
Rethinking Joint Maximum Mean Discrepancy for Domain Adaptation

Wei Wang¹ Haifeng Xia¹ Chao Huang^{1*} Zhengming Ding²

Cong Wang³ Haojie Li⁴ Xiaochun Cao¹

¹Shenzhen Campus of Sun Yat-sen University ²Department of Computer Science, Tulane University

³University of California, San Francisco ⁴Shandong University of Science and Technology

{wangwei29, xiahf5, huangch253, caoxiaochun}@mail.sysu.edu.cn zding1@tulane.edu

supercong94@gmail.com hjli@sdust.edu.cn

Contents

1	A Concise JMMD in an RKHS	1
2	A Concise JMMD in a Projected RKHS	3
3	Probability Distribution Distances	6
3.1	Maximum Mean Discrepancy	6
3.2	Class-Wise Maximum Mean Discrepancy	6
3.3	Weighted Class-Wise Maximum Mean Discrepancy	7
4	The Uniformity of JMMD	7
5	Reproducing Kernels	8
6	Experiments	9

1 A Concise JMMD in an RKHS

Theorem 1 *In an RKHS, JMMD could be rewritten as the following concise form,*

$$\begin{aligned}\mathbb{D}_{\mathcal{H}}\left(\mathcal{P}^s(\mathbf{X}^s, \mathbf{Y}^s), \mathcal{P}^t(\mathbf{X}^t, \mathbf{Y}^t)\right) &= \left\| \frac{1}{n^s} \sum_{i=1}^{n^s} \left(\psi(\mathbf{x}_i^s) \otimes \phi(\mathbf{y}_i^s) \right) - \frac{1}{n^t} \sum_{j=1}^{n^t} \left(\psi(\mathbf{x}_j^t) \otimes \phi(\mathbf{y}_j^t) \right) \right\|_{\mathcal{H}}^2 \\ &= \text{tr}\left(\mathbf{K}^{\text{XX}}(\mathbf{K}^{\text{YY}} \odot \mathbf{M}^{\text{J}})\right),\end{aligned}\tag{1}$$

where $\mathbb{D}_{\mathcal{H}}$ denotes a distance metric between two joint probability distributions, *i.e.*, \mathcal{P}^s and \mathcal{P}^t . We empirically estimate JMMD with the following steps: i) we utilize ψ and ϕ to map features and labels from the source domain and target domain to the RKHS, respectively; ii) we calculate the mean of

*Corresponding author.

the tensor product between feature and label for each domain; iii) we compute the difference between these two means. Notably, \mathbf{x}_i and \mathbf{x}_j are the i -th and j -th column vectors of \mathbf{X} .

Proof:

For convenience, we define $\Gamma^s(\mathbf{x}^s, \mathbf{y}^s)$ and $\Gamma^t(\mathbf{x}^t, \mathbf{y}^t)$ as shown in the following equations,

$$\begin{aligned}\Gamma^s(\mathbf{x}^s, \mathbf{y}^s) &= [\psi(\mathbf{x}_1^s) \otimes \phi(\mathbf{y}_1^s), \dots, \psi(\mathbf{x}_{n^s}^s) \otimes \phi(\mathbf{y}_{n^s}^s)] \in \mathbb{R}^{\infty \times n^s}, \\ \Gamma^t(\mathbf{x}^t, \mathbf{y}^t) &= [\psi(\mathbf{x}_1^t) \otimes \phi(\mathbf{y}_1^t), \dots, \psi(\mathbf{x}_{n^t}^t) \otimes \phi(\mathbf{y}_{n^t}^t)] \in \mathbb{R}^{\infty \times n^t}.\end{aligned}\quad (2)$$

Then, (1) could be rewritten as below,

$$\begin{aligned}\mathbb{D}_{\mathcal{H}}(\mathcal{P}^s(\mathbf{X}^s, \mathbf{Y}^s), \mathcal{P}^t(\mathbf{X}^t, \mathbf{Y}^t)) &= \left\| \frac{1}{n^s} \sum_{i=1}^{n^s} (\psi(\mathbf{x}_i^s) \otimes \phi(\mathbf{y}_i^s)) - \frac{1}{n^t} \sum_{j=1}^{n^t} (\psi(\mathbf{x}_j^t) \otimes \phi(\mathbf{y}_j^t)) \right\|_{\mathcal{H}}^2 \\ &= \text{tr} \left(\begin{bmatrix} \Gamma^s(\mathbf{x}^s, \mathbf{y}^s)^\top \Gamma^s(\mathbf{x}^s, \mathbf{y}^s) & \Gamma^s(\mathbf{x}^s, \mathbf{y}^s)^\top \Gamma^t(\mathbf{x}^t, \mathbf{y}^t) \\ \Gamma^t(\mathbf{x}^t, \mathbf{y}^t)^\top \Gamma^s(\mathbf{x}^s, \mathbf{y}^s) & \Gamma^t(\mathbf{x}^t, \mathbf{y}^t)^\top \Gamma^t(\mathbf{x}^t, \mathbf{y}^t) \end{bmatrix} \begin{bmatrix} \frac{\mathbf{1}_{n^s \times 1} \mathbf{1}_{n^s \times 1}^\top}{n^s n^s} & \frac{-\mathbf{1}_{n^s \times 1} \mathbf{1}_{n^t \times 1}^\top}{n^s n^t} \\ \frac{-\mathbf{1}_{n^t \times 1} \mathbf{1}_{n^s \times 1}^\top}{n^t n^s} & \frac{\mathbf{1}_{n^t \times 1} \mathbf{1}_{n^t \times 1}^\top}{n^t n^t} \end{bmatrix} \right).\end{aligned}\quad (3)$$

where $\mathbf{1}_{n^s \times 1}$ and $\mathbf{1}_{n^t \times 1}$ are two column vectors whose elements are all ones with sizes of n^s and n^t .

