In this supplementary material, we provide the proofs of all theorems and propositions in the paper.

**Definition 1 (Hilbert Sinkhorn divergence, HSD).** Given measures \( \mu, \nu \in \mathbb{P}(X) \) and elements \( u, v \in \mathcal{H} \), the Hilbert Sinkhorn divergence between embedding \( \phi_*\mu \) and \( \phi_*\nu \) is written as

\[
S_\epsilon (\phi_*\mu, \phi_*\nu) = \inf_{\pi_{\phi}} \int_{\mathcal{H} \times \mathcal{H}} c_{\phi}(u, v) d\pi_{\phi}(u, v) + \epsilon \Phi(\pi_{\phi}) \tag{1}
\]

where \( \pi_{\phi} \in \Pi (\phi_*\mu, \phi_*\nu) \) is a joint probability measure with two marginals \( \phi_*\mu \) and \( \phi_*\nu \), and

\[
c_{\phi}(u, v) = \|u - v\|_\mathcal{H}^2 \\
\Phi(\pi_{\phi}) = \log \left( \frac{d\pi_{\phi}}{d(\phi_*\mu) d(\phi_*\nu)}(u, v) \right)
\]

**Definition 2 (Hilbert embedding).** Let \( \mathbb{P}(X) \) be the set of probability measures on sample set \( X \) and \( \mathbb{P}(\mathcal{H}) \) be the set of probability measures on reproducing kernel Hilbert space \( \mathcal{H} \). Given a probability measure \( \mu \in \mathbb{P}(X) \), the implicit feature map \( \phi : X \to \mathcal{H} \) will induce the Hilbert embedding of \( \mu \):

\[
\phi_* : \mathbb{P}(X) \to \mathbb{P}(\mathcal{H}), \mu \mapsto \phi_*\mu = \int_X \phi(x) d\mu(x) \tag{2}
\]

For the map \( (\phi, \phi) : X \times X \to \mathcal{H} \times \mathcal{H} \), we similarly have

\[
(\phi, \phi)_* : (\mu, \nu) \mapsto (\phi_*\mu, \phi_*\nu) \tag{3}
\]

**Definition 3 (Hilbert Sinkhorn divergence).** Given measures \( \mu, \nu \in \mathbb{P}(X) \) and elements \( u, v \in \mathcal{H} \), the Hilbert Sinkhorn divergence between embedding \( \phi_*\mu \) and \( \phi_*\nu \) is written as

\[
S_\epsilon (\phi_*\mu, \phi_*\nu) = \inf_{\pi_{\phi}} \int_{\mathcal{H} \times \mathcal{H}} c_{\phi}(u, v) d\pi_{\phi}(u, v) + \epsilon \Phi(\pi_{\phi}) \tag{4}
\]

where \( \pi_{\phi} \in \Pi (\phi_*\mu, \phi_*\nu) \) is a joint probability measure with two marginals \( \phi_*\mu \) and \( \phi_*\nu \), and

\[
c_{\phi}(u, v) = \|u - v\|_\mathcal{H}^2 \\
\Phi(\pi_{\phi}) = \log \left( \frac{d\pi_{\phi}}{d(\phi_*\mu) d(\phi_*\nu)}(u, v) \right)
\]

**Definition 4.** Given measurable spaces \( (X_1, \Sigma_1) \) and \( (X_2, \Sigma_2) \), a measurable mapping \( f : X_1 \to X_2 \) and a measure \( \mu : \Sigma_1 \to [0, +\infty] \), the pushforward of \( \mu \) is defined to be the measure \( f_*(\mu) : \Sigma_2 \to [0, +\infty] \) given by

\[
(f_*(\mu))(B) = \mu \left( f^{-1}(B) \right) \text{ for } B \in \Sigma_2 \tag{5}
\]

*Equal contribution
†Corresponding author
1. Proving Theorem 1

Theorem 1. Given two measures \( \mu, \nu \in \mathbb{P}(X) \), we write

\[
S_{H,\epsilon}(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{X \times X} c_H(x, y) d\pi(x, y) + \epsilon H(\pi)
\]

(6)

where \( \pi \in \Pi(\mu, \nu) \) is the joint probability measure on \( X \times X \) with marginals \( \mu \) and \( \nu \), and

\[
c_H(x, y) = \| \phi(x) - \phi(y) \|_H^2 = k(x, x) + k(y, y) - 2k(x, y)
\]

Then we have the following conclusions:

- \( S_{H,\epsilon}(\mu, \nu) = S_{\epsilon}(\phi_*\mu, \phi_*\nu) \)

- If \( \pi^* \) is a minimizer of (6), its Hilbert embedding \( (\phi, \phi)_*,\pi^* \) is a minimizer of (4).

