Hilbert Sinkhorn Divergence for Optimal Transport Supplementary

Qian Li[1*] Zhichao Wang[2*†] Gang Li[3] Jun Pang[4] Guandong Xu[1]

[1] Faculty of Engineering and Information Technology, University of Technology Sydney, Australia

[2] School of Electrical Engineering and Telecommunications, University of New South Wales, Australia

[3] Centre for Cyber Security Research and Innovation, Deakin University, Geelong, VIC 3216, Australia

[4] Faculty of Science, Technology and Medicine, University of Luxembourg

{qian.li, guandong.xu}@uts.edu.au, zchaoking@gmail.com, gang.li@deakin.edu.au, jun.pang@uni.lu

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

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

$$\mathcal{S}_\epsilon\left(\phi_*\mu, \phi_*\nu\right) = \inf_{\pi_\phi} \int_{\mathcal{H}\times\mathcal{H}} c_\phi(u,v) d\pi_\phi(u,v) + \epsilon\Phi(\pi_\phi) \tag{1}$$

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

$$c_\phi(u,v) = \|u - v\|_{\mathcal{H}}^2$$

$$\Phi(\pi_\phi) = \log\left(\frac{d\pi_\phi}{d\left(\phi_*\mu\right) d\left(\phi_*\nu\right)}(u,v)\right)$$

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

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

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

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

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

$$\mathcal{S}_\epsilon\left(\phi_*\mu, \phi_*\nu\right) = \inf_{\pi_\phi} \int_{\mathcal{H}\times\mathcal{H}} c_\phi(u,v) d\pi_\phi(u,v) + \epsilon\Phi(\pi_\phi) \tag{4}$$

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

$$c_\phi(u,v) = \|u - v\|_{\mathcal{H}}^2$$

$$\Phi(\pi_\phi) = \log\left(\frac{d\pi_\phi}{d\left(\phi_*\mu\right) d\left(\phi_*\nu\right)}(u,v)\right)$$

Definition 4. *Given measurable spaces (X_1, Σ_1) and (X_2, Σ_2), a measurable mapping $f : X_1 \to X_2$ and a measure $\mu : \Sigma_1 \to [0, +\infty]$, the pushforward of μ is defined to be the measure $f_*(\mu) : \Sigma_2 \to [0, +\infty]$ given by*

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

*Equal contribution

†Corresponding author

1. Proving Theorem 1

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

$$\mathcal{S}_{\mathcal{H},\epsilon}(\mu,\nu) = \inf_{\pi} \int_{\mathcal{X}\times\mathcal{X}} c_{\mathcal{H}}(x,y)d\pi(x,y) + \epsilon H(\pi) \tag{6}$$

where $\pi \in \Pi(\mu,\nu)$ is the joint probability measure on $\mathcal{X}\times\mathcal{X}$ with marginals μ and ν, and

$$c_{\mathcal{H}}(x,y) = \|\phi(x)-\phi(y)\|_{\mathcal{H}}^2 = k(x,x)+k(y,y)-2k(x,y)$$
$$H(\pi) = \log\left(\frac{d\pi}{d\mu\, d\nu}(x,y)\right)$$

Then we have the following conclusions:

- $\mathcal{S}_{\mathcal{H},\epsilon}(\mu,\nu) = \mathcal{S}_{\epsilon}\left(\phi_*\mu, \phi_*\nu\right)$
- *If π^* 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 $\left((\phi,\phi)_*\pi\right)(u,v) = \pi\left(\phi^{-1}(u),\phi^{-1}(v)\right) = \pi(x,y)$. Thus, the HSD is reformulated as