Moreover, we have the following equations,

$$\Gamma^s(\mathbf{x}^s, \mathbf{y}^s)^\top \Gamma^s(\mathbf{x}^s, \mathbf{y}^s) = \mathbf{K}^{X^s X^s} \odot \mathbf{K}^{Y^s Y^s}, \quad (4)$$

$$\Gamma^t(\mathbf{x}^t, \mathbf{y}^t)^\top \Gamma^t(\mathbf{x}^t, \mathbf{y}^t) = \mathbf{K}^{X^t X^t} \odot \mathbf{K}^{Y^t Y^t}, \quad (5)$$

$$\Gamma^s(\mathbf{x}^s, \mathbf{y}^s)^\top \Gamma^t(\mathbf{x}^t, \mathbf{y}^t) = \mathbf{K}^{X^s X^t} \odot \mathbf{K}^{Y^s Y^t}, \quad (6)$$

$$\Gamma^t(\mathbf{x}^t, \mathbf{y}^t)^\top \Gamma^s(\mathbf{x}^s, \mathbf{y}^s) = \mathbf{K}^{X^t X^s} \odot \mathbf{K}^{Y^t Y^s}. \quad (7)$$

where $\mathbf{K}^{X^s X^s}, \dots, \mathbf{K}^{Y^s Y^s}, \dots \in \mathbb{R}^{n^s \times n^s}, \dots \mathbb{R}^{n^s \times n^s}, \dots$ are the kernel matrices and they are computed by $k_{ij}^{X^s X^s} = k^X(\mathbf{x}_i^s, \mathbf{x}_j^{s\top}), \dots, k_{ij}^{Y^s Y^s} = k^Y(\mathbf{y}_i^s, \mathbf{y}_j^{s\top}), \dots$. Here k^X and k^Y are feature and label kernels.

Therefore, we can rewrite (3) using the feature kernel matrix \mathbf{K}^{XX} and the label kernel matrix \mathbf{K}^{YY} as below,

$$\begin{aligned}\mathbb{D}_{\mathcal{H}}(\mathcal{P}^s(\mathbf{X}^s, \mathbf{Y}^s), \mathcal{P}^t(\mathbf{X}^t, \mathbf{Y}^t)) &= \text{tr} \left(\begin{bmatrix} \mathbf{K}^{X^s X^s} & \mathbf{K}^{X^s X^t} \\ \mathbf{K}^{X^t X^s} & \mathbf{K}^{X^t X^t} \end{bmatrix} \odot \begin{bmatrix} \mathbf{K}^{Y^s Y^s} & \mathbf{K}^{Y^s Y^t} \\ \mathbf{K}^{Y^t Y^s} & \mathbf{K}^{Y^t Y^t} \end{bmatrix} \begin{bmatrix} \frac{\mathbf{1}_{n^s \times 1} \mathbf{1}_{n^s \times 1}^\top}{n^s n^s} & \frac{-\mathbf{1}_{n^s \times 1} \mathbf{1}_{n^t \times 1}^\top}{n^s n^t} \\ \frac{-\mathbf{1}_{n^t \times 1} \mathbf{1}_{n^s \times 1}^\top}{n^t n^s} & \frac{\mathbf{1}_{n^t \times 1} \mathbf{1}_{n^t \times 1}^\top}{n^t n^t} \end{bmatrix} \right) \\ &= \text{tr}(\mathbf{K}^{XX} \odot \mathbf{K}^{YY} \mathbf{M}^J),\end{aligned}\quad (8)$$

where $\mathbf{K}^{XX} \in \mathbb{R}^{n \times n}$ and $\mathbf{K}^{YY} \in \mathbb{R}^{n \times n}$ are feature and label kernel matrices for all source and target domains, and $n = n^s + n^t$.

According to $\text{tr}(\mathbf{A} \odot \mathbf{BC}) = \text{tr}(\mathbf{A}(\mathbf{B} \odot \mathbf{C}))$ where \mathbf{A}, \mathbf{B} and \mathbf{C} are symmetric matrices [1], thus $\text{tr}(\mathbf{K}^{XX} \odot \mathbf{K}^{YY} \mathbf{M}^J) = \text{tr}(\mathbf{K}^{XX}(\mathbf{K}^{YY} \odot \mathbf{M}^J))$. \mathbf{M}^J is calculated as below,

$$m_{ij}^J = \begin{cases} 1/(n^s n^s), & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}^s \\ 1/(n^t n^t), & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}^t \\ -1/(n^s n^t), & \text{otherwise.} \end{cases} \quad (9)$$

where \mathcal{D}^s and \mathcal{D}^t denote source and target domains.

□

2 A Concise JMMD in a Projected RKHS

Theorem 2 *In a projected RKHS, JMMD could be rewritten as the following concise form,*

$$\begin{aligned} & \mathbb{D}_{\mathcal{H}}(\mathcal{P}^s(\mathbf{X}^s, \mathbf{Y}^s), \mathcal{P}^t(\mathbf{X}^t, \mathbf{Y}^t)) \\ &= \left\| \frac{1}{n^s} \sum_{i=1}^{n^s} \left(\mathbf{T}^\top \psi(\mathbf{x}_i^s) \otimes \phi(\mathbf{y}_i^s) \right) - \frac{1}{n^t} \sum_{j=1}^{n^t} \left(\mathbf{T}^\top \psi(\mathbf{x}_j^t) \otimes \phi(\mathbf{y}_j^t) \right) \right\|_{\mathcal{H}}^2 \\ &= \left\| \frac{1}{n^s} \sum_{i=1}^{n^s} \left(\sum_{l=1}^n (\mathbf{b}_l \psi(\mathbf{x}_i^s)^\top \psi(\mathbf{x}_i^s)) \otimes \phi(\mathbf{y}_i^s) \right) - \frac{1}{n^t} \sum_{j=1}^{n^t} \left(\sum_{l=1}^n (\mathbf{b}_l \psi(\mathbf{x}_j^t)^\top \psi(\mathbf{x}_j^t)) \otimes \phi(\mathbf{y}_j^t) \right) \right\|_{\mathcal{H}}^2 \\ &= \text{tr}(\mathbf{B}^\top \mathbf{K}^{\text{XX}} (\mathbf{K}^{\text{YY}} \odot \mathbf{M}^J) \mathbf{K}^{\text{XX}} \mathbf{B}). \end{aligned} \quad (10)$$

where $\mathbf{T} \in \mathbb{R}^{\infty \times d}$ is the feature projection matrix and ‘d’ is the dimension in the embedded subspace. Different from (1), we project $\psi(\mathbf{x}_i^s)$ and $\psi(\mathbf{x}_j^t)$ into an embedded subspace, and then empirically estimate JMMD.

