# Conditional Bures Metric for Domain Adaptation (Supplementary Material)

You-Wei Luo[1] Chuan-Xian Ren[1,2*]
[1]School of Mathematics, Sun Yat-Sen University, China
[2]Pazhou Lab, Guangzhou, China
luoyw28@mail2.sysu.edu.cn, rchuanx@mail.sysu.edu.cn

## Abstract

*This supplementary material contains the proofs of theorems and some details on the experiment setting: 1) we present the discussions on the proposed method; 2) we show some details on the proposed conditional distribution matching network; 3) we present an additional comparison experiment; 4) we introduce some basic definition and conclusions on the bounded operators and Hilbert spaces as a preliminary of proofs; 5) we present the proofs of all theorems and propositions in the paper.*

## S.1. Discussions

**How CKB helps knowledge transfer?**

As we aim at the classification-oriented transfer, where the task knowledge in $P_{X|Y}$ needs to be considered during the transfer of $X$, the aligned conditional distributions ($P^s_{X|Y} = P^t_{X|Y}$) are essential to the successful generalization of source predictor. To achieve this, we develop a rigorously defined conditional discrepancy metric (*i.e.*, CKB) which has not been explored in transfer learning. Theoretically, recent advancement [37] shows the target generalization error is bounded by the discrepancy between the optimal classifiers on two domains. Since the error rate of source classifier will be small on the target domain with aligned conditional distribution $P_{X|Y}$, the CKB-based conditional alignment is equivalent to minimize the discrepancy between the classifiers.

**Relation to other metrics.**

**Connection:** MMD, kernel Bures, and CKB are all kernel embedding metrics. Besides, CKB metric is essentially the minimized transport cost of the class-wise kernel OT.

**Difference:** CKB metric is directly built on the conditional distributions, while MMD and kernel Bures are the marginal distribution embedding metrics. Compared with the class-wise MMD [38] and OT, CKB is not only well-defined with infinite conditions (*i.e.*, $Y$ is continuous), but also estimates the conditional distribution with data from all conditions. As a closed-form solution, CKB also avoids the burdens of optimization in class-wise OT. Note the class-wise computation strategy estimates the conditional discrepancy separately and will be inefficient when $Y$ is continuous or $|\mathcal{Y}|$ is large. Therefore, CKB is more effective when there are more categories.

**About label shift and limitations.**

Label shift investigates the shifting joint distribution $P_{XY}$ via $P_{XY} = P_Y P_{X|Y}$. Current label shift and GAN-like methods [11, 22, 28, 35] usually give an implicit approximate solution or make a strong assumption on $P_{X|Y}$. These methods commonly assume that there exists a transformation or generator $F(\cdot) : X \to Z$ s.t. $P^s_{Z|Y} = P^t_{Z|Y}$. However, neither the practical sufficient condition for the existence of $F(\cdot)$ nor its explicit modeling are explored. To overcome this bottleneck, our work builds an explicit rule for $F(\cdot)$ with the conditional discrepancy metric CKB, which helps the classification knowledge transfer by minimizing the CKB distance between domains. Further, the MMD on $P_Y$ is applied to align the shifting $P_{XY}$, which is actually less flexible and is the main limitation.

*Corresponding Author.

## S.2. Experimental Details

We implement the proposed conditional distribution matching network in PyTorch [1] platform. The framework of the conditional distribution matching network is presented in Figure 1. For ImageCLEF-DA and Office-Home datasets, we use ResNet-50 [2] as the backbone DNNs in Figure 1. Follow the standard protocol, AlexNet [3] and modified LeNet [4] are adopted as the backbones in Office10 and Digits datasets. We use Adam Optimizer ($lr = 0.0002$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) with batch size of 40 for model training. Based on the trial-and-error approach, the hyper-parameters $\lambda_1$ and $\lambda_2$ are set as $5e$-2 and $1e0$, respectively. The "Bures" and "Kernel Bures" in **Ablation** experiment mean that we replace the $\mathcal{L}_{\text{CKB}}$ in CKB model with Eq. (6) and Eq. (7), respectively. We report their best results via the same grid search as shown in the paper. For all the datasets, we randomly repeat the experiments for 10 times. All experiments are performed on an Ubuntu 18.04 operating system PC with an Intel Core i7-6950X 3.00GHz CPU PC, 64G RAM and an NVIDIA TITAN Xp GPU.

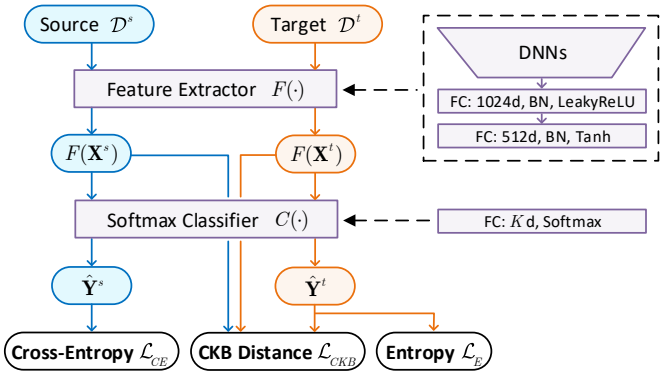


Figure 1. Network architecture of the proposed conditional distribution matching network. "FC" means the fully connected layer and its following elements respectively represent the dimensions after the projection, batch normalization and activation.

## S.3. Experiment on Refurbished Office-31

Refurbished Office-31 [5] is recently released for extending the major problems of the Office-31 dataset. Specifically, Refurbished Office-31 replaces a total of 834 on the Amazon domain in the original Office-31 dataset, while the sample size of Amazon domain is still 2,817. For the DSLR and Webcam domains, they are kept the same as in the original Office-31 dataset. We denote the Amazon, DSLR and Webcam domains in Refurbished Office-31 as $\mathbf{A_{ref}}$, **W**, **D**, respectively. The results in Table 1 show that the proposed methods are competitive with other SOTA UDA methods. Especially, the proposed methods achieve the highest accuracies on tasks **D**→**W** and **W**→**D**. Since there are less ground-truth labels in tasks **D**→$\mathbf{A_{ref}}$ and **W**→$\mathbf{A_{ref}}$, the empirical estimation of CKB metric may be less effective and the classification accuracies are lower. Defining a more accurate pseudo-labeling strategy will further boost the CKB-based adaptation models.