Proof. Applying the pushforward map in (5), we have

\[
\begin{aligned}
\left( \phi, \phi \right)_* \pi(u, v) &= \pi \left( \phi^{-1}(u), \phi^{-1}(v) \right) = \pi(x, y).
\end{aligned}
\]

Thus, the HSD is reformulated as

\[
S_{H,\epsilon}(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{X \times X} \| \phi(x) - \phi(y) \|_H^2 d\pi(x, y) + \epsilon \log \left( \frac{d\pi}{d\mu d\nu} (x, y) \right)
\]

(7)

on the other hand, for all \( \pi \in \Pi(\mu, \nu) \),

\[
\begin{aligned}
\int_{H \times H} (\| u - v \|_H^2) d\pi(u, v) + \epsilon \log \left( \frac{d\pi}{d\phi_*\mu} (u, v) \right) = & \int_{H \times H} (\| u - v \|_H^2 + \epsilon \log \left( \frac{d\pi}{d\phi_*\mu} (u, v) \right) ) d\phi_*\pi(u, v) \\
= & \int_{X \times X} (\| \phi(x) - \phi(y) \|_H^2) d\pi(x, y) + \epsilon \log \left( \frac{d\pi}{d\mu d\nu} (x, y) \right) \\
\geq & \inf_{\pi \in \Pi(\phi_*\mu, \phi_*\nu)} \int_{X \times X} (\| \phi(x) - \phi(y) \|_H^2) d\pi(x, y) + \epsilon \log \left( \frac{d\pi}{d\mu d\nu} (x, y) \right)
\end{aligned}
\]

(8)

Take the infimum on \( \pi_\phi \in \Pi(\phi_*\mu, \phi_*\nu) \) over domain \( H \times H \), the inequality (8) remains hold. That is \( S_{\epsilon}(\phi_*\mu, \phi_*\nu) \geq S_{H,\epsilon}(\mu, \nu) \). Therefore, combining (7) and (8) achieves

\[
S_{\epsilon}(\phi_*\mu, \phi_*\nu) = S_{H,\epsilon}(\mu, \nu)
\]

(9)

If \( \pi^* \) is a minimizer of (6), then

\[
S_{H,\epsilon}(\mu, \nu) = \int_{X \times X} \| \phi(x) - \phi(y) \|_H^2 d\pi^*(x, y) + \epsilon \log \left( \frac{d\pi^*}{d\mu d\nu} (x, y) \right)
\]

(10)
which implies that $(\phi, \phi)^* \pi^*$ is a minimizer of (4).

2. Proving Proposition 1

Proposition 1 (Variational representation). The KL divergence admits the following variational representation in the reproducing kernel Hilbert space:

$$S_\epsilon (\phi_* \mu, \phi_* \nu) = \epsilon \left( 1 + \min_{\pi_{\phi}} [T] - \log (E_{\xi_{\phi}} [e^T]) \right)$$

(11)

where the infimum is taken over $\pi_{\phi} \in \Pi (\phi_* \mu, \phi_* \nu)$, $\xi_{\phi} (x, y) = e^{-d(x, y)/\epsilon}$ and function $T = \log \frac{d\pi_{\phi}}{d\xi_{\phi}} + C$ for some constant $C \in \mathbb{R}$.