Proof:

We begin by introducing the Representer theorem [2] as below,

Theorem 3 (Representer theorem) *It says that any function can be decomposed into finite values of a kernel function with corresponding coefficients [2].*

$$\begin{aligned} \mathbf{T}^\top \psi(\mathbf{x}) &= \sum_{i=1}^n (\mathbf{b}_i k^X(\mathbf{x}, \mathbf{x}_i)) = \\ &= \sum_{i=1}^n (\mathbf{b}_i \langle \psi(\mathbf{x}), \psi(\mathbf{x}_i) \rangle) = \sum_{i=1}^n (\mathbf{b}_i \psi(\mathbf{x}_i)^\top \psi(\mathbf{x})), \end{aligned} \quad (11)$$

where $\mathbf{b}_i \in \mathbb{R}^{d \times 1}$ and we define a new projection matrix $\mathbf{B} = [\mathbf{b}_1^\top; \dots; \mathbf{b}_n^\top] \in \mathbb{R}^{n \times d}$.

For convenience, we define $\Theta^s(\mathbf{x}^s, \mathbf{y}^s)$ and $\Theta^t(\mathbf{x}^t, \mathbf{y}^t)$ as shown in the following equations according to the Representer theorem,

$$\begin{aligned} \Theta^s(\mathbf{x}^s, \mathbf{y}^s) &= \\ &= \left[\left(\sum_{l=1}^n \mathbf{b}_l \psi(\mathbf{x}_l)^\top \right) \psi(\mathbf{x}_1^s) \otimes \phi(\mathbf{y}_1^s), \dots, \left(\sum_{l=1}^n \mathbf{b}_l \psi(\mathbf{x}_l)^\top \right) \psi(\mathbf{x}_{n^s}^s) \otimes \phi(\mathbf{y}_{n^s}^s) \right] \in \mathbb{R}^{\infty \times n^s}, \\ \Theta^t(\mathbf{x}^t, \mathbf{y}^t) &= \\ &= \left[\left(\sum_{l=1}^n \mathbf{b}_l \psi(\mathbf{x}_l)^\top \right) \psi(\mathbf{x}_1^t) \otimes \phi(\mathbf{y}_1^t), \dots, \left(\sum_{l=1}^n \mathbf{b}_l \psi(\mathbf{x}_l)^\top \right) \psi(\mathbf{x}_{n^t}^t) \otimes \phi(\mathbf{y}_{n^t}^t) \right] \in \mathbb{R}^{\infty \times n^t}. \end{aligned} \quad (12)$$

where $\mathbf{b}_i \in \mathbb{R}^{d \times 1}$ and we define a new projection matrix $\mathbf{B} = [\mathbf{b}_1^\top; \dots; \mathbf{b}_n^\top] \in \mathbb{R}^{n \times d}$ ($n = n^s + n^t$).

Then, (10) could be rewritten as below,

$\mathbb{D}_{\mathcal{H}}$

$$\begin{aligned}
&= \left\| \frac{1}{n^s} \sum_{i=1}^{n^s} \left(\left(\sum_{l=1}^n \mathbf{b}_l \psi(\mathbf{x}_l)^\top \right) \psi(\mathbf{x}_i^s) \otimes \phi(\mathbf{y}_i^s) \right) - \frac{1}{n^t} \sum_{j=1}^{n^t} \left(\left(\sum_{l=1}^n \mathbf{b}_l \psi(\mathbf{x}_l)^\top \right) \psi(\mathbf{x}_j^t) \otimes \phi(\mathbf{y}_j^t) \right) \right\|_{\mathcal{H}}^2 \\
&= \text{tr} \left(\begin{bmatrix} \Theta^s(\mathbf{x}^s, \mathbf{y}^s) & \Theta^t(\mathbf{x}^t, \mathbf{y}^t) \end{bmatrix} \begin{bmatrix} \frac{\mathbf{1}_{n^s \times 1} \mathbf{1}_{n^s \times 1}^\top}{n^s n^s} & \frac{-\mathbf{1}_{n^s \times 1} \mathbf{1}_{n^t \times 1}^\top}{n^s n^t} \\ \frac{-\mathbf{1}_{n^t \times 1} \mathbf{1}_{n^s \times 1}^\top}{n^t n^s} & \frac{\mathbf{1}_{n^t \times 1} \mathbf{1}_{n^t \times 1}^\top}{n^t n^t} \end{bmatrix} \begin{bmatrix} \Theta^s(\mathbf{x}^s, \mathbf{y}^s)^\top \\ \Theta^t(\mathbf{x}^t, \mathbf{y}^t)^\top \end{bmatrix} \right) \\
&= \text{tr} \left(\begin{bmatrix} \Theta^s(\mathbf{x}^s, \mathbf{y}^s)^\top \\ \Theta^t(\mathbf{x}^t, \mathbf{y}^t)^\top \end{bmatrix} \begin{bmatrix} \Theta^s(\mathbf{x}^s, \mathbf{y}^s) & \Theta^t(\mathbf{x}^t, \mathbf{y}^t) \end{bmatrix} \begin{bmatrix} \frac{\mathbf{1}_{n^s \times 1} \mathbf{1}_{n^s \times 1}^\top}{n^s n^s} & \frac{-\mathbf{1}_{n^s \times 1} \mathbf{1}_{n^t \times 1}^\top}{n^s n^t} \\ \frac{-\mathbf{1}_{n^t \times 1} \mathbf{1}_{n^s \times 1}^\top}{n^t n^s} & \frac{\mathbf{1}_{n^t \times 1} \mathbf{1}_{n^t \times 1}^\top}{n^t n^t} \end{bmatrix} \right) \\
&= \text{tr} \left(\begin{bmatrix} \Theta^s(\mathbf{x}^s, \mathbf{y}^s)^\top \Theta^s(\mathbf{x}^s, \mathbf{y}^s) & \Theta^s(\mathbf{x}^s, \mathbf{y}^s)^\top \Theta^t(\mathbf{x}^t, \mathbf{y}^t) \\ \Theta^t(\mathbf{x}^t, \mathbf{y}^t)^\top \Theta^s(\mathbf{x}^s, \mathbf{y}^s) & \Theta^t(\mathbf{x}^t, \mathbf{y}^t)^\top \Theta^t(\mathbf{x}^t, \mathbf{y}^t) \end{bmatrix} \begin{bmatrix} \frac{\mathbf{1}_{n^s \times 1} \mathbf{1}_{n^s \times 1}^\top}{n^s n^s} & \frac{-\mathbf{1}_{n^s \times 1} \mathbf{1}_{n^t \times 1}^\top}{n^s n^t} \\ \frac{-\mathbf{1}_{n^t \times 1} \mathbf{1}_{n^s \times 1}^\top}{n^t n^s} & \frac{\mathbf{1}_{n^t \times 1} \mathbf{1}_{n^t \times 1}^\top}{n^t n^t} \end{bmatrix} \right). \tag{13}
\end{aligned}$$

