcal{G}_k$. It follows from Theorem 2 that with probability at least $1 - \frac{\delta}{2}$ over the choice of $\{X_j\}_{j=1}^n$, we have

$$\gamma_K(S, f) \leq \frac{2C}{\sqrt{n}} + \sqrt{\frac{2C^2}{n} \log \frac{2}{\delta}}.$$

Combining all the above results, we have that with probability at least $1-\delta$ over the choice of $\{X_j\}_{j=1}^n$, the following holds:

$$\begin{aligned}
\gamma_K(f, g_{\text{emp}}) - \inf_{g \in \mathcal{G}} \gamma_K(f, g) \leq \frac{4C}{\sqrt{n}} + \sqrt{\frac{8C^2}{n} \log \frac{2}{\delta}} \\
+ \frac{2C}{\sqrt{k}} + \varepsilon. \qquad (9)
\end{aligned}$$

Letting $\varepsilon \to 0$ in (9) yields the result in (6).

Since

$$\gamma_K(S, g_{\text{emp}}) - \inf_{g \in \mathcal{G}} \gamma_K(f, g) \leq A_2 + A_3 + A_4 + \varepsilon,$$

the upper bound in (7) follows by taking $\varepsilon \to 0$. On the other hand, since $\gamma_K(f, g_{\text{emp}}) - \inf_{g \in \mathcal{G}} \gamma_K(f, g) \geq 0$, we have

$$\begin{aligned}
\gamma_K(S, g_{\text{emp}}) - \inf_{g \in \mathcal{G}} \gamma_K(f, g) &\geq \gamma_K(S, g_{\text{emp}}) - \gamma_K(f, g_{\text{emp}}) \\
&\geq -\gamma_K(S, f)
\end{aligned}$$

and the lower bound in (7) follows from Theorem 2. ∎

Although γ_K provides a nice result that does not make any assumptions on f and $\mathcal{C}$, one would be intrigued about what makes γ_K special. In fact, can't we use any other distance metric on $M_+^1(\mathcal{X})$ to obtain a similar result? This can be understood by carefully noticing the proof of Theorem 3 wherein it should be clear that the terms, A_1, A_3 and A_4 are bounded by $\gamma_K(S, f)$—note that $\gamma_K(\tilde{g}_k, g_\varepsilon)$, which is a bound on A_4 is also of the form $\gamma_K(S, f)$ where S is drawn from $f = g_\varepsilon$. This means, Theorem 3 hinges completely on Theorem 2. Therefore, if we use a distance measure that behaves similar to γ_K in the sense shown in Theorem 2, we obtain similar results as obtained with γ_K. However, it is not clear what distance measures other than γ_K exhibit this behavior.

IV. Conclusion & Discussion

In this work, we have studied the problem of mixture density estimation using the notion of embedding probability measures into a reproducing kernel Hilbert space. We showed that this approach does not make any assumptions on the unknown density or the base class of densities while achieving rates that are better to those obtained using KL divergence. While we proved results for the M-estimator, in future, we would like to investigate the bounds for a greedy procedure similar to the one considered in [1] and [2].

References

[1] J. Li and A. Barron, "Mixture density estimation," in *Advances in Neural Information Processing Systems 12*, S. A. Solla, T. K. Leon, and K.-R. Muller, Eds. San Mateo, CA: Morgan Kaufmann Publishers, 1999, pp. 279–285.

[2] J. Li, "Estimation of mixture models," Ph.D. dissertation, Department of Statistics, Yale University, 1999.

[3] A. Rakhlin, D. Panchenko, and S. Mukherjee, "Risk bounds for mixture density estimation," *ESAIM: Probability and Statistics*, vol. 9, pp. 220–229, 2005.

[4] S. van de Geer, "Rates of convergence for the maximum likelihood estimator in mixture models," *Nonparametric Statistics*, vol. 6, pp. 293–310, 1996.

[5] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, pp. 337–404, 1950.

[6] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. London, UK: Kluwer Academic Publishers, 2004.

[7] A. J. Smola, A. Gretton, L. Song, and B. Schölkopf, "A Hilbert space embedding for distributions," in *Proc. 18th International Conference on Algorithmic Learning Theory*. Springer-Verlag, Berlin, Germany, 2007, pp. 13–31.

[8] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet, "Hilbert space embeddings and metrics on probability measures," *Journal of Machine Learning Research*, vol. 11, pp. 1517–1561, 2010.

[9] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, "Kernel measures of conditional dependence," in *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. Cambridge, MA: MIT Press, 2008, pp. 489–496.

[10] K. Fukumizu, B. K. Sriperumbudur, A. Gretton, and B. Schölkopf, "Characteristic kernels on groups and semigroups," in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., 2009, pp. 473–480.

[11] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel method for the two sample problem," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. MIT Press, 2007, pp. 513–520.

[12] A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*. New York: Springer-Verlag, 1996.

[13] I. Steinwart and A. Christmann, *Support Vector Machines*. Springer, 2008.