$$\begin{aligned}
\mathcal{S}_{\mathcal{H},\epsilon}(\mu,\nu) &= \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} \|\phi(x)-\phi(y)\|_{\mathcal{H}}^2 d\pi(x,y) + \epsilon\log\left(\frac{d\pi}{d\mu d\nu}(x,y)\right)\\
&= \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{H}\times\mathcal{H}} \left(\|u-v\|_{\mathcal{H}}^2\right) d(\phi,\phi)_*\pi(u,v) + \epsilon\log\left(\frac{d(\phi,\phi)_*\pi}{d\left(\phi_*\mu\right)d\left(\phi_*\nu\right)}(u,v)\right)\\
&= \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{H}\times\mathcal{H}} \left(\|u-v\|_{\mathcal{H}}^2\right) d\pi_\phi(u,v) + \epsilon\log\left(\frac{d\pi_\phi}{d\left(\phi_*\mu\right)d\left(\phi_*\nu\right)}(u,v)\right)\\
&\geq \inf_{\pi_\phi\in\Pi(\phi_*\mu,\phi_*\nu)} \int_{\mathcal{H}\times\mathcal{H}} \left(\|u-v\|_{\mathcal{H}}^2\right) d\pi_\phi(u,v) + \epsilon\log\left(\frac{d\pi_\phi}{d\left(\phi_*\mu\right)d\left(\phi_*\nu\right)}(u,v)\right)\\
&= \mathcal{S}_{\epsilon}\left(\phi_*\mu,\phi_*\nu\right)
\end{aligned} \tag{7}$$

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

$$\begin{aligned}
&\int_{\mathcal{H}\times\mathcal{H}} \left(\|u-v\|_{\mathcal{H}}^2\right) d\pi_\phi(u,v) + \epsilon\log\left(\frac{d\pi_\phi}{d\left(\phi_*\mu\right)d\left(\phi_*\mu\right)}(u,v)\right)\\
=&\int_{\mathcal{H}\times\mathcal{H}} \left(\|u-v\|_{\mathcal{H}}^2 + \epsilon\log\left(\frac{d(\phi,\phi)_*\pi}{d\left(\phi_*\mu\right)d\left(\phi_*\nu\right)}(u,v)\right)\right) d(\phi,\phi)_*\pi(u,v)\\
=&\int_{\mathcal{X}\times\mathcal{X}} \left(\|\phi(x)-\phi(y)\|_{\mathcal{H}}^2\right) d\pi(x,y) + \epsilon\log\left(\frac{d\pi}{d\mu d\nu}(x,y)\right)\\
\geq&\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} \left(\|\phi(x)-\phi(y)\|_{\mathcal{H}}^2\right) d\pi(x,y) + \epsilon\log\left(\frac{d\pi}{d\mu d\nu}(x,y)\right)\\
=&\mathcal{S}_{\mathcal{H},\epsilon}(\mu,\nu)
\end{aligned} \tag{8}$$

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

$$\mathcal{S}_{\epsilon}\left(\phi_*\mu,\phi_*\nu\right) = \mathcal{S}_{\mathcal{H},\epsilon}(\mu,\nu) \tag{9}$$

If π^* is a minimizer of (6), then

$$\begin{aligned}
&\mathcal{S}_{\mathcal{H},\epsilon}(\mu,\nu)\\
&= \int_{\mathcal{X}\times\mathcal{X}} \|\phi(x)-\phi(y)\|_{\mathcal{H}}^2 d\pi^*(x,y) + \epsilon\log\left(\frac{d\pi^*}{d\mu d\nu}(x,y)\right)\\
&= \int_{\mathcal{H}\times\mathcal{H}} \left(\|u-v\|_{\mathcal{H}}^2\right) d(\phi,\phi)_*\pi^*(u,v) + \epsilon\log\left(\frac{d(\phi,\phi)_*\pi^*}{d\left(\phi_*\mu\right)d\left(\phi_*\nu\right)}(u,v)\right)\\
&= \mathcal{S}_{\epsilon}\left(\phi_*\mu,\phi_*\nu\right)
\end{aligned} \tag{10}$$

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

2. Proving Proposition 1

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

$$\mathcal{S}_\epsilon\left(\phi_* \mu, \phi_* \nu\right) = \epsilon\left(1 + \min_{\pi_\phi} \mathbb{E}_{\pi_\phi}[T] - \log\left(\mathbb{E}_{\xi_\phi}\left[e^T\right]\right)\right) \tag{11}$$

where the infimum is taken over $\pi_\phi \in \Pi\left(\phi_* \mu, \phi_* \nu\right)$, $\xi_\phi(x, y) = e^{-d(x,y)/\epsilon}$ *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 ξ_ϕ on $\mathcal{H} \times \mathcal{H}$, we define the Kullback-Leibler (KL) divergence

$$\mathrm{KL}(\pi_\phi \mid \xi_\phi) = \int_{\mathcal{H}\times\mathcal{H}} \pi(u, v)\left[\ln \frac{\pi(u, v)}{\xi(u, v)} - 1\right] \mathrm{d}u \mathrm{d}v \tag{12}$$