Similar to the proof of Theorem 1, we rewrite $\Theta^s(\mathbf{x}^s, \mathbf{y}^s)^\top \Theta^s(\mathbf{x}^s, \mathbf{y}^s)$, $\Theta^s(\mathbf{x}^s, \mathbf{y}^s)^\top \Theta^t(\mathbf{x}^t, \mathbf{y}^t)$, $\Theta^t(\mathbf{x}^t, \mathbf{y}^t)^\top \Theta^s(\mathbf{x}^s, \mathbf{y}^s)$, \dots using feature and label kernels. First,

$$\begin{aligned}
&\Theta^s(\mathbf{x}^s, \mathbf{y}^s)^\top \Theta^s(\mathbf{x}^s, \mathbf{y}^s) \\
&= \begin{bmatrix} \left\langle \left(\sum_{l=1}^n \mathbf{b}_l \psi(\mathbf{x}_l)^\top \right) \psi(\mathbf{x}_1^s) \otimes \phi(\mathbf{y}_1^s), \left(\sum_{l=1}^n \mathbf{b}_l \psi(\mathbf{x}_l)^\top \right) \psi(\mathbf{x}_1^s) \otimes \phi(\mathbf{y}_1^s) \right\rangle & \cdots & \cdots \\ \left\langle \left(\sum_{l=1}^n \mathbf{b}_l \psi(\mathbf{x}_l)^\top \right) \psi(\mathbf{x}_2^s) \otimes \phi(\mathbf{y}_2^s), \left(\sum_{l=1}^n \mathbf{b}_l \psi(\mathbf{x}_l)^\top \right) \psi(\mathbf{x}_1^s) \otimes \phi(\mathbf{y}_1^s) \right\rangle & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ \left\langle \left(\sum_{l=1}^n \mathbf{b}_l \psi(\mathbf{x}_l)^\top \right) \psi(\mathbf{x}_{n^s}^s) \otimes \phi(\mathbf{y}_{n^s}^s), \left(\sum_{l=1}^n \mathbf{b}_l \psi(\mathbf{x}_l)^\top \right) \psi(\mathbf{x}_1^s) \otimes \phi(\mathbf{y}_1^s) \right\rangle & \cdots & \cdots \end{bmatrix}, \tag{14}
\end{aligned}$$

where $\langle \bullet, \bullet \rangle$ denotes the inner product between two vectors. Moreover, we have the following equation,

$$\begin{aligned}
& \left\langle \left(\sum_{l=1}^n \mathbf{b}_l \psi(\mathbf{x}_l)^\top \right) \psi(\mathbf{x}_i^s) \otimes \phi(\mathbf{y}_j^s), \left(\sum_{l=1}^n \mathbf{b}_l \psi(\mathbf{x}_l)^\top \right) \psi(\mathbf{x}_k) \otimes \phi(\mathbf{y}_m) \right\rangle \\
&= \left(\left(\sum_{l=1}^n \mathbf{b}_l \psi(\mathbf{x}_l)^\top \right) \psi(\mathbf{x}_i^s) \otimes \phi(\mathbf{y}_j^s) \right)^\top \left(\left(\sum_{l=1}^n \mathbf{b}_l \psi(\mathbf{x}_l)^\top \right) \psi(\mathbf{x}_k) \otimes \phi(\mathbf{y}_m) \right) \\
&= \left(\psi(\mathbf{x}_i^s)^\top \left(\sum_{l=1}^n \psi(\mathbf{x}_l) \mathbf{b}_l^\top \right) \otimes \phi(\mathbf{y}_j^s)^\top \right) \left(\left(\sum_{l=1}^n \mathbf{b}_l \psi(\mathbf{x}_l)^\top \right) \psi(\mathbf{x}_k) \otimes \phi(\mathbf{y}_m) \right) \\
&= \left(\psi(\mathbf{x}_i^s)^\top \left(\sum_{l=1}^n \psi(\mathbf{x}_l) \mathbf{b}_l^\top \right) \right) \left(\sum_{l=1}^n \mathbf{b}_l \psi(\mathbf{x}_l)^\top \right) \psi(\mathbf{x}_k) \otimes \left(\phi(\mathbf{y}_j^s)^\top \phi(\mathbf{y}_m) \right) \\
&= \left[k^X(\mathbf{x}_i^s, \mathbf{x}_1), k^X(\mathbf{x}_i^s, \mathbf{x}_2), \dots, k^X(\mathbf{x}_i^s, \mathbf{x}_n) \right] \mathbf{B} \mathbf{B}^\top \left[k^X(\mathbf{x}_1, \mathbf{x}_k), k^X(\mathbf{x}_2, \mathbf{x}_k), \dots, k^X(\mathbf{x}_n, \mathbf{x}_k) \right]^\top \otimes k^Y(\mathbf{y}_j^s, \mathbf{y}_m) \\
&= \left[k^X(\mathbf{x}_i^s, \mathbf{x}_1), k^X(\mathbf{x}_i^s, \mathbf{x}_2), \dots, k^X(\mathbf{x}_i^s, \mathbf{x}_n) \right] \mathbf{B} \mathbf{B}^\top \left[k^X(\mathbf{x}_1, \mathbf{x}_k), k^X(\mathbf{x}_2, \mathbf{x}_k), \dots, k^X(\mathbf{x}_n, \mathbf{x}_k) \right]^\top k^Y(\mathbf{y}_j^s, \mathbf{y}_m) \\
&= \mathbf{K}_{(i, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, k)}^{XX} k^Y(\mathbf{y}_j, \mathbf{y}_m),
\end{aligned} \tag{15}$$