Table 1. Accuracies (%) on Refurbished Office-31 (ResNet-50).

| **Image-CLEF-DA** | $A_{ref} \to D$ | $A_{ref} \to W$ | $D \to A_{ref}$ | $D \to W$ | $W \to A_{ref}$ | $W \to D$ | Mean |
|---|---|---|---|---|---|---|---|
| Source [2] | 79.2 ± 0.6 | 76.8 ± 1.0 | 73.5 ± 1.2 | 96.3 ± 0.2 | 74.1 ± 0.5 | 99.1 ± 0.3 | 83.2 |
| RSDA-DANN [6] | 90.9 ± 1.3 | 91.8 ± 0.5 | 87.3 ± 0.6 | 98.8 ± 0.2 | 90.5 ± 0.9 | 99.9 ± 0.1 | 93.2 |
| RSDA-MSTN [6] | 93.2 ± 1.0 | 92.2 ± 0.3 | 91.7 ± 0.9 | 99.1 ± 0.2 | **93.0** ± 0.6 | **100.0** ± 0.0 | **94.9** |
| SymNet [7] | 92.4 ± 0.4 | 91.0 ± 0.2 | 90.6 ± 0.4 | 98.0 ± 0.1 | 89.2 ± 0.4 | 99.8 ± 0.0 | 93.5 |
| CAN [8] | **94.4** ± 0.3 | **92.8** ± 0.5 | **92.3** ± 0.8 | 98.5 ± 0.2 | 90.9 ± 1.0 | 99.7 ± 0.2 | 94.8 |
| CKB | 93.0 ± 1.0 | 92.3 ± 0.3 | 88.9 ± 1.1 | **99.3** ± 0.3 | 86.2 ± 0.9 | **100.0** ± 0.0 | 93.3 |
| CKB+MMD | 92.8 ± 1.2 | 91.8 ± 0.4 | 88.9 ± 0.9 | **99.3** ± 0.2 | 86.0 ± 1.1 | **100.0** ± 0.0 | 93.1 |

## S.4. Proofs

### S.4.1. Preliminary: Operators on Hilbert Space

Before the proofs, we briefly review the operators on Hilbert space. Let $T$ be a operator on Hilbert space $\mathcal{H}$, the adjoint of $T$ is denoted by $T^*$ and $T$ is called self-adjoint if $T^* = T$. The following lemma proves that there exists a unique square root for any positive operator.

**Lemma S.1 ([9], Theorem VI.9, square root lemma)** *Let $T$ be a positive operator on $\mathcal{H}$. Then there is a unique $\sqrt{T}$ on $\mathcal{H}$ such that $\sqrt{T}$ is positive, self-adjoint and $(\sqrt{T})^2 = T$.*

With the square root, we define the absolute value of $T$ as $|T| = \sqrt{T^*T}$. Let $\{\varphi_n\}_{n=1}^{\infty}$ be an orthonormal basis in $\mathcal{H}$. Note that the absolute value of a self-adjoint operator $T$ is the operator itself, i.e., $|T| = \sqrt{T^*T} = \sqrt{T^2} = T$. For any positive operator $T$, its trace is defined by $\text{tr}(T) = \sum_n \langle \varphi_n, T\varphi_n \rangle$. An operator $T$ is called trace class and Hilbert-Schmidt class if and only if its trace-norm $\|T\|_1 = \text{tr}(\sqrt{T^*T}) = \text{tr}(|T|)$ (which is also known as nuclear norm $\|\cdot\|_*$ in vector space) and Hilbert-Schmidt-norm $\|T\|_2 = \sqrt{\text{tr}(T^*T)}$ are finite, respectively. The sets of all trace class and Hilbert-Schmidt class operators are denoted by $\mathscr{B}_1$ and $\mathscr{B}_2$, respectively. We denote $T_1 \geq T_2$ if $T_1 - T_2$ is positive.

**Lemma S.2 ([9], Theorem VI.20, VI.22)** *For any $T_1, T_2 \in \mathbb{S}^+(\mathcal{H})$,*
*(a) $\mathscr{B}_1$ is a Banach space with norm $\|\cdot\|_1$.*
*(b) $\mathscr{B}_2$ is a Hilbert space with inner product $\langle T_1, T_2 \rangle_2 = \text{tr}(T_1^* T_2)$.*
*(c) $\|\cdot\|_2 \leq \|\cdot\|_1$ and $\|T_1 T_2\|_1 \leq \|T_1\|_2 \|T_2\|_2$.*
*(d) $\|T_1\|_1 = \|T_1^*\|_1$ and $\|T_1\|_2 = \|T_1^*\|_2$.*

The above lemma shows that $\mathscr{B}_1$ and $\mathscr{B}_2$ are vector space, and $T_1 T_2 \in \mathscr{B}_1$ if $T_1, T_2 \in \mathscr{B}_2$. The next lemma shows that the distance between the square roots of any two positive operators is bounded by the square distance between the operators. This lemma will be used to guarantee the convergence of the square root of conditional covariance operator in the proof of Theorem 4.

**Lemma S.3 ([10], Lemma 4.1)** *Let $T_1$ and $T_2$ be positive operators on Hilbert space $\mathcal{H}$. Then*

$$\|\sqrt{T_1} - \sqrt{T_2}\|_2^2 \leq \|T_1 - T_2\|_1.$$

Based on the above lemmas, we derive some properties of the operators on $\mathbb{S}^+(\mathcal{H})$ which is the set of all positive, self-adjoint and trace class operator on $\mathcal{H}$. The following properties will be used in the proofs of Proposition 1 and Theorem 4.

**Corollary S.1** *Let $T_1, T_2 \in \mathbb{S}^+(\mathcal{H})$, then*
*(a) $\|T_1\|_1 = \text{tr}(T_1)$.*
*(b) $\|T_1\|_1 = \|\sqrt{T_1}\|_2^2$ which means $\sqrt{T_1} \in \mathscr{B}_2$.*
*(c) $\|\sqrt{T_1}\sqrt{T_2}\|_1 \leq \|\sqrt{T_1}\|_2 \|\sqrt{T_2}\|_2$ which means $\sqrt{T_1}\sqrt{T_2} \in \mathscr{B}_1$.*

**Proof** *(a) As $T_1$ is self-adjoint, $\|T_1\|_1 = \text{tr}(|T_1|) = \text{tr}(T_1)$.*
*(b) From (a) and Lemma S.1 (c) we know that $\sqrt{T_1}$ is also positive and self-adjoint. Then we have*

$$\|T_1\|_1 = \text{tr}(T_1) = \text{tr}(\sqrt{T_1}\sqrt{T_1}) = \text{tr}(\sqrt{T_1}^*\sqrt{T_1}) = \|\sqrt{T_1}\|_2^2.$$

*Since $T_1$ is trace class (i.e., $T_1 \in \mathscr{B}_1$), then $\|\sqrt{T_1}\|_2^2 = \|T_1\|_1 < \infty$ and $\sqrt{T_1} \in \mathscr{B}_2$.*
*(c) This can be proved from (b) and Lemma S.2 (c).*

### S.4.2. Proof of Theorem 1

**Theorem 1** *Let $(\mathcal{X}, \mathcal{B}_\mathcal{X})$ be the locally compact and Hausdorff measurable space and the reproducing kernel $k$ be $c_0$-universal. Assuming that $(\phi(X), \psi(Y))$ is a Gaussian random variable in $\mathcal{H}_\mathcal{X} \oplus \mathcal{H}_\mathcal{Y}$. For any $P^s_{X|Y}, P^t_{X|Y} \in \text{Pr}^s(\mathcal{X}|\mathcal{Y})$, we have*

$$d_{\text{CKB}}(\mathbf{R}^s_{XX|Y}, \mathbf{R}^t_{XX|Y}) = 0 \quad \Longrightarrow \quad P^s_{X|Y} = P^t_{X|Y}.$$

We first introduce an important property called 3-splitting.