Proof. Step 1: Given an absolutely continuous measure $\pi_{\phi} \in \mathbb{P}(\mathcal{H} \times \mathcal{H})$ and a positive function $\xi_{\phi}$ on $\mathcal{H} \times \mathcal{H}$, we define the Kullback-Leibler (KL) divergence

$$\text{KL}(\pi_{\phi} \mid \xi_{\phi}) = \int_{\mathcal{H} \times \mathcal{H}} \pi(u, v) \left[ \ln \frac{\pi(u, v)}{\xi(u, v)} - 1 \right] \text{d}u \text{d}v$$

(12)

We can associate $\|u - v\|_{\mathcal{H}}^2$ to a Gibbs distribution $\xi_{\phi}$ defined by $dG = \frac{1}{Z} e^T d\xi_{\phi}$, where $Z = E_{\xi_{\phi}} [e^T]$.

By combining KL divergence (12) and Gibbs distribution (13) algebraically, Hilbert Sinkhorn divergence (4) can be computed as the smallest KL divergence between coupling $\pi_{\phi}$ and Gibbs distribution $\xi_{\phi}$ in the reproducing kernel Hilbert space:

$$S_\epsilon (\phi_* \mu, \phi_* \nu) = \epsilon \left( 1 + \min_{\pi_{\phi} \in \Pi(\phi_* \mu, \phi_* \nu)} \text{KL} (\pi_{\phi} \mid \xi_{\phi}) \right)$$

(14)

Step 2. We use Donsker-Varahan representation for KL divergence

$$\text{KL} (\pi_{\phi} \mid \xi_{\phi}) = \sup_{T : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}} E_{\pi_{\phi}} [T] - \log (E_{\xi_{\phi}} [e^T])$$

(15)

A simple proof of (15) is as follows: for a given function $T$, let us consider the Gibbs distribution $G$ defined by $dG = \frac{1}{Z} e^T d\xi_{\phi}$, where $Z = E_{\xi_{\phi}} [e^T]$. By construction

$$E_{\pi_{\phi}} [T] - \log Z = E_{\pi_{\phi}} \left[ \log \frac{dG}{d\xi_{\phi}} \right]$$

(16)

Let $\Delta$ be the gap,

$$\Delta := \text{KL} (\pi_{\phi} \mid \xi_{\phi}) - (E_{\pi_{\phi}} [T] - \log (E_{\xi_{\phi}} [e^T]))$$

(17)

Using Eq. (16), we can write $\Delta$ as the KL-divergence:

$$\Delta = E_{\pi_{\phi}} \left[ \log \frac{d\pi_{\phi}}{d\xi_{\phi}} - \log \frac{dG}{d\xi_{\phi}} \right] = E_{\pi_{\phi}} \log \frac{d\pi_{\phi}}{dG} = \text{KL}(\pi_{\phi} \parallel G)$$

(18)

For KL-divergence, we have $\Delta \geq 0$ in (17). Thus, it can be shown that for any $T$,

$$\text{KL}(\pi_{\phi} \parallel \xi_{\phi}) \geq E_{\pi_{\phi}} [T] - \log (E_{\xi_{\phi}} [e^T])$$

(19)

and the inequality also holds for taking the supremum on the right side. Finally, the identity (18) also shows that $\Delta = 0$ whenever $G = \pi_{\phi}$, i.e., optimal functions $T$ has the form

$$T = \log \frac{d\pi_{\phi}}{d\xi_{\phi}} + C$$

(20)

for some constant $C \in \mathbb{R}$. Combining (14), (15) and (20), we achieve the conclusion.
3. Proving Proposition 2

**Proposition 2** (Lower bound). The Hilbert Sinkhorn distance has the following lower bound:

\[ S_\epsilon (\varphi_* \mu, \varphi_* \nu) \geq \epsilon \left( 1 + \min_{\pi \in \Pi(\mu, \nu)} \mathbb{E}_\pi [k] - \log \left( \mathbb{E}_{\xi} \left[ e^k \right] \right) \right) \]

where \( \epsilon > 0 \), \( \varphi_* \mu \) and \( \varphi_* \nu \) are Hilbert embedding in Eq. (2) and \( k \) is a kernel function.