We can associate $\|u - v\|_{\mathcal{H}}^2$ to a Gibbs distribution

$$\xi_\phi(u, v) = e^{-\|u-v\|_{\mathcal{H}}^2/\epsilon}, \text{ then } \|u - v\|_{\mathcal{H}}^2 = -\epsilon \ln \xi_\phi(u, v) \tag{13}$$

By combining KL divergence (12) and Gibbs distribution (13) algebraically, Hilbert Sinkhorn divergence (4) can be computed as the smallest KL divergence between coupling π_ϕ and Gibbs distribution ξ_ϕ in the reproducing kernel Hilbert space:

$$\mathcal{S}_\epsilon\left(\phi_* \mu, \phi_* \nu\right) = \epsilon\left(1 + \min_{\pi_\phi \in \Pi(\phi_* \mu, \phi_* \nu)} : \mathrm{KL}\left(\pi_\phi \mid \xi_\phi\right)\right) \tag{14}$$

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

$$\mathrm{KL}\left(\pi_\phi \mid \xi_\phi\right) = \sup_{T:\mathcal{H}\times\mathcal{H}\to\mathbb{R}} \mathbb{E}_{\pi_\phi}[T] - \log\left(\mathbb{E}_{\xi_\phi}\left[e^T\right]\right) \tag{15}$$

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

$$\mathbb{E}_{\pi_\phi}[T] - \log Z = \mathbb{E}_{\pi_\phi}\left[\log \frac{d\mathbb{G}}{d\xi_\phi}\right] \tag{16}$$

Let Δ be the gap,

$$\Delta := \mathrm{KL}\left(\pi_\phi \mid \xi_\phi\right) - \left(\mathbb{E}_{\pi_\phi}[T] - \log\left(\mathbb{E}_{\xi_\phi}\left[e^T\right]\right)\right) \tag{17}$$

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

$$\Delta = \mathbb{E}_{\pi_\phi}\left[\log \frac{d\pi_\phi}{d\xi_\phi} - \log \frac{d\mathbb{G}}{d\xi_\phi}\right] = \mathbb{E}_{\pi_\phi} \log \frac{d\pi_\phi}{d\mathbb{G}} = \mathrm{KL}(\pi_\phi \| \mathbb{G}) \tag{18}$$

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

$$\mathrm{KL}(\pi_\phi \| \xi_\phi) \geq \mathbb{E}_{\pi_\phi}[T] - \log\left(\mathbb{E}_{\xi_\phi}\left[e^T\right]\right) \tag{19}$$

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

$$T = \log \frac{d\pi_\phi}{d\xi_\phi} + C \tag{20}$$

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

3. Proving Proposition 2

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

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

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

Proof. KL divergence (15) in product space $\mathcal{H} \times \mathcal{H}$ satisfies

$$\begin{aligned}
\mathrm{KL}\left(\pi_\phi \mid \xi_\phi\right) &= \sup_{T:\mathcal{H}\times\mathcal{H}\to\mathbb{R}} \mathbb{E}_{\pi_\phi}[T] - \log\left(\mathbb{E}_{\xi_\phi}\left[e^T\right]\right) \\
&= \sup_{T:\mathcal{H}\times\mathcal{H}\to\mathbb{R}} \int_{\mathcal{H}\times\mathcal{H}} T(u,v)d\pi_\phi(u,v) - \log\int_{\mathcal{H}\times\mathcal{H}} e^T d\xi_\phi(u,v) \\
&= \sup_{T:\mathcal{H}\times\mathcal{H}\to\mathbb{R}} \int_{\mathcal{H}\times\mathcal{H}} T(u,v)d\left((\phi,\phi)_*\pi\right)(u,v) - \log\int_{\mathcal{H}\times\mathcal{H}} e^{T(u,v)} d\left((\phi,\phi)_*\xi\right)(u,v) \\
&\overset{(5)}{=} \sup_{T:\mathcal{H}\times\mathcal{H}\to\mathbb{R}} \int_{\Omega\times\Omega} T(\phi(x),\phi(y))d\pi(x,y) - \log\int_{\Omega\times\Omega} e^{T(\phi(x),\phi(y))} d\xi
\end{aligned} \tag{21}$$