**Definition S.1 (3-splitting property)** *A probability measure $P_X \in \text{Pr}(\mathcal{X})$ satisfies the 3-splitting property if there exist three disjoint subsets $\Omega_1, \Omega_2, \Omega_3 \subset \mathcal{X}$, which satisfy $P_X(\Omega_1), P_X(\Omega_2), P_X(\Omega_3) > 0$ and $\Omega_1 \cup \Omega_2 \cup \Omega_3 = \mathcal{X}$.*

The 3-splitting property is vital for the injectiveness of $P_X \to \mathbf{R}_{XX}$. Specifically, the 3-splitting property guarantees the unique covariance embedding $\mathbf{R}_{XX}$ for a certain probability measure $P_X$, i.e., $P^s_X \neq P^t_X \Longrightarrow \mathbf{R}^s_{XX} \neq \mathbf{R}^t_{XX}$. For more detailed discussions of the 3-splitting property, please refer to literature [11]. With the 3-splitting property, we present a lemma.

**Lemma S.4 ([11], Theorem 5)** *Let the measurable space $(\mathcal{Z}, \mathcal{B}_{\mathcal{Z}})$ be locally compact and Hausdorff. Let $k$ be a $c_0$-universal reproducing kernel. Then, the embedding $P_Z \to \mathbf{R}_{ZZ}, \forall P_Z \in \Pr(\mathcal{Z})$ is injective.*

Now we begin the proof of Theorem 1.

**Proof** *As $d_{CKB}(\mathbf{R}^s_{XX|Y}, \mathbf{R}^t_{XX|Y}) = 0$ and the CKB metric defines a metric on $\mathbb{S}^+(\mathcal{H}_{\mathcal{X}})$, we have $\mathbf{R}^s_{XX|Y} = \mathbf{R}^t_{XX|Y}$. Since $(\phi(X), \psi(Y))$ is a Gaussian random variable in $\mathcal{H}_{\mathcal{X}} \oplus \mathcal{H}_{\mathcal{Y}}$, then the conditional covariance $\mathbf{R}_{XX|Y=y}$ on RKHS is independent of the condition $Y = y$ [12], i.e.,*

$$\mathbf{R}_{XX|Y=y_1} = \mathbf{R}_{XX|Y=y_2}, \forall y_1, y_2, \in \mathcal{Y}. \tag{S.1}$$

*Further, we show the connection between the conditional covariance operator $\mathbf{R}_{XX|Y}$ and conditional covariance $\mathbf{R}_{XX|Y=y}$ with fixed condition $y$. We first review the kernel mean embedding operator $\mathcal{C}_{X|Y} = \mathbf{R}_{XY}\mathbf{R}^{-1}_{YY}$ [12, 13], which satisfies*

$$\mu_{X|Y=y} = \mathbb{E}_{X|Y=y}\left[\phi(X)|Y = y\right] = \mu_X + \mathcal{C}_{X|Y}(\psi(y) - \mu_Y), \quad \mathbf{R}_{XY} = \mathcal{C}_{X|Y}\mathbf{R}_{YY}.$$

*Note $\forall y_i \in \mathcal{Y}$, $\mathbf{R}_{XX|Y}$ can be rewritten as*

$$\begin{aligned}
\mathbf{R}_{XX|Y} &= \mathbb{V}_X\left[\phi(X)\right] - \mathbf{R}_{XY}\mathbf{R}^{-1}_{YY}\mathbf{R}_{YX} \\
&= \mathbb{V}_X\left[\phi(X)\right] - \mathcal{C}_{X|Y}\mathbf{R}_{YY}\mathcal{C}^*_{X|Y} \\
&= \mathbb{V}_X\left[\phi(X)\right] - \mathbb{E}_Y\left[\mathcal{C}_{X|Y}(\psi(y) - \mu_y) \otimes (\psi(y) - \mu_y)\mathcal{C}^*_{X|Y}\right] \\
&= \mathbb{V}_X\left[\phi(X)\right] - \mathbb{E}_Y\left[(\mu_{X|Y=y} - \mu_X) \otimes (\mu_{X|Y=y} - \mu_X)\right] \\
&= \mathbb{V}_X\left[\phi(X)\right] - \mathbb{V}_Y\left[\mu_{X|Y=y}\right] \\
&= \mathbb{V}_X\left[\phi(X)\right] - \mathbb{V}_Y\left[\mathbb{E}_{X|Y}\left[\phi(X)|Y\right]\right] \\
&= \mathbb{E}_Y\left[\mathbb{V}_{X|Y}\left[\phi(X)|Y\right]\right] \qquad \text{(S.2)} \\
&= \mathbb{E}_Y\left[\mathbf{R}_{XX|Y}\right] \\
&= \mathbf{R}_{XX|Y=y_i}, \qquad \text{(S.3)}
\end{aligned}$$

*where Eq. (S.3) is concluded from Eq. (S.1) and Eq. (S.2) is obtained from the Eve's law which is written as*

$$\mathbb{V}\left[\phi(X)\right] = \mathbb{E}\left[\mathbb{V}\left[\phi(X)|Y\right]\right] + \mathbb{V}\left[\mathbb{E}\left[\phi(X)|Y\right]\right].$$

*The above equations show that $\mathbf{R}_{XX|Y}$ is exactly the expectation of $\mathbf{R}_{XX|Y=y}$ over $Y$, and $\mathbf{R}_{XX|Y} = \mathbf{R}_{XX|Y=y_i}$, $\forall y_i \in \mathcal{Y}$. Now, for every fixed $y$, denote $\mathcal{Z} = \mathcal{X} \times \{y\}$. Since the finite space $\{y\}$ is always compact and the product of locally compact spaces is locally compact, the measure space $\mathcal{Z}$ is locally compact. Note $\mathcal{Z}$ is also Hausdorff, since for any two distinct points $x_1, x_2 \in \mathcal{X}$, there exist neighbourhoods of each which are disjoint and this result also holds for $(x_1, y), (x_2, y) \in \mathcal{Z}$. Then it can be concluded from Theorem S.4 that $P_{X|Y=y} \to \mathbf{R}_{XX|Y=y}$ is injective. Recall that $\mathbf{R}^s_{XX|Y} = \mathbf{R}^t_{XX|Y}$, thus, we have $\mathbf{R}^s_{XX|Y=y} = \mathbf{R}^t_{XX|Y=y}$ from Eq. (S.3) and $P^s_{X|Y=y} = P^t_{X|Y=y}$ for all $y \in \mathcal{Y}$, i.e., $P^s_{X|Y} = P^t_{X|Y}$.*