**Proof.** KL divergence \( (15) \) in product space \( H \times H \) satisfies

\[
\text{KL} (\pi_\phi \mid \xi_\phi) = \sup_{T: H \times H \to \mathbb{R}} \mathbb{E}_{\pi_\phi} [T] - \log \left( \mathbb{E}_{\xi_\phi} \left[ e^T \right] \right)
\]

\[
= \sup_{T: H \times H \to \mathbb{R}} \int_{H \times H} T(u, v) d\pi_\phi (u, v) - \log \int_{H \times H} e^T d\xi_\phi (u, v)
\]

\[
= \sup_{T: H \times H \to \mathbb{R}} \int_{H \times H} T(u, v) d((\phi, \phi)_x) (u, v) - \log \int_{H \times H} e^{T(u,v)} d((\phi, \phi)_x) (u, v)
\]

\[
\geq \sup_{T \in \mathcal{M}} \int_{\Omega \times \Omega} T(\phi(x), \phi(y)) d\pi(x, y) - \log \int_{\Omega \times \Omega} e^{T(\phi(x), \phi(y))} d\xi(x, y)
\]

By (22) and (14), given a kernel function \( k \) we have the lower bound

\[
S_\epsilon (\varphi_* \mu, \varphi_* \nu) = \epsilon \left( 1 + \min_{\pi_\phi \in \pi_\phi} \text{KL} (\pi_\phi \mid \xi_\phi) \right)
\]

\[
\geq \epsilon \left( 1 + \min_{\pi \in \Pi} \sup_{k: X \times X \to \mathbb{R}} \mathbb{E}_\pi [k] - \log \left( \mathbb{E}_{\xi} \left[ e^k \right] \right) \right)
\]

\[
\geq \epsilon \left( 1 + \min_{\pi \in \Pi} \mathbb{E}_\pi [k] - \log \left( \mathbb{E}_{\xi} \left[ e^k \right] \right) \right)
\]

Since \( S_{H, \epsilon} (\mu, \nu) = S_\epsilon (\varphi_* \mu, \varphi_* \nu) \) as provided in Theorem \( \text{[1]} \) As a consequence, we directly have the following result by using Proposition \( \text{[1]} \) and \( \text{[2]} \).

**Corollary 1.** The reformulation \( (6) \) admits the following variational representation and lower bound:

\[
S_{H, \epsilon} (\mu, \nu) = \epsilon \left( 1 + \min_{\pi_\phi \in \pi_\phi} \mathbb{E}_{\pi_\phi} [T] - \log \left( \mathbb{E}_{\xi_\phi} \left[ e^T \right] \right) \right)
\]

\[
S_{H, \epsilon} (\mu, \nu) \geq \epsilon \left( 1 + \min_{\pi \in \Pi(\mu, \nu)} \mathbb{E}_\pi [k] - \log \left( \mathbb{E}_{\xi} \left[ e^k \right] \right) \right)
\]

The related notations are defined in Prop \( \text{[7]} \) and \( \text{[2]} \)
4. Proving Theorem 2

**Theorem 2** (Strong consistency). Given empirical measures \( \mu_n, \nu_n \) and \( \epsilon, \eta > 0 \), there exists \( N > 0 \) such that

\[
\forall n \geq N, \ P \left( |S_{H,\epsilon}(\mu_n, \nu_n) - S_{H,\epsilon}(\mu, \nu)| \leq \epsilon \eta \right) = 1
\]  

(24)

**Proof.** We assume that \( \pi_\phi \) is the optimal of problem [14], and \( \pi_{\phi, n} \) is the optimal of problem

\[
S_{H,\epsilon}(\mu_n, \nu_n) = \epsilon \left( 1 + \min_{\pi_{\phi, n} \in \Pi_{\phi, n}} : \text{KL}(\pi_{\phi, n} \mid \xi_{\phi, n}) \right)
\]  

(25)

Then we start by using (15) and the triangular inequality to write,

\[
|S_{H,\epsilon}(\mu_n, \nu_n) - S_{H,\epsilon}(\mu, \nu)| \leq \epsilon \left( \sup_{T : H \times H \to \mathbb{R}} |E_{\pi_{\phi, n}}[T] - E_{\pi_{\phi}}[T]| + \sup_{T : H \times H \to \mathbb{R}} |\log(E_{\xi_{\phi}}[e^T]) - \log(E_{\xi_{\phi, n}}[e^T])| \right)
\]  

(26)