where the subscripts (i, \bullet) and (\bullet, k) denote the i -th row vector and the k -th column vector of a given matrix, respectively. Then, we can obtain the following equation,

$$\begin{aligned}
& \Theta^s(\mathbf{x}^s, \mathbf{y}^s)^\top \Theta^s(\mathbf{x}^s, \mathbf{y}^s) = \\
& \begin{bmatrix} \mathbf{K}_{(1, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, 1)}^{XX} k^Y(\mathbf{y}_1, \mathbf{y}_1) & \dots & \mathbf{K}_{(1, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, n^s)}^{XX} k^Y(\mathbf{y}_1, \mathbf{y}_{n^s}) \\ \mathbf{K}_{(2, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, 1)}^{XX} k^Y(\mathbf{y}_2, \mathbf{y}_1) & \dots & \mathbf{K}_{(2, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, n^s)}^{XX} k^Y(\mathbf{y}_2, \mathbf{y}_{n^s}) \\ \dots & \dots & \dots \\ \mathbf{K}_{(n^s, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, 1)}^{XX} k^Y(\mathbf{y}_{n^s}, \mathbf{y}_1) & \dots & \mathbf{K}_{(n^s, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, n^s)}^{XX} k^Y(\mathbf{y}_{n^s}, \mathbf{y}_{n^s}) \end{bmatrix}.
\end{aligned} \tag{16}$$

Similarly, $\Theta^t(\mathbf{x}^t, \mathbf{y}^t)^\top \Theta^t(\mathbf{x}^t, \mathbf{y}^t) =$

$$\begin{bmatrix} \mathbf{K}_{(n^s+1, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, n^s+1)}^{XX} k^Y(\mathbf{y}_{n^s+1}, \mathbf{y}_{n^s+1}) & \dots & \mathbf{K}_{(n^s+1, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, n)}^{XX} k^Y(\mathbf{y}_{n^s+1}, \mathbf{y}_n) \\ \mathbf{K}_{(n^s+2, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, n^s+1)}^{XX} k^Y(\mathbf{y}_{n^s+2}, \mathbf{y}_{n^s+1}) & \dots & \mathbf{K}_{(n^s+2, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, n)}^{XX} k^Y(\mathbf{y}_{n^s+2}, \mathbf{y}_n) \\ \dots & \dots & \dots \\ \mathbf{K}_{(n, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, n^s+1)}^{XX} k^Y(\mathbf{y}_n, \mathbf{y}_{n^s+1}) & \dots & \mathbf{K}_{(n, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, n)}^{XX} k^Y(\mathbf{y}_n, \mathbf{y}_n) \end{bmatrix}. \tag{17}$$

$\Theta^s(\mathbf{x}^s, \mathbf{y}^s)^\top \Theta^t(\mathbf{x}^t, \mathbf{y}^t) =$

$$\begin{bmatrix} \mathbf{K}_{(1, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, n^s+1)}^{XX} k^Y(\mathbf{y}_1, \mathbf{y}_{n^s+1}) & \dots & \mathbf{K}_{(1, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, n)}^{XX} k^Y(\mathbf{y}_1, \mathbf{y}_n) \\ \mathbf{K}_{(2, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, n^s+1)}^{XX} k^Y(\mathbf{y}_2, \mathbf{y}_{n^s+1}) & \dots & \mathbf{K}_{(2, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, n)}^{XX} k^Y(\mathbf{y}_2, \mathbf{y}_n) \\ \dots & \dots & \dots \\ \mathbf{K}_{(n^s, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, n^s+1)}^{XX} k^Y(\mathbf{y}_{n^s}, \mathbf{y}_{n^s+1}) & \dots & \mathbf{K}_{(n^s, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, n)}^{XX} k^Y(\mathbf{y}_{n^s}, \mathbf{y}_n) \end{bmatrix}. \tag{18}$$

$\Theta^t(\mathbf{x}^t, \mathbf{y}^t)^\top \Theta^s(\mathbf{x}^s, \mathbf{y}^s) =$

$$\begin{bmatrix} \mathbf{K}_{(n^s+1, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, 1)}^{XX} k^Y(\mathbf{y}_{n^s+1}, \mathbf{y}_1) & \cdots & \mathbf{K}_{(n^s+1, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, n^s)}^{XX} k^Y(\mathbf{y}_{n^s+1}, \mathbf{y}_{n^s}) \\ \mathbf{K}_{(n^s+2, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, 1)}^{XX} k^Y(\mathbf{y}_{n^s+2}, \mathbf{y}_1) & \cdots & \mathbf{K}_{(n^s+2, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, n^s)}^{XX} k^Y(\mathbf{y}_{n^s+2}, \mathbf{y}_{n^s}) \\ \cdots & \cdots & \cdots \\ \mathbf{K}_{(n, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, 1)}^{XX} k^Y(\mathbf{y}_n, \mathbf{y}_1) & \cdots & \mathbf{K}_{(n, \bullet)}^{XX} \mathbf{B} \mathbf{B}^\top \mathbf{K}_{(\bullet, n^s)}^{XX} k^Y(\mathbf{y}_n, \mathbf{y}_{n^s}) \end{bmatrix}. \quad (19)$$