### S.4.3. Proof of Proposition 1

**Proposition 1** *The CKB metric $d_{\text{CKB}}(\cdot, \cdot)$ defines a metric on $\mathbb{S}^+(\mathcal{H}_{\mathcal{X}})$.*

**Proof** *Let $T_1, T_2, T_3 \in \mathbb{S}^+(\mathcal{H}_{\mathcal{X}})$. We first rewrite the CKB metric as*

$$\begin{aligned}
d^2_{\text{CKB}}(T_1, T_2) &= \text{tr}\left(T_1 + T_2 - 2\sqrt{\sqrt{T_1}T_2\sqrt{T_1}}\right) \\
&= \|T_1\|_1 + \|T_2\|_1 - 2\text{tr}\left(\sqrt{\sqrt{T_1}\sqrt{T_2}\sqrt{T_2}\sqrt{T_1}}\right) \qquad \text{(S.4)} \\
&= \|T_1\|_1 + \|T_2\|_1 - 2\text{tr}\left(\sqrt{(\sqrt{T_2}\sqrt{T_1})^*\sqrt{T_2}\sqrt{T_1}}\right) \\
&= \|T_1\|_1 + \|T_2\|_1 - 2\|\sqrt{T_2}\sqrt{T_1}\|_1, \qquad \text{(S.5)}
\end{aligned}$$

*where Eq. (S.4) holds from Corollary S.1 (a). It can be deduced from Corollary S.1 (c) that* $\|\sqrt{T_2}\sqrt{T_1}\|_1 \leq \infty$ *and* $d^2_{\text{CKB}}(\cdot,\cdot) \leq \infty$*. With the reformulation Eq. (S.5), we derive an important inequality for the CKB metric as follows.*

$$\begin{aligned} d^2_{\text{CKB}}(T_1,T_2) &= \|T_1\|_1 + \|T_2\|_1 - 2\|\sqrt{T_2}\sqrt{T_1}\|_1 \\ &= \|\sqrt{T_1}\|_2^2 + \|\sqrt{T_1}\|_2^2 - 2\|\sqrt{T_2}\sqrt{T_1}\|_1 \qquad \text{(S.6)} \\ &\geq \|\sqrt{T_1}\|_2^2 + \|\sqrt{T_1}\|_2^2 - 2\|\sqrt{T_2}\|_2\|\sqrt{T_1}\|_2 \qquad \text{(S.7)} \\ &= \left(\|\sqrt{T_1}\|_2 - \|\sqrt{T_2}\|_2\right)^2, \qquad \text{(S.8)} \end{aligned}$$

*where Eq. (S.6) holds from Corollary S.1 (b) and Eq. (S.7) from Corollary S.1 (c). The above inequality shows that* $d_{\text{CKB}}(T_1,T_2) \geq \left|\|\sqrt{T_1}\|_2 - \|\sqrt{T_2}\|_2\right|$*, where* $\left|\|\sqrt{T_1}\|_2 - \|\sqrt{T_2}\|_2\right|$ *is the absolute value of the real number* $\|\sqrt{T_1}\|_2 - \|\sqrt{T_2}\|_2$*. Now we begin to prove the metric properties.*
*(i) As* $d_{\text{CKB}}(T_1,T_2) \geq \left|\|\sqrt{T_1}\|_2 - \|\sqrt{T_2}\|_2\right| \geq 0$*, the CKB metric is nonnegative.*
*(ii) On the one hand, if* $T_1 = T_2$*, then it can be deduced from Eq. (S.5) that*

$$d^2_{\text{CKB}}(T_1,T_2) = \|T_1\|_1 + \|T_2\|_1 - 2\|\sqrt{T_2}\sqrt{T_1}\|_1 = \|T_1\|_1 + \|T_1\|_1 - 2\|\sqrt{T_1}\sqrt{T_1}\|_1 = 0.$$

*On the other hand, without loss of generality, assuming that* $T_1 \geq T_2$*. If* $d_{\text{CKB}}(T_1,T_2) = 0$*, then*

$$\begin{aligned} d_{\text{CKB}}(T_1,T_2) = 0 &\xrightarrow{\text{Eq. (S.8)}} \left|\|\sqrt{T_1}\|_2 - \|\sqrt{T_2}\|_2\right| = 0 \\ &\Longrightarrow |\text{tr}(T_1 - T_2)| = 0 \\ &\Longrightarrow \|T_1 - T_2\|_1 = 0 \\ &\Longrightarrow T_1 = T_2, \end{aligned}$$

*where* $\text{tr}(T_1 - T_2) = \|T_1 - T_2\|_1 = 0$ *as the sum of self-adjoint operators is still self-adjoint. Thus we have* $d_{\text{CKB}}(T_1,T_2) = 0$ *if and only if* $T_1 = T_2$*.*
*(iii) The following equations prove the symmetry property of the CKB metric.*

$$\begin{aligned} d^2_{\text{CKB}}(T_1,T_2) &= \|T_1\|_1 + \|T_2\|_1 - 2\|\sqrt{T_2}\sqrt{T_1}\|_1 \\ &= \|T_1\|_1 + \|T_2\|_1 - 2\|\,(\sqrt{T_2}\sqrt{T_1})^*\|_1 \qquad \text{(S.9)} \\ &= \|T_2\|_1 + \|T_1\|_1 - 2\|\sqrt{T_1}\sqrt{T_2}\|_1 \\ &= d^2_{\text{CKB}}(T_2,T_1), \end{aligned}$$

*where Eq.(S.9) holds from Lemma S.1 (d). Since* $d_{\text{CKB}} \geq 0$*,* $d_{\text{CKB}}(T_1,T_2) = d_{\text{CKB}}(T_2,T_1)$*.*
*(iv) The following equations prove the triangle inequality of the CKB metric.*

$$\begin{aligned} d_{\text{CKB}}(T_1,T_2) + d_{\text{CKB}}(T_2,T_3) &\geq \left|\|\sqrt{T_1}\|_2 - \|\sqrt{T_2}\|_2\right| + \left|\|\sqrt{T_2}\|_2 - \|\sqrt{T_3}\|_2\right| \\ &\geq \left|\|\sqrt{T_1}\|_2 - \|\sqrt{T_2}\|_2 + \|\sqrt{T_2}\|_2 - \|\sqrt{T_3}\|_2\right| \\ &= \left|\|\sqrt{T_1}\|_2 - \|\sqrt{T_3}\|_2\right| \\ &= d_{\text{CKB}}(T_1,T_3). \end{aligned}$$

*The proof is completed by combing (i)-(iv).*

### S.4.4. Proof of Proposition 2

**Proposition 2** *If* $k_{\mathcal{Y}}$ *is positive definite kernel, then* $\mathbf{B}_s$ *and* $\mathbf{B}_t$ *are positive definite for any* $\varepsilon > 0$*.*

**Proof** *To simplify the notation, we consider the following matrix:*

$$\mathbf{B} = \mathbf{I}_n - \frac{1}{n\varepsilon}\left[\mathbf{G}_Y - \mathbf{G}_Y\left(\mathbf{G}_Y + \varepsilon n\mathbf{I}_n\right)^{-1}\mathbf{G}_Y\right].$$