It is reasonable to assume that functions \( T \) are uniformly bounded by a constant \( M \), i.e \( \|T\|_H \leq M \) in reproducing kernel Hilbert space. Since \( \log \) is Lipschitz continuous with constant \( e^M \) in the interval \( [e^{-M}, e^M] \), we have

\[
|\log(E_{\xi_{\phi}}[e^T]) - \log(E_{\xi_{\phi, n}}[e^T])| \leq e^M |E_{\xi_{\phi}}[e^T] - E_{\xi_{\phi, n}}[e^T]|
\]  

(27)

The families of functions \( T \) and \( e^T \) satisfy the uniform law of large numbers [3][4][5]. Given \( \eta > 0 \), we can thus choose \( N \in \mathbb{N} \) such that \( \forall n \geq N \) and with probability one,

\[
\sup_{T : H \times H \to \mathbb{R}} |E_{\pi_{\phi, n}}[T] - E_{\pi_{\phi}}[T]| \leq \frac{\eta}{2} \quad \text{and} \quad \sup_{T : H \times H \to \mathbb{R}} |\log(E_{\xi_{\phi}}[e^T]) - \log(E_{\xi_{\phi, n}}[e^T])| \leq \frac{\eta}{2} e^{-M}
\]  

(28)

Substituting Eqs. (28) and (28) into (26) leads to

\[
\forall n \geq N, \quad |S_{H,\epsilon}(\mu_n, \nu_n) - S_{H,\epsilon}(\mu, \nu)| \leq \frac{\epsilon \eta}{2} + \frac{\epsilon \eta}{2} + \epsilon \eta
\]  

(29)

with probability one. \( \square \)

5. Proving Proposition 3

**Proposition 3** (approximation error). Let sample space \( X \) be a subset of \( \mathbb{R}^d \) and diameter be \( |X| = \sup \{ \|x - y\| \mid x, y \in X \} \), we have

\[
|S_{H,\epsilon}(\mu, \nu) - W_{\epsilon}(\mu, \nu)| \leq \epsilon \eta
\]

\[
|S_{H,\epsilon}(\mu, \nu) - W(\mu, \nu)| \leq \epsilon \left( \eta + 2d \log(e^{2LD} / \sqrt{\epsilon}) \right)
\]

(30)

where \( \epsilon > 0 \), \( D \geq |X| \) and \( L \) is a Lipschitz constant.

**Proof.** \textbf{Step 1.} Notice that we can follow the idea of Proposition [1] to construct the following representation in Euclidean space

\[
W_{\epsilon}(\mu, \nu) = \epsilon \left( 1 + \min_{\pi \in \mathcal{H}(\mu, \nu)} E_{\pi}[T] - \log(E_{\xi}[e^T]) \right)
\]

(31)

where \( \xi(x, y) = e^{-d(x,y)/\epsilon} \) and function \( T = \log \frac{d\pi}{d\pi'} \). By construction, \( T \) satisfies \( E_{\xi}[e^T] = \int d\pi = 1 \).

Without loss of generality, we assume that \( \pi \) makes the minimum of \( E_{\pi}[K] - \log(E_{\xi}[e^K]) \) appeared in (??). Then

\[
W_{\epsilon}(\mu, \nu) - S_{H,\epsilon}(\mu, \nu) \leq \epsilon \left( E_{\pi}[T] - \log(E_{\xi}[e^T]) \right) - \epsilon \left( E_{\pi}[K] - \log(E_{\xi}[e^K]) \right) \text{ by (31) and (??)}
\]

\[
= \epsilon \left( (E_{\pi}[T] - \log 1) - E_{\pi}[K] + \log(E_{\xi}[e^K]) \right)
\]

\[
= \epsilon \left( (E_{\pi}(T - K) + \log(E_{\xi}[e^K]) \right)
\]

\[
\leq \epsilon \left( (E_{\pi}(T) - E_{\pi}(K) + (E_{\xi}[e^K] - 1)) \right)
\]

(32)
where we used the inequality $\log x \leq x - 1$.