According to (15) ~ (18), we could obtain the following equation,

$$\begin{aligned} \mathbb{D}_{\mathcal{H}} &= \text{tr} \left(\begin{bmatrix} \Theta^s(\mathbf{x}^s, \mathbf{y}^s)^\top \Theta^s(\mathbf{x}^s, \mathbf{y}^s) & \Theta^s(\mathbf{x}^s, \mathbf{y}^s)^\top \Theta^t(\mathbf{x}^t, \mathbf{y}^t) \\ \Theta^t(\mathbf{x}^t, \mathbf{y}^t)^\top \Theta^s(\mathbf{x}^s, \mathbf{y}^s) & \Theta^t(\mathbf{x}^t, \mathbf{y}^t)^\top \Theta^t(\mathbf{x}^t, \mathbf{y}^t) \end{bmatrix} \begin{bmatrix} \frac{\mathbf{1}_{n^s \times 1} \mathbf{1}_{n^s \times 1}^\top}{n^s n^s} & \frac{-\mathbf{1}_{n^s \times 1} \mathbf{1}_{n^t \times 1}^\top}{n^s n^t} \\ \frac{-\mathbf{1}_{n^t \times 1} \mathbf{1}_{n^s \times 1}^\top}{n^t n^s} & \frac{\mathbf{1}_{n^t \times 1} \mathbf{1}_{n^t \times 1}^\top}{n^t n^t} \end{bmatrix} \right) \\ &= \text{tr}(\mathbf{B}^\top \mathbf{K}^{XX} (\mathbf{K}^{YY} \odot \mathbf{M}^J) \mathbf{K}^{XX} \mathbf{B}). \end{aligned} \quad (20)$$

□

3 Probability Distribution Distances

3.1 Maximum Mean Discrepancy

The maximum mean discrepancy (MMD) [3] establishes the mean embedding of the marginal probability distribution in a RKHS endowed by the kernel k^X (feature mapping ψ), and using finite samples to empirically estimate the distance between μ_{X^s} (mean embedding of source domain) and μ_{X^t} (mean embedding of target domain) with the Hilbert-Schmidt norm as the following equation,

$$\begin{aligned} \mathbb{D}_{\mathcal{H}}(\mathcal{P}^s(\mathbf{X}^s), \mathcal{P}^t(\mathbf{X}^t)) &= \left\| \mathbb{E}(\psi(\mathbf{X}^s)) - \mathbb{E}(\psi(\mathbf{X}^t)) \right\|_{\mathcal{H}}^2 = \left\| \mu_{X^s} - \mu_{X^t} \right\|_{\mathcal{H}}^2 \\ &= \left\| \frac{1}{n^s} \sum_{i=1}^{n^s} \psi(\mathbf{x}_i) - \frac{1}{n^t} \sum_{j=1}^{n^t} \psi(\mathbf{x}_j) \right\|_{\mathcal{H}}^2 = \text{tr}(\mathbf{K}^{XX} \mathbf{M}^M), \end{aligned} \quad (21)$$

where $\mathbf{K}^{XX} = \psi(\mathbf{X})^\top \psi(\mathbf{X}) \in \mathbb{R}^{n \times n}$ and $k_{ij}^{XX} = k^X(\mathbf{x}_i, \mathbf{x}_j)$. Besides, the MMD matrix \mathbf{M}^M can be computed as below,

$$m_{ij}^M = \begin{cases} 1/(n^s n^s), & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}^s \\ 1/(n^t n^t), & \mathbf{x}_i, \mathbf{x}_j \in \mathcal{D}^t \\ -1/(n^s n^t), & \text{otherwise.} \end{cases} \quad (22)$$

Moreover, the MMD in a projected RKHS is $\text{tr}(\mathbf{B}^\top \mathbf{K}^{XX} \mathbf{M}^m \mathbf{K}^{XX} \mathbf{B})$.

3.2 Class-Wise Maximum Mean Discrepancy

The class-wise maximum mean discrepancy (CMMD) [4] constructs the sum of MMD for each specific class as the following equation,

$$\begin{aligned}
& \mathbb{D}_{\mathcal{H}}\left(\mathcal{P}^s(\mathbf{X}^s|\mathbf{Y}^s), \mathcal{P}^t(\mathbf{X}^t|\mathbf{Y}^t)\right) \\
&= \sum_{c=1}^C \left\| \mathbb{E}\left(\psi(\mathbf{X}^{s,c})\right) - \mathbb{E}\left(\psi(\mathbf{X}^{t,c})\right) \right\|_{\mathcal{H}}^2 = \sum_{c=1}^C \left\| \mu_{X^{s,c}} - \mu_{X^{t,c}} \right\|_{\mathcal{H}}^2 \\
&= \sum_{c=1}^C \left\| \frac{1}{n^{s,c}} \sum_{i=1}^{n^{s,c}} \psi(\mathbf{x}_i) - \frac{1}{n^{t,c}} \sum_{j=1}^{n^{t,c}} \psi(\mathbf{x}_j) \right\|_{\mathcal{H}}^2 = \sum_{c=1}^C \text{tr}(\mathbf{K}^{XX} \mathbf{M}^{C,c}),
\end{aligned} \tag{23}$$

where the MMD matrix $\mathbf{M}^{C,c}$ can be computed as below,

$$m_{ij}^{C,c} = \begin{cases} 1/(n^{s,c}n^{s,c}), & \mathbf{x}_i \in \mathcal{D}^{s,c}, \mathbf{x}_j \in \mathcal{D}^{s,c} \\ 1/(n^{t,c}n^{t,c}), & \mathbf{x}_i \in \mathcal{D}^{t,c}, \mathbf{x}_j \in \mathcal{D}^{t,c} \\ -1/(n^{s,c}n^{t,c}), & \mathbf{x}_i \in \mathcal{D}^{s,c}, \mathbf{x}_j \in \mathcal{D}^{t,c} \\ -1/(n^{t,c}n^{s,c}), & \mathbf{x}_i \in \mathcal{D}^{t,c}, \mathbf{x}_j \in \mathcal{D}^{s,c} \\ 0, & \text{otherwise.} \end{cases} \tag{24}$$

Similarly, the CMMD in a projected RKHS is $\sum_{c=1}^C \text{tr}(\mathbf{B}^\top \mathbf{K}^{XX} \mathbf{M}^{C,c} \mathbf{K}^{XX} \mathbf{B})$.