*If $k_{\mathcal{Y}}$ is positive-definite kernel, then the kernel matrix $\mathbf{K}_{YY}$ is positive definite. Recall that $\mathbf{G}_Y = \mathbf{H}_n\mathbf{K}_{YY}\mathbf{H}_n$, so $\mathbf{G}_Y$ is a real positive semi-definite matrix whose Eigenvalue Decomposition (EVD) $\mathbf{G}_Y = \mathbf{U}\mathbf{D}\mathbf{U}^T$ always exists. Note that $\mathbf{U}$ is orthogonal and $\mathbf{D}$ is a diagonal matrix with non-negative entries. Then we can rewrite $\mathbf{B}$ as*

$$\begin{aligned}
\mathbf{B} =& \mathbf{I}_n - \frac{1}{n\varepsilon}\left[\mathbf{U}\mathbf{D}\mathbf{U}^T - \mathbf{U}\mathbf{D}\mathbf{U}^T\left(\mathbf{U}\mathbf{D}\mathbf{U}^T + \varepsilon n\mathbf{I}_n\right)^{-1}\mathbf{U}\mathbf{D}\mathbf{U}^T\right] \\
=& \mathbf{I}_n - \frac{1}{n\varepsilon}\left[\mathbf{U}\mathbf{D}\mathbf{U}^T - \mathbf{U}\mathbf{D}\mathbf{U}^T\left(\mathbf{U}(\mathbf{D} + \varepsilon n\mathbf{I}_n)\mathbf{U}^T\right)^{-1}\mathbf{U}\mathbf{D}\mathbf{U}^T\right] \\
=& \mathbf{I}_n - \frac{1}{n\varepsilon}\left[\mathbf{U}\mathbf{D}\mathbf{U}^T - \mathbf{U}\mathbf{D}\left(\mathbf{D} + \varepsilon n\mathbf{I}_n\right)^{-1}\mathbf{D}\mathbf{U}^T\right] \\
=& \mathbf{I}_n - \frac{1}{n\varepsilon}\mathbf{U}\left[\mathbf{D} - \mathbf{D}\left(\mathbf{D} + \varepsilon n\mathbf{I}_n\right)^{-1}\mathbf{D}\right]\mathbf{U}^T \\
=& \mathbf{U}\left[\mathbf{I}_n - \frac{1}{n\varepsilon}\left(\mathbf{D} - \mathbf{D}\left(\mathbf{D} + \varepsilon n\mathbf{I}_n\right)^{-1}\mathbf{D}\right)\right]\mathbf{U}^T \\
=& \mathbf{U}\mathbf{D}'\mathbf{U}^T
\end{aligned}$$

*where $\mathbf{D}' = \mathbf{I}_n - \frac{1}{n\varepsilon}\left(\mathbf{D} - \mathbf{D}\left(\mathbf{D} + \varepsilon n\mathbf{I}_n\right)^{-1}\mathbf{D}\right)$. It is clear that $\mathbf{D}'$ is a diagonal matrix and $\mathbf{U}\mathbf{D}'\mathbf{U}^T$ is the EVD of $\mathbf{B}$. Let $d_i'$ and $d_i \geq 0$ be the $i$-th diagonal entries of $\mathbf{D}'$ and $\mathbf{D}$, respectively. From the definition of $\mathbf{D}'$, we have*

$$d_i' = 1 - \frac{1}{n\varepsilon}\left(d_i - \frac{d_i^2}{d_i + n\varepsilon}\right) = 1 - \frac{d_i}{d_i + n\varepsilon} = \frac{n\varepsilon}{d_i + n\varepsilon} > 0,$$

*which means $\mathbf{B}$ has entirely positive eigenvalues. Especially, we have*

$$\mathbf{B} = \varepsilon n\mathbf{U}(\mathbf{D} + \varepsilon n\mathbf{I}_n)^{-1}\mathbf{U}^T = \varepsilon n\left(\mathbf{G}_Y + \varepsilon n\mathbf{I}_n\right)^{-1}.$$

*Thus, we have prove that $\mathbf{B}$ is positive definite.*

### S.4.5. Proof of Theorem 2

**Theorem 2** *The empirical estimation of the CKB metric is computed as*

$$\begin{aligned}
\hat{d}^2_{\text{CKB}}(\hat{\mathbf{R}}^s_{XX|Y}, \hat{\mathbf{R}}^t_{XX|Y}) = \varepsilon\text{tr}\left[\mathbf{G}^s_X\left(\varepsilon n\mathbf{I}_n + \mathbf{G}^s_Y\right)^{-1}\right] + \varepsilon\text{tr}\left[\mathbf{G}^t_X\left(\varepsilon m\mathbf{I}_m + \mathbf{G}^t_Y\right)^{-1}\right] \\
- \frac{2}{\sqrt{nm}}\left\|\left(\mathbf{H}_m\mathbf{C}_t\right)^T\mathbf{K}^{ts}_{XX}\left(\mathbf{H}_n\mathbf{C}_s\right)\right\|_*,
\end{aligned}$$

*where $\mathbf{C}_{s/t}$ satisfies the decomposition $\mathbf{B}_{s/t} = \mathbf{C}_{s/t}\mathbf{C}^T_{s/t}$ and $\|\cdot\|_*$ is the nuclear norm.*

To prove Theorem 2, we use the following lemma to reformulate the inverse of the implicit feature map matrix.

**Lemma S.5 (Sherman-Morrison-Woodbury [14])** *Suppose $\mathbf{A} \in \mathbb{R}^{n\times n}$ is nonsingular and $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{n\times k}$. If $(\mathbf{I}_k + \mathbf{V}^T\mathbf{A}^{-1}\mathbf{U})$ is nonsingular, then*

$$\left(\mathbf{A} + \mathbf{U}\mathbf{V}^T\right)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{I}_k + \mathbf{V}^T\mathbf{A}^{-1}\mathbf{U}\right)^{-1}\mathbf{V}^T\mathbf{A}^{-1}.$$

Now we begin the proof of Theorem 2.