Let the set of inner products be $\mathcal{M} = \{\langle\cdot,\cdot\rangle : \mathcal{H}\times\mathcal{H}\to\mathbb{R}\}$. Clearly, $\mathcal{M} \subseteq \{T : \mathcal{H}\times\mathcal{H}\to\mathbb{R}\}$ for all functional maps T define on $\mathcal{H}\times\mathcal{H}$. If functional map $T \in \mathcal{M}$, then we have $T(\phi(x),\phi(y)) = \langle\phi(x),\phi(y)\rangle = k(x,y)$ where $k : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ is a kernel function. Hence, we continue (21) to get

$$\begin{aligned}
&\geq \sup_{T\in\mathcal{M}} \int_{\Omega\times\Omega} T(\phi(x),\phi(y))d\pi(x,y) - \log\int_{\Omega\times\Omega} e^{T(\phi(x),\phi(y))} d\xi(x,y) \\
&= \sup_k \int_{\Omega\times\Omega} k(x,y)d\pi(x,y) - \log\int_{\Omega\times\Omega} e^{K(x,y)} d\xi(x,y) \\
&= \sup_{K:\mathcal{X}\times\mathcal{X}\to\mathbb{R}} \mathbb{E}_\pi[k] - \log\left(\mathbb{E}_\xi\left[e^k\right]\right)
\end{aligned} \tag{22}$$

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

$$\begin{aligned}
\mathcal{S}_\epsilon\left(\phi_*\mu, \phi_*\nu\right) &= \epsilon\left(1 + \min_{\pi_\phi\in\pi_\phi} \mathrm{KL}\left(\pi_\phi \mid \xi_\phi\right)\right) \\
&\geq \epsilon\left(1 + \min_{\pi\in\Pi} \sup_{k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}} \mathbb{E}_\pi[k] - \log\left(\mathbb{E}_\xi\left[e^k\right]\right)\right) \\
&\geq \epsilon\left(1 + \min_{\pi\in\Pi} \mathbb{E}_\pi[k] - \log\left(\mathbb{E}_\xi\left[e^k\right]\right)\right)
\end{aligned} \tag{23}$$

$\square$

Since $\mathcal{S}_{\mathcal{H},\epsilon}(\mu,\nu) = \mathcal{S}_\epsilon\left(\phi_*\mu, \phi_*\nu\right)$ as provided in Theorem 1. As a consequence, we directly have the following result by using Proposition 1 and 2.

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

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

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

The related notations are defined in Prop 1 and 2.

4. Proving Theorem 2

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

$$\forall n \geq N,\ \mathbb{P}\left(\left|\mathcal{S}_{\mathcal{H},\epsilon}(\mu_n, \nu_n) - \mathcal{S}_{\mathcal{H},\epsilon}(\mu, \nu)\right| \leq \epsilon\eta\right) = 1 \tag{24}$$

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

$$\mathcal{S}_{\mathcal{H},\epsilon}\left(\mu_n, \nu_n\right) = \epsilon\left(1 + \min_{\pi_{\phi,n} \in \Pi_{\phi,n}} : \mathrm{KL}\left(\pi_{\phi,n} \mid \xi_{\phi,n}\right)\right) \tag{25}$$

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

$$\left|S_{\mathcal{H},\epsilon}\left(\mu_n, \nu_n\right) - S_{\mathcal{H},\epsilon}(\mu, \nu)\right| \leq \epsilon\left(\sup_{T:\mathcal{H}\times\mathcal{H}\to\mathbb{R}} \left|\mathbb{E}_{\pi_{\phi,n}}[T] - \mathbb{E}_{\pi_\phi}[T]\right| + \sup_{T:\mathcal{H}\times\mathcal{H}\to\mathbb{R}} \left|\log\left(\mathbb{E}_{\xi_\phi}\left[e^T\right]\right) - \log\left(\mathbb{E}_{\xi_{\phi,n}}\left[e^T\right]\right)\right|\right) \tag{26}$$

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

$$\left|\log\left(\mathbb{E}_{\xi_\phi}\left[e^T\right]\right) - \log\left(\mathbb{E}_{\xi_{\phi,n}}\left[e^T\right]\right)\right| \leq e^M \left|\mathbb{E}_{\xi_\phi}\left[e^T\right] - \mathbb{E}_{\xi_{\phi,n}}\left[e^T\right]\right| \tag{27}$$