3.3 Weighted Class-Wise Maximum Mean Discrepancy

To deal with class imbalanced dataset, the weighted class-wise maximum mean discrepancy (WCMMMD) introduces the class prior probability $\mathcal{P}(\mathbf{Y})$ into the CMMD [5], which pays more attention on the large-size categories and is formulated as the following equation,

$$\begin{aligned}
& \sum_{c=1}^C \left\| \frac{\mathcal{P}^s(\mathbf{y}^s=c)}{n^{s,c}} \sum_{i=1}^{n^{s,c}} \psi(\mathbf{x}_i) - \frac{\mathcal{P}^t(\mathbf{y}^t=c)}{n^{t,c}} \sum_{j=1}^{n^{t,c}} \psi(\mathbf{x}_j) \right\|_{\mathcal{H}}^2 \\
&= \sum_{c=1}^C \left\| \frac{1}{n^s} \sum_{i=1}^{n^{s,c}} \psi(\mathbf{x}_i) - \frac{1}{n^t} \sum_{j=1}^{n^{t,c}} \psi(\mathbf{x}_j) \right\|_{\mathcal{H}}^2 = \sum_{c=1}^C \text{tr}(\mathbf{K}^{XX} \mathbf{M}^{WC,c}),
\end{aligned} \tag{25}$$

where $\mathbf{M}^{WC,c}$ can be computed with the following equation,

$$m_{ij}^{WC,c} = \begin{cases} 1/(n^s n^s), & \mathbf{x}_i \in \mathcal{D}^{s,c}, \mathbf{x}_j \in \mathcal{D}^{s,c} \\ 1/(n^t n^t), & \mathbf{x}_i \in \mathcal{D}^{t,c}, \mathbf{x}_j \in \mathcal{D}^{t,c} \\ -1/(n^s n^t), & \mathbf{x}_i \in \mathcal{D}^{s,c}, \mathbf{x}_j \in \mathcal{D}^{t,c} \\ -1/(n^t n^s), & \mathbf{x}_i \in \mathcal{D}^{t,c}, \mathbf{x}_j \in \mathcal{D}^{s,c} \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

Similarly, the WCMMMD in a projected RKHS is $\sum_{c=1}^C \text{tr}(\mathbf{B}^\top \mathbf{K}^{XX} \mathbf{M}^{WC,c} \mathbf{K}^{XX} \mathbf{B})$.

4 The Uniformity of JMMD

Theorem 4 *The marginal, class conditional and weighted class conditional probability distribution distances are three special cases of JMMD with label reproducing kernels \mathbf{K}^1 , \mathbf{K}^2 and \mathbf{K}^3 . $\mathbf{K}^1 = \mathbf{I}_{n \times n}$ is a matrix whose elements are all 1 with the size of $n \times n$, and \mathbf{K}^2 , \mathbf{K}^3 are defined as below,*

$$k_{ij}^2 = \begin{cases} (n^s n^s)/(n^{s,c} n^{s,c}), & \mathbf{x}_i \in \mathcal{D}^{s,c}, \mathbf{x}_j \in \mathcal{D}^{s,c} \\ (n^t n^t)/(n^{t,c} n^{t,c}), & \mathbf{x}_i \in \mathcal{D}^{t,c}, \mathbf{x}_j \in \mathcal{D}^{t,c} \\ (n^s n^t)/(n^{s,c} n^{t,c}), & \mathbf{x}_i \in \mathcal{D}^{s,c}, \mathbf{x}_j \in \mathcal{D}^{t,c} \\ (n^t n^s)/(n^{t,c} n^{s,c}), & \mathbf{x}_i \in \mathcal{D}^{t,c}, \mathbf{x}_j \in \mathcal{D}^{s,c} \\ 0, & \text{otherwise,} \end{cases} \quad (27)$$

$$k_{ij}^3 = \begin{cases} 1, & \mathbf{x}_i \in \mathcal{D}^{s,c}, \mathbf{x}_j \in \mathcal{D}^{s,c} \\ 1, & \mathbf{x}_i \in \mathcal{D}^{t,c}, \mathbf{x}_j \in \mathcal{D}^{t,c} \\ 1, & \mathbf{x}_i \in \mathcal{D}^{s,c}, \mathbf{x}_j \in \mathcal{D}^{t,c} \\ 1, & \mathbf{x}_i \in \mathcal{D}^{t,c}, \mathbf{x}_j \in \mathcal{D}^{s,c} \\ 0, & \text{otherwise,} \end{cases} \quad (28)$$

where the superscript ‘s/t,c’ denotes data points from the c-th class in the source/target domain.

Proof:

As proved before, the formulations of concise JMMD are $\text{tr}(\mathbf{K}^{XX}(\mathbf{K}^{YY} \odot \mathbf{M}^J))$ (in a RKHS) and $\text{tr}(\mathbf{B}^\top \mathbf{K}^{XX}(\mathbf{K}^{YY} \odot \mathbf{M}^J) \mathbf{K}^{XX} \mathbf{B})$ (in a projected RKHS). Moreover, the formulations of marginal probability distribution distance are $\text{tr}(\mathbf{K}^{XX} \mathbf{M}^M)$ (in a RKHS) and $\text{tr}(\mathbf{B}^\top \mathbf{K}^{XX} \mathbf{M}^M \mathbf{K}^{XX} \mathbf{B})$ (in a projected RKHS). The formulations of class conditional probability distribution distance are $\text{tr}(\mathbf{K}^{XX} \mathbf{M}^{C,c})$ (in a RKHS) and $\text{tr}(\mathbf{B}^\top \mathbf{K}^{XX} \mathbf{M}^{C,c} \mathbf{K}^{XX} \mathbf{B})$ (in a projected RKHS). The formulations of weighted class conditional probability distribution distance are $\text{tr}(\mathbf{K}^{XX} \mathbf{M}^{WC,c})$ (in a RKHS) and $\text{tr}(\mathbf{B}^\top \mathbf{K}^{XX} \mathbf{M}^{WC,c} \mathbf{K}^{XX} \mathbf{B})$ (in a projected RKHS). It is easy to verify that $\mathbf{K}^1 \odot \mathbf{M}^J = \mathbf{M}^M$, $\mathbf{K}^2 \odot \mathbf{M}^J = \mathbf{M}^{C,c}$ and $\mathbf{K}^3 \odot \mathbf{M}^J = \mathbf{M}^{WC,c}$. Therefore, the marginal, class conditional and weighted class conditional probability distribution distances are three special cases of JMMD with different label reproducing kernels \mathbf{K}^1 , \mathbf{K}^2 and \mathbf{K}^3 . We will prove \mathbf{K}^1 , \mathbf{K}^2 and \mathbf{K}^3 are the reproducing kernels in next Subsection. \square