**Proof** *The trace of conditional covariance operators $R^{s/t}_{XX|Y}$ have been investigated in [15] and can be written as*

$$\mathrm{tr}\left(\hat{\mathbf{R}}^{s}_{XX|Y}\right)=\mathrm{tr}\left[\mathbf{G}^{s}_{X}\left(\varepsilon n\mathbf{I}_n+\mathbf{G}^{s}_{Y}\right)^{-1}\right],\mathrm{tr}\left(\hat{\mathbf{R}}^{t}_{XX|Y}\right)=\mathrm{tr}\left[\mathbf{G}^{t}_{X}\left(\varepsilon n\mathbf{I}_m+\mathbf{G}^{t}_{Y}\right)^{-1}\right]. \tag{S.10}$$

*In terms of the third term in conditional kernel Bures metric, we first reformulate $R^{s}_{XX|Y}$ ($R^{t}_{XX|Y}$ is the same).*

$$\begin{aligned}\hat{\mathbf{R}}^{s}_{XX|Y}&=\frac{1}{n}\mathbf{\Phi}_s\mathbf{H}_n^2\mathbf{\Phi}_s^T-\frac{1}{n^2}\mathbf{\Phi}_s\mathbf{H}_n^2\mathbf{\Psi}_s^T\left(\hat{\mathbf{R}}^{s}_{YY}+\varepsilon\mathbf{I}\right)^{-1}\mathbf{\Psi}_s\mathbf{H}_n^2\mathbf{\Phi}_s^T\\&=\frac{1}{n}\mathbf{\Phi}_s\mathbf{H}_n\left[\mathbf{I}_n-\frac{1}{n}\mathbf{H}_n\mathbf{\Psi}_s^T\left(\hat{\mathbf{R}}^{s}_{YY}+\varepsilon\mathbf{I}\right)^{-1}\mathbf{\Psi}_s\mathbf{H}_n\right]\mathbf{H}_n\mathbf{\Phi}_s^T.\end{aligned} \tag{S.11}$$

*For $\left(\hat{\mathbf{R}}_{YY}+\varepsilon\mathbf{I}\right)^{-1}$, we apply the Sherman-Morrison-Woodbury formula as*

$$\begin{aligned}\left(\hat{\mathbf{R}}^{s}_{YY}+\varepsilon\mathbf{I}\right)^{-1}&=n\left(\varepsilon n\mathbf{I}+\mathbf{\Psi}_s\mathbf{H}_n\mathbf{H}_n\mathbf{\Psi}_s^T\right)^{-1}\\&=n\left[\frac{1}{\varepsilon n}\mathbf{I}-\frac{1}{\varepsilon^2n^2}\mathbf{\Psi}_s\mathbf{H}_n\left(\mathbf{I}_n+\frac{1}{\varepsilon n}\mathbf{H}_n\mathbf{\Psi}_s^T\mathbf{\Psi}_s\mathbf{H}_n\right)^{-1}\mathbf{H}_n\mathbf{\Psi}_s^T\right]\\&=\frac{1}{\varepsilon}\left[\mathbf{I}-\mathbf{\Psi}_s\mathbf{H}_n\left(\varepsilon n\mathbf{I}_n+\mathbf{G}^{s}_{Y}\right)^{-1}\mathbf{H}_n\mathbf{\Psi}_s^T\right].\end{aligned}$$

*By substituting this inverse to Eq. (S.11), then we have*

$$\begin{aligned}\hat{\mathbf{R}}^{s}_{XX|Y}&=\frac{1}{n}\mathbf{\Phi}_s\mathbf{H}_n\left[\mathbf{I}_n-\frac{1}{n\varepsilon}\left(\mathbf{G}^{s}_{Y}-\mathbf{G}^{s}_{Y}\left(\mathbf{G}^{s}_{Y}+\varepsilon n\mathbf{I}_n\right)^{-1}\mathbf{G}^{s}_{Y}\right)\right]\mathbf{H}_n\mathbf{\Phi}_s^T\\&=\frac{1}{n}\mathbf{\Phi}_s\mathbf{H}_n\mathbf{B}_s\mathbf{H}_n\mathbf{\Phi}_s^T.\end{aligned}$$

*Similarly, one can show that $\hat{\mathbf{R}}^{t}_{XX|Y}=\frac{1}{m}\mathbf{\Phi}_t\mathbf{H}_m\mathbf{B}_t\mathbf{H}_m\mathbf{\Phi}_t^T$.*

*Based on above reformulations, we consider the third term $\mathrm{tr}(\hat{\mathbf{R}}^{st}_{XX|Y})$. Let $\mathbf{C}_{s/t}$ be any matrix satisfy the decomposition $\mathbf{B}_{s/t}=\mathbf{C}_{s/t}\mathbf{C}_{s/t}^T$. Denote $\hat{\mathbf{H}}_n\triangleq\mathbf{H}_n\mathbf{C}_s$ and $\hat{\mathbf{H}}_m\triangleq\mathbf{H}_m\mathbf{C}_t$, then*

$$\begin{aligned}\mathrm{tr}\left(\hat{\mathbf{R}}^{st}_{XX|Y}\right)&=\mathrm{tr}\left(\sqrt{\sqrt{\frac{1}{n}\mathbf{\Phi}_s\mathbf{H}_n\mathbf{B}_s\mathbf{H}_n\mathbf{\Phi}_s^T}\left(\frac{1}{m}\mathbf{\Phi}_t\mathbf{H}_m\mathbf{B}_t\mathbf{H}_m\mathbf{\Phi}_t^T\right)\sqrt{\frac{1}{n}\mathbf{\Phi}_s\mathbf{H}_n\mathbf{B}_s\mathbf{H}_n\mathbf{\Phi}_s^T}}\right)\\&=\frac{1}{\sqrt{nm}}\sqrt{\sqrt{\mathbf{\Phi}_s\hat{\mathbf{H}}_n\hat{\mathbf{H}}_n^T\mathbf{\Phi}_s^T}\left(\mathbf{\Phi}_t\hat{\mathbf{H}}_m\hat{\mathbf{H}}_m^T\mathbf{\Phi}_t^T\right)\sqrt{\mathbf{\Phi}_s\hat{\mathbf{H}}_n\hat{\mathbf{H}}_n^T\mathbf{\Phi}_s^T}}\\&=\frac{1}{\sqrt{nm}}\sqrt{\left(\sqrt{\mathbf{\Phi}_s\hat{\mathbf{H}}_n\hat{\mathbf{H}}_n^T\mathbf{\Phi}_s^T}\mathbf{\Phi}_t\hat{\mathbf{H}}_m\right)\left(\sqrt{\mathbf{\Phi}_s\hat{\mathbf{H}}_n\hat{\mathbf{H}}_n^T\mathbf{\Phi}_s^T}\mathbf{\Phi}_t\hat{\mathbf{H}}_m\right)^T}\\&=\frac{1}{\sqrt{nm}}\sqrt{\left(\sqrt{\mathbf{\Phi}_s\hat{\mathbf{H}}_n\hat{\mathbf{H}}_n^T\mathbf{\Phi}_s^T}\mathbf{\Phi}_t\hat{\mathbf{H}}_m\right)^T\left(\sqrt{\mathbf{\Phi}_s\hat{\mathbf{H}}_n\hat{\mathbf{H}}_n^T\mathbf{\Phi}_s^T}\mathbf{\Phi}_t\hat{\mathbf{H}}_m\right)}\\&=\frac{1}{\sqrt{nm}}\sqrt{\hat{\mathbf{H}}_m^T\mathbf{\Phi}_t^T\left(\mathbf{\Phi}_s\hat{\mathbf{H}}_n\hat{\mathbf{H}}_n^T\mathbf{\Phi}_s^T\right)\mathbf{\Phi}_t\hat{\mathbf{H}}_m}\\&=\frac{1}{\sqrt{nm}}\sqrt{\left(\hat{\mathbf{H}}_m^T\mathbf{K}^{ts}_{XX}\hat{\mathbf{H}}_n\right)\left(\hat{\mathbf{H}}_m^T\mathbf{K}^{ts}_{XX}\hat{\mathbf{H}}_n\right)^T}\\&=\frac{1}{\sqrt{nm}}\left\|\left(\mathbf{H}_m\mathbf{C}_t\right)^T\mathbf{K}^{ts}_{XX}\left(\mathbf{H}_n\mathbf{C}_s\right)\right\|_*.\end{aligned} \tag{S.12}$$