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

$$\sup_{T:\mathcal{H}\times\mathcal{H}\to\mathbb{R}} \left|\mathbb{E}_{\pi_{\phi,n}}[T] - \mathbb{E}_{\pi_\phi}[T]\right| \leq \frac{\eta}{2} \quad \text{and} \quad \sup_{T:\mathcal{H}\times\mathcal{H}\to\mathbb{R}} \left|\log\left(\mathbb{E}_{\xi_\phi}\left[e^T\right]\right) - \log\left(\mathbb{E}_{\xi_{\phi,n}}\left[e^T\right]\right)\right| \leq \frac{\eta}{2} e^{-M} \tag{28}$$

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

$$\forall n \geq N, \quad \left|\mathcal{S}_{\mathcal{H},\epsilon}(\mu_n, \nu_n) - \mathcal{S}_{\mathcal{H},\epsilon}(\mu, \nu)\right| \leq \frac{\epsilon\eta}{2} + \frac{\epsilon\eta}{2} = \epsilon\eta \tag{29}$$

with probability one. □

5. Proving Proposition 3

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

$$\begin{aligned} &\left|\mathcal{S}_{\mathcal{H},\epsilon}(\mu, \nu) - \mathcal{W}_\epsilon(\mu, \nu)\right| \leq \epsilon\eta \\ &\left|\mathcal{S}_{\mathcal{H},\epsilon}(\mu, \nu) - \mathcal{W}(\mu, \nu)\right| \leq \epsilon\left(\eta + 2d\log\frac{e^2 LD}{\sqrt{d}\epsilon}\right) \end{aligned} \tag{30}$$

where $\epsilon > 0$, $D \geq |\mathcal{X}|$ and L is a Lipschitz constant.

Proof. **Step 1**. Notice that we can follow the idea of Proposition 1 to construct the following representation in Euclidean space

$$\mathcal{W}_\epsilon(\mu, \nu) = \epsilon\left(1 + \min_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_\pi[T] - \log\left(\mathbb{E}_\xi\left[e^T\right]\right)\right) \tag{31}$$

where $\xi(x, y) = e^{-d(x,y)/\epsilon}$ and function $T = \log\frac{d\pi}{d\xi}$. By construction, T satisfies $\mathbb{E}_\xi\left[e^T\right] = \int d\pi = 1$.

Without loss of generality, we assume that π makes the minimum of $\mathbb{E}_\pi[K] - \log\left(\mathbb{E}_\xi\left[e^K\right]\right)$ appeared in (**??**). Then

$$\begin{aligned} \mathcal{W}_\epsilon(\mu, \nu) - \mathcal{S}_{\mathcal{H},\epsilon}(\mu, \nu) &\leq \epsilon\left(\mathbb{E}_\pi[T] - \log\left(\mathbb{E}_\xi\left[e^T\right]\right)\right) - \epsilon\left(\mathbb{E}_\pi[K] - \log\left(\mathbb{E}_\xi\left[e^K\right]\right)\right) \text{ by (31) and (??)} \\ &= \epsilon\left(\left(\mathbb{E}_\pi[T] - \log 1\right) - \mathbb{E}_\pi[K] + \log\left(\mathbb{E}_\xi\left[e^K\right]\right)\right) \\ &= \epsilon\left(\left(\mathbb{E}_\pi[T] - \mathbb{E}_\pi[K]\right) + \log\left(\mathbb{E}_\xi\left[e^K\right]\right)\right) \\ &\leq \epsilon\left(\left(\mathbb{E}_\pi[T] - \mathbb{E}_\pi[K]\right) + \left(\mathbb{E}_\xi\left[e^K\right] - 1\right)\right) \\ &= \epsilon\left(\left(\mathbb{E}_\pi[T - K]\right) + \left(\mathbb{E}_\xi\left[e^K - e^T\right]\right)\right) \end{aligned} \tag{32}$$

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 [2], we can choose a kernel function $K \leq M$ such that

$$\mathbb{E}_\pi |T - K| \leq \frac{\eta}{2} \quad \text{and} \quad \mathbb{E}_\xi |T - K| \leq \frac{\eta}{2} e^{-M} \tag{33}$$

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

$$\mathbb{E}_\xi \left|e^T - e^K\right| \leq e^M \mathbb{E}_\xi |T - K| \leq \frac{\eta}{2} \tag{34}$$

From (32)-(34) and the triangular inequality, we then obtain

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

which proves (30).