**Step 2.** Fix $\eta > 0$. We first consider the case where $|T| \leq M$ is bounded. By the universal approximation theorem, we can choose a kernel function $K \leq M$ such that

$$\mathbb{E}_x [T - K] \leq \frac{\eta}{2} \quad \text{and} \quad \mathbb{E}_x |T - K| \leq \frac{\eta}{2} e^{-M} \tag{33}$$

Since $\exp$ is Lipschitz continuous with constant $e^{M}$ on $(-\infty, M]$, we have

$$\mathbb{E}_x [e^T - e^K] \leq e^{M} \mathbb{E}_x |T - K| \leq \frac{\eta}{2} \tag{34}$$

From (33)-(34) and the triangular inequality, we then obtain

$$|W_\epsilon(\mu, \nu) - S_{H, \epsilon}(\mu, \nu)| \leq \epsilon \left(\mathbb{E}_x |T - K| + \mathbb{E}_x [e^T - e^K]\right) \leq \epsilon \eta \tag{35}$$

which proves (30).

**Step 3.** In this step, we are interested in bounding the error made when approximating $W(\mu, \nu)$ with $S_{H, \epsilon}(\mu, \nu)$. Assume $\mathcal{X}$ is the subsets of $\mathbb{R}^d$, the diameter $|\mathcal{X}| = \sup \{\|x - x'\| \mid x, x' \in \mathcal{X}\} \leq D$ and the cost function is $L$-Lipschitz. Then it holds

$$W_{\epsilon}(\mu, \nu) - W(\mu, \nu) \leq 2\epsilon d \log \frac{e^{2LD}}{\sqrt{d\epsilon}} \tag{36}$$

From (35), (36) and the triangular inequality, we then obtain

$$|S_{H, \epsilon}(\mu, \nu) - W(\mu, \nu)| \leq \epsilon \left(\eta + 2d \log \frac{e^{2LD}}{\sqrt{d\epsilon}}\right) \tag{37}$$

\[ \Box \]

6. Proving Theorem 3

**Theorem 3** (asymptotic bound). The Hilbert Sinkhorn estimator $S_{H, \epsilon}(\mu_n, \nu_n)$ approximates the Wasserstein distance $W(\mu, \nu)$ with the following bound,

$$\forall n \geq N, \mathbb{P}(|S_{H, \epsilon}(\mu_n, \nu_n) - W(\mu, \nu)| \leq \zeta) = 1 \tag{38}$$

where $\zeta = 2e \left(\eta + d \log \frac{e^{2LD}}{\sqrt{d\epsilon}}\right)$.

**Proof.** Let $\eta > 0$. We find a kernel function and $N > 0$ such that (24) and (30) hold. By the triangular inequality, for all $n \geq N$ and with probability one, we have:

$$|S_{H, \epsilon}(\mu_n, \nu_n) - W(\mu, \nu)| \leq |S_{H, \epsilon}(\mu_n, \nu_n) - S_{H, \epsilon}(\mu, \nu)| + |S_{H, \epsilon}(\mu, \nu) - W(\mu, \nu)| \leq \zeta$$

where $\zeta = 2e \left(\eta + d \log \frac{e^{2LD}}{\sqrt{d\epsilon}}\right)$

\[ \Box \]

7. Proving Theorem 4

**Lemma 1.** [10] We assume that arbitrary function $f \in \mathcal{H}$ is bounded (i.e., $\|f\|_{\mathcal{H}} \leq M$). Given the covering disk $B_\eta = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq \eta\}$, the covering number of $\mathcal{H}$ is

$$N(\mathcal{H}, \eta) \leq \left(\frac{3M}{\eta}\right)^m \tag{39}$$

where $m$ is the number of basis that span the function $f$.

**Theorem 4.** Given the desired accuracy parameters $\eta, \epsilon > 0$ and the confidence parameter $\eta$, we have,

$$\mathbb{P}(|S_{H, \epsilon}(\mu, \nu) - S_{H, \epsilon}(\mu_n, \nu_n)| \leq \epsilon \eta) \geq 1 - \delta, \tag{40}$$

whenever the number $n$ of samples satisfies

$$n \geq \frac{2M^2(\log(2/\delta) + m \log(24M/\eta))}{\eta^2} \tag{41}$$

where $m$ and $M$ are given in Lem. [7]
We deduce from (45) and L-Lipschtiz in reproducing kernel Hilbert space. By Hoeffding inequality, for all function $|f| \leq M$

$$\Pr \left( |E_{\mu}f - E_{\mu,n}f| > \frac{\eta}{4} \right) \leq 2 \exp \left( -\frac{\eta^2 n}{2M^2} \right)$$