*Combining Eq. (S.10) and (S.12), the empirical estimation of conditional kernel Bures metric is proved.*

### S.4.6. Proof of Theorem 3

Before the proof, we first review some basic definitions used in asymptotic statistics [16]. Let $d(\cdot,\cdot)$ be a distance function, and $Z_n$ a sequence of random variable. $Z_n$ is said to *converge to $Z$ in probability* (which is denoted by $Z_n \xrightarrow{\mathrm{P}} Z$) if $\forall \varepsilon > 0$,

$$\mathrm{P}\left(d(Z_n, Z) > \varepsilon\right) \to 0 \quad (n \to \infty).$$

This convergence is equivalent to $d(Z_n, Z) \to 0$. Now we introduce the stochastic $o$ symbols. The notation $o_P(1)$ means that a sequence of random vectors converges to 0 in probability, i.e.,

$$Z_n = o_P(1) \quad \Longleftrightarrow \quad Z_n \xrightarrow{\mathrm{P}} 0.$$

With the notation $o_P(\cdot)$, the convergence can be generalized as

$$Z_n = o_P(S_n) \quad \Longleftrightarrow \quad Z_n = Y_n S_n \text{ and } Y_n \xrightarrow{\mathrm{P}} 0,$$

where $S_n$ is the so-called *rate*.

**Lemma S.6 ([16], Theorem 2.3, Continuous mapping)** *Let $g$ be continuous at every point of a set $C$ such that $\mathrm{P}(Z \in C) = 1$. If $Z_n \xrightarrow{\mathrm{P}} Z$, then $g(Z_n) \xrightarrow{\mathrm{P}} g(Z)$.*

**Lemma S.7 ([15], Proposition 7)** *Let the regularization parameter $\varepsilon$ in $\mathbf{R}_{XX|Y}$ be a series related to $n$, i.e., $\varepsilon_n$. Assuming $\varepsilon_n$ satisfies that $\varepsilon_n \to 0$ and $\varepsilon_n\sqrt{n} \to \infty$ $(n \to \infty)$, then we have*

$$|\mathrm{tr}(\hat{\mathbf{R}}^{(n)}_{XX|Y}) - \mathrm{tr}(\mathbf{R}_{XX|Y})| = o_P\left(\frac{1}{\varepsilon_n\sqrt{n}}\right)$$

*as $n \to \infty$.*

Lemma S.6 shows that the continuous mapping is guaranteed to preserve the convergence of a sequence. Lemma S.7 show the convergence of empirical conditional covariance estimation. We will use these lemmas to prove the convergence the square root of the conditional covariance operator.

**Theorem 3** *Let the regularization parameter $\varepsilon$ in $\mathbf{R}_{XX|Y}$ be a series related to $n'$, i.e., $\varepsilon_{n'}$. Assuming $\varepsilon_{n'}$ satisfies that $\varepsilon_{n'} \to 0$ and $\varepsilon_{n'}\sqrt{n'} \to \infty$ $(n' \to \infty)$, then we have*

$$|\hat{D}^{(n')}_{\mathrm{CKB}} - D_{\mathrm{CKB}}| \to 0 \quad (n' \to \infty)$$

*in probability with rate $(\frac{1}{\varepsilon'_n\sqrt{n'}})^{\frac{1}{2}}$.*

**Proof** *To simplify the notation, let $\mathbf{R}^{s^{(n)}}$ and $\mathbf{R}^{t^{(m)}}$ be the conditional covariance operator drawn i.i.d. from distribution $P^s_{XY}$ and $P^t_{XY}$ with sample size $n$ and $m$, respectively. Recall that $n' = \min\{n, m\}$, we first split $|\hat{D}^{(n')}_{\mathrm{CKB}} - D_{\mathrm{CKB}}|$ into three terms, i.e.,*

$$|\hat{D}^{(n')}_{\mathrm{CKB}} - D_{\mathrm{CKB}}| \le |\mathrm{tr}(\mathbf{R}^{s^{(n)}}) - \mathrm{tr}(\mathbf{R}^s)| + |\mathrm{tr}(\mathbf{R}^{t^{(m)}}) - \mathrm{tr}(\mathbf{R}^t)| + 2|\mathrm{tr}(\mathbf{R}^{st^{(n,m)}}) - \mathrm{tr}(\mathbf{R}^{st})|. \tag{S.13}$$

*According to Lemma S.7, the first two terms in Eq. (S.13) are guaranteed to converge to 0 as $n' \to \infty$ with rate $o_P\left(\frac{1}{\varepsilon_{n'}\sqrt{n'}}\right)$. Now we focus on the third term $|\mathrm{tr}(\mathbf{R}^{st^{(n,m)}}) - \mathrm{tr}(\mathbf{R}^{st})|$, which can be reformulated as*

$$
\begin{aligned}
& |\mathrm{tr}(\mathbf{R}^{st^{(n,m)}}) - \mathrm{tr}(\mathbf{R}^{st})| \\
= & \left| \left\| \sqrt{\mathbf{R}^{t^{(m)}}} \sqrt{\mathbf{R}^{s^{(n)}}} \right\|_1 - \left\| \sqrt{\mathbf{R}^{t}} \sqrt{\mathbf{R}^{s}} \right\|_1 \right| \\
\leq & \left\| \sqrt{\mathbf{R}^{t^{(m)}}} \sqrt{\mathbf{R}^{s^{(n)}}} - \sqrt{\mathbf{R}^{t}} \sqrt{\mathbf{R}^{s}} \right\|_1 \\
= & \left\| \sqrt{\mathbf{R}^{t^{(m)}}} \sqrt{\mathbf{R}^{s^{(n)}}} - \sqrt{\mathbf{R}^{t}} \sqrt{\mathbf{R}^{s^{(n)}}} + \sqrt{\mathbf{R}^{t}} \sqrt{\mathbf{R}^{s^{(n)}}} - \sqrt{\mathbf{R}^{t}} \sqrt{\mathbf{R}^{s}} \right\|_1 \\
\leq & \left\| \sqrt{\mathbf{R}^{t^{(m)}}} \sqrt{\mathbf{R}^{s^{(n)}}} - \sqrt{\mathbf{R}^{t}} \sqrt{\mathbf{R}^{s^{(n)}}} \right\|_1 + \left\| \sqrt{\mathbf{R}^{t}} \sqrt{\mathbf{R}^{s^{(n)}}} - \sqrt{\mathbf{R}^{t}} \sqrt{\mathbf{R}^{s}} \right\|_1 \\
\leq & \left\| \sqrt{\mathbf{R}^{t^{(m)}}} - \sqrt{\mathbf{R}^{t}} \right\|_2 \left\| \sqrt{\mathbf{R}^{s^{(n)}}} \right\|_2 + \left\| \sqrt{\mathbf{R}^{t}} \right\|_2 \left\| \sqrt{\mathbf{R}^{s^{(n)}}} - \sqrt{\mathbf{R}^{s}} \right\|_2 \\
\leq & \sqrt{\left\| \mathbf{R}^{t^{(m)}} - \mathbf{R}^{t} \right\|_1} \left\| \sqrt{\mathbf{R}^{s^{(n)}}} \right\|_2 + \left\| \sqrt{\mathbf{R}^{t}} \right\|_2 \sqrt{\left\| \mathbf{R}^{s^{(n)}} - \mathbf{R}^{s} \right\|_1}, \qquad \text{(S.14)}
\end{aligned}
$$