Step 3. In this step, we are interested in bounding the error made when approximating $W(\mu, \nu)$ with $\mathcal{S}_{\mathcal{H},\epsilon}(\mu, \nu)$. Assume $\mathcal{X}$ is the subsets of $\mathbb{R}^d$, the diameter $|\mathcal{X}| = \sup \{\|x - x'\| \mid x, x' \in \mathcal{X}\} \leqslant D$ and the cost function is L-Lipschitz. Then it holds [1]

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

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

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

□

6. Proving Theorem 3

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

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

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

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:

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

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

7. Proving Theorem 4

Lemma 1. *[6] 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*

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

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 η, we have,*

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

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. 1.

Proof. Assume functional T is M-bounded and L-Lipschtiz in reproducing kernel Hilbert space. By Hoeffding inequality, for all function $|f| \leq M$

$$\Pr\left(|\mathbb{E}_{\mu} f - \mathbb{E}_{\mu_n} f| > \frac{\eta}{4}\right) \leq 2\exp\left(-\frac{\eta^2 n}{2M^2}\right) \tag{42}$$

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 η. We need to choose a minimal cover number of the domain $B_R = \{T \in \mathcal{H} : \|T\|_{\mathcal{H}} \leq R\}$ by a finite set of small balls with radius η 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

$$\mathcal{N}(\mathcal{H}, \eta) \leq \left(\frac{3M}{\eta}\right)^m \tag{43}$$

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

$$\Pr\left(\sup_j \left|\mathbb{E}_{\pi_\phi}[T_j] - \mathbb{E}_{\pi_{\phi,n}}[T_j]\right| \geq \frac{\eta}{4}\right) \leq 2\mathcal{N}(\mathcal{H}, \eta)\exp\left(-\frac{\eta^2 n}{2M^2}\right) < \delta \tag{44}$$

which gives

$$\Pr\left(\sup_j \left|\mathbb{E}_{\pi_\phi}[T_j] - \mathbb{E}_{\pi_{\phi,n}}[T_j]\right| \leq \frac{\eta}{4}\right) > 1 - \delta \tag{45}$$

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

$$n \geq \frac{2M^2(\log(2/\delta) + m\log(24ML/\eta))}{\eta^2} \tag{46}$$

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

$$\begin{aligned}
\left|\mathbb{E}_{\pi_\phi}[T] - \mathbb{E}_{\pi_{\phi,n}}[T]\right| &\leq \left|\mathbb{E}_{\pi_\phi}[T] - \mathbb{E}_{\pi_\phi}[\mathcal{I}_j]\right| + \left|\mathbb{E}_{\pi_\phi}[T_j] - \mathbb{E}_{\pi_{\phi,n}}[T_j]\right| + \left|\mathbb{E}_{\pi_{\phi,n}}[T_j] - \mathbb{E}_{\pi_{n,\mathcal{H}}}[T]\right| \\
&\leq \frac{\eta}{8} + \frac{\eta}{4} + \frac{\eta}{8} \\
&= \frac{\eta}{2}
\end{aligned} \tag{47}$$

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

$$\left|\log \mathbb{E}_{\xi_\phi}\left[e^T\right] - \log \mathbb{E}_{\xi_{\phi,n}}\left[e^T\right]\right| \leq \frac{\eta}{2} \tag{48}$$

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

$$\begin{aligned}
&|S_{\mathcal{H},\epsilon}(\mu_n, \nu_n) - S_{\mathcal{H},\epsilon}(\mu, \nu)| \\
\leq&\epsilon\left(\sup_T \left|\mathbb{E}_{\pi_{\phi,n}}[T] - \mathbb{E}_{\pi_\phi}[T]\right| + \sup_T \left|\log\left(\mathbb{E}_{\xi_\phi}\left[e^T\right]\right) - \log\left(\mathbb{E}_{\xi_{\phi,n}}\left[e^T\right]\right)\right|\right) \\
\leq&\epsilon\left(\frac{\eta}{2} + \frac{\eta}{2}\right) \\
=&\epsilon\eta
\end{aligned} \tag{49}$$

□

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.