(42)

To extend this inequality to a uniform inequality over all functions $T$ and $e^T$, the standard technique is to choose a minimal cover of Hilbert space by a finite set of small balls with radius $\eta$. We need to choose a minimal cover number of the domain $B_R = \{ T \in \mathcal{H} : \|T\|_H \leq R \}$ by a finite set of small balls with radius $\eta$ such that $B_R \subset \bigcup_j B_\eta \{ T_j \}$. As given in Lemma [1] the minimal cardinality of such covering is bounded by the covering number such that

$$\mathcal{N}(\mathcal{H}, \eta) \leq \left( \frac{3M}{\eta} \right)^m$$

(43)

Successively applying the union bound in (42) with the set of functions $\{ T_j \}$ to get

$$\Pr \left( \sup_j |E_{\pi_o} \{ T_j \} - E_{\pi_o,n} \{ T_j \}| \geq \frac{\eta}{4} \right) \leq 2\mathcal{N}(\mathcal{H}, \eta) \exp \left( -\frac{\eta^2 n}{2M^2} \right) < \delta$$

(44)

which gives

$$\Pr \left( \sup_j |E_{\pi_o} \{ T_j \} - E_{\pi_o,n} \{ T_j \}| \leq \frac{\eta}{4} \right) > 1 - \delta$$

(45)

We now choose small ball radius $\lambda = \eta/8L$. Then solving $2\mathcal{N}(\mathcal{H}, \eta) \exp \left( -\frac{\eta^2 n}{2M^2} \right) \leq \delta$ for sample number $n$ in (44) to get

$$n \geq \frac{2M^2(\log(2/\delta) + m \log(24ML/\eta))}{\eta^2}$$

(46)

We deduce from (45) and L-Lipschtiz of $|T - T_j| \leq L\eta = \epsilon/8$, with probability $1 - \delta$, for all $T$ and $T_j$

$$|E_{\pi_o} \{ T \} - E_{\pi_o,n} \{ T \}| \leq |E_{\pi_o} \{ T \} - E_{\pi_o} \{ T_j \}| + |E_{\pi_o} \{ T_j \} - E_{\pi_o,n} \{ T_j \}| + |E_{\pi_o,n} \{ T_j \} - E_{\pi_o,n} \{ T \}|$$

$$\leq \frac{\eta}{8} + \frac{\eta}{4} + \frac{\eta}{8}$$

$$= \frac{\eta}{2}$$

(47)

Similarly, we also obtain that for all functions $e^T$, with probability at least $1 - \delta$,

$$|\log E_{\xi_o} \{ e^T \} - \log E_{\xi_o,n} \{ e^T \}| \leq \frac{\eta}{2}$$

(48)

Finally, using (26), (47) and (48), for all $T$

$$|S_{\mathcal{H},c} \{ \mu_n, \nu_n \} - S_{\mathcal{H},c} \{ \mu, \nu \}|$$

$$\leq \epsilon \left( \sup_T |E_{\pi_o} \{ T \} - E_{\pi_o} \{ T \}| + \sup_T |\log (E_{\xi_o} \{ e^T \}) - \log (E_{\xi_o,n} \{ e^T \})| \right)$$

$$\leq \epsilon \left( \frac{\eta}{2} + \frac{\eta}{2} \right)$$

$$= \epsilon \eta$$

(49)

\[ \square \]

References

[1] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of sinkhorn divergences. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 1574–1583. PMLR, 2019.

[2] C. A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. Journal of Machine Learning Research, 7(Dec):2651–2667, 2006.

[3] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyond. arXiv preprint arXiv:1605.09522, 2016.
[4] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

[5] I. Tolstikhin, B. K. Sriperumbudur, and K. Muandet. Minimax estimation of kernel mean embeddings. *The Journal of Machine Learning Research*, 18(1):3002–3048, 2017.

[6] D.-X. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 49(7):1743–1752, 2003.