*where Eq. (S.14) holds from Lemma S.3. Since the conditional covariance operators are trace class,* $\left\| \sqrt{\mathbf{R}^{s^{(n)}}} \right\|_2$ *and* $\left\| \sqrt{\mathbf{R}^{t}} \right\|_2$ *are finite from Corollary S.1 (b). Therefore, the estimation error is bounded by* $\sqrt{\left\| \mathbf{R}^{t^{(m)}} - \mathbf{R}^{t} \right\|_1}$ *and* $\sqrt{\left\| \mathbf{R}^{s^{(n)}} - \mathbf{R}^{s} \right\|_1}$. *For* $\sqrt{\left\| \mathbf{R}^{t^{(m)}} - \mathbf{R}^{t} \right\|_1}$, *since* $\mathbf{R}^{t^{(m)}} - \mathbf{R}^{t}$ *is self-adjoint, we have*

$$
d^{t^{(m)}} \triangleq \left\| \mathbf{R}^{t^{(m)}} - \mathbf{R}^{t} \right\|_1 = |\mathrm{tr}(\mathbf{R}^{t^{(m)}} - \mathbf{R}^{t})| = |\mathrm{tr}(\mathbf{R}^{t^{(m)}}) - \mathrm{tr}(\mathbf{R}^{t})|.
$$

*Lemma S.7 shows that* $d^{t^{(m)}} = o_P\left(\frac{1}{\varepsilon_m \sqrt{m}}\right)$ *as* $n \to \infty$, *which means* $\varepsilon_m \sqrt{m} d^{t^{(m)}} \xrightarrow{\mathrm{P}} 0$. *Denote the estimation error sequence* $Z_m = \varepsilon_m \sqrt{m} d^{t^{(m)}}$, $Z = 0$ *and the set* $C = [0, \infty)$, *then* $\mathrm{P}(Z \in C) = 1$ *and* $g(\cdot) = \sqrt{\cdot}$ *is continuous on* $C$ *since* $Z_n, Z \in C$. *Then from Lemma S.6, we have*

$$
g(Z_n) = \sqrt{Z_n} = \left( \sqrt{\varepsilon_m \sqrt{m}} \right) \sqrt{d^{t^{(m)}}} \xrightarrow{\mathrm{P}} 0,
$$

*which means* $\sqrt{d^{t^{(m)}}} = \sqrt{\left\| \mathbf{R}^{t^{(m)}} - \mathbf{R}^{t} \right\|_1} = o_P\left(\frac{1}{\sqrt{\varepsilon_m \sqrt{m}}}\right)$ *as* $m \to \infty$. *Similarly, we have* $\sqrt{\left\| \mathbf{R}^{s^{(n)}} - \mathbf{R}^{s} \right\|_1} = o_P\left(\frac{1}{\sqrt{\varepsilon_n \sqrt{n}}}\right)$ *as* $n \to \infty$. *Finally, the estimation error of the CKB metric is*

$$
|\hat{D}_{\mathrm{CKB}}^{(n')} - D_{\mathrm{CKB}}| = o_P\left(\frac{1}{\varepsilon_{n'} \sqrt{n'}}\right) + o_P\left(\frac{1}{\sqrt{\varepsilon_m \sqrt{m}}}\right) + o_P\left(\frac{1}{\sqrt{\varepsilon_n \sqrt{n}}}\right) = o_P\left(\frac{1}{\sqrt{\varepsilon_n' \sqrt{n'}}}\right)
$$

*as* $n' \to \infty$.

## References


[1] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *NeurIPS*, pp. 8026–8037, 2019. 2

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, pp. 770–778, 2016. 2

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, pp. 1097–1105, 2012. 2

[4] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *ICML*, pp. 1989–1998, 2018. 2

[5] T. Ringwald and R. Stiefelhagen, "Adaptiope: A modern benchmark for unsupervised domain adaptation," in *WACV*, pp. 101–110, 2021. 2

[6] X. Gu, J. Sun, and Z. Xu, "Spherical space domain adaptation with robust pseudo-label loss," in *CVPR*, pp. 9101–9110, 2020. 2

[7] Y. Zhang, B. Deng, H. Tang, L. Zhang, and K. Jia, "Unsupervised multi-class domain adaptation: Theory, algorithms, and practice," *IEEE TPAMI*, 2020. 2

[8] G. Kang, L. Jiang, Y. Wei, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for single-and multi-source domain adaptation," *IEEE TPAMI*, 2020. 2

[9] M. Reed and B. Simon, *Methods of Modern Mathematical Physics: Functional analysis*. Academic Press, 1981. 2, 3

[10] R. T. Powers and E. Størmer, "Free states of the canonical anticommutation relations," *Communications in Mathematical Physics*, vol. 16, no. 1, pp. 1–33, 1970. 3

[11] Z. Zhang, M. Wang, and A. Nehorai, "Optimal transport in reproducing kernel hilbert spaces: Theory and applications," *IEEE TPAMI*, 2019. 3, 4

[12] I. Klebanov, I. Schuster, and T. Sullivan, "A rigorous theory of conditional mean embeddings," *SIAM Journal on Mathematics of Data Science*, vol. 2, no. 3, pp. 583–606, 2020. 4

[13] L. Song, J. Huang, A. Smola, and K. Fukumizu, "Hilbert space embeddings of conditional distributions with applications to dynamical systems," in *ICML*, pp. 961–968, 2009. 4

[14] M. A. Woodbury, "Woodbury, inverting modified matrices, memorandum rept. 42," *Statistical Research Group, Princeton University, Princeton, NJ*, 1950. 6

[15] K. Fukumizu, F. R. Bach, M. I. Jordan, *et al.*, "Kernel dimension reduction in regression," *The Annals of Statistics*, vol. 37, no. 4, pp. 1871–1905, 2009. 7, 8

[16] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge university press, 2000. 8