5 Reproducing Kernels

Theorem 5 \mathbf{K}^1 , \mathbf{K}^2 , \mathbf{K}^3 and \mathbf{K}^4 are the reproducing kernels, where \mathbf{K}^1 , \mathbf{K}^2 , \mathbf{K}^3 are defined in Theorem 4.

Proof:

\mathbf{K}^1 , \mathbf{K}^2 and \mathbf{K}^3 : According to the Mercer’s theorem [2], we only have to prove that the Gram matrices \mathbf{G}^1 , \mathbf{G}^2 and \mathbf{G}^3 corresponding to \mathbf{K}^1 , \mathbf{K}^2 and \mathbf{K}^3 are semi-positive definite matrices. In fact, the Gram matrices \mathbf{G}^2 and \mathbf{G}^3 can be decomposed into the following equation,

$$\mathbf{G}^2 = \sum_{c=1}^C \mathbf{G}^{2,c}, \quad \mathbf{G}^3 = \sum_{c=1}^C \mathbf{G}^{3,c}. \quad (29)$$

It is obvious that the sum of several semi-positive definite matrices is also a semi-positive definite matrix, thus we only have to prove that the Gram matrices \mathbf{G}^1 , $\mathbf{G}^{2,c}$ and $\mathbf{G}^{3,c}$ are semi-positive definite. \mathbf{G}^1 , $\mathbf{G}^{2,c}$ and $\mathbf{G}^{3,c}$ can be decomposed into the following equations,

$$\mathbf{G}^1 = \mathbf{p}^1 \mathbf{p}^{1\top}, \quad \mathbf{G}^{2,c} = \mathbf{p}^{2,c} \mathbf{p}^{2,c\top}, \quad \mathbf{G}^{3,c} = \mathbf{p}^{3,c} \mathbf{p}^{3,c\top}, \quad (30)$$

where $\mathbf{p}^1 = \mathbf{1}_n$ is a column vector whose elements are all 1, and $\mathbf{p}^{2,c} \in \mathbb{R}^n$, $\mathbf{p}^{3,c} \in \mathbb{R}^n$ could be defined as below,

$$p_i^{2,c} = \begin{cases} n^s/n^{s,c}, & \mathbf{x}_i \in \mathcal{D}^{s,c} \\ n^t/n^{t,c}, & \mathbf{x}_i \in \mathcal{D}^{t,c} \\ 0, & \text{otherwise,} \end{cases} \quad (31)$$

$$p_i^{3,c} = \begin{cases} 1, & \mathbf{x}_i \in \mathcal{D}^{s,c} \\ 1, & \mathbf{x}_i \in \mathcal{D}^{t,c} \\ 0, & \text{otherwise,} \end{cases} \quad (32)$$

where $p_i^{2/3,c}$ is the value of the i -th component of $\mathbf{l}^{2/3,c}$. For $\forall \mathbf{x} \in \mathbb{R}^n$ and $\mathbf{x} \neq 0$, we have,

$$\begin{aligned} \mathbf{x}^\top \mathbf{G}^1 \mathbf{x} &= \mathbf{x}^\top \mathbf{p}^1 \mathbf{p}^{1\top} \mathbf{x} = \mathbf{x}^\top \mathbf{1}_n \mathbf{1}_n^\top \mathbf{x} \\ &= (x_1 + x_2 + \dots + x_n)^2 \geq 0, \end{aligned} \quad (33)$$

$$\begin{aligned} \mathbf{x}^\top \mathbf{G}^{2,c} \mathbf{x} &= \mathbf{x}^\top \mathbf{p}^{2,c} \mathbf{p}^{2,c\top} \mathbf{x} \\ &= (x_1 p_1^{2,c} + x_2 p_2^{2,c} + \dots + x_n p_n^{2,c})^2 \geq 0. \end{aligned} \quad (34)$$

and,

$$\begin{aligned} \mathbf{x}^\top \mathbf{G}^{3,c} \mathbf{x} &= \mathbf{x}^\top \mathbf{p}^{3,c} \mathbf{p}^{3,c\top} \mathbf{x} \\ &= (x_1 p_1^{3,c} + x_2 p_2^{3,c} + \dots + x_n p_n^{3,c})^2 \geq 0. \end{aligned} \quad (35)$$

Therefore, \mathbf{G}^1 , \mathbf{G}^2 and \mathbf{G}^3 are semi-positive definite matrices. Then, \mathbf{K}^1 , \mathbf{K}^2 and \mathbf{K}^3 are the reproducing kernels.

\mathbf{K}^4 : For $\forall \mathbf{x} \in \mathbb{R}^n$ and $\mathbf{x} \neq 0$, due to $w_{ij} \geq 0$, we have,

$$\mathbf{x}^\top \mathbf{G}^4 \mathbf{x} = \sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - x_j)^2 \geq 0, \quad (36)$$

where \mathbf{G}^4 is the Gram matrix of \mathbf{K}^4 and \mathbf{W} is defined as below,

$$w_{ij} = \begin{cases} (n^s n^s)/(n^{s,c} n^{s,c}), & \mathbf{x}_i \in \mathcal{D}^{s,c}, \mathbf{x}_j \in \mathcal{D}^{s,c} \\ (n^t n^t)/(n^{t,c} n^{t,c}), & \mathbf{x}_i \in \mathcal{D}^{t,c}, \mathbf{x}_j \in \mathcal{D}^{t,c} \\ 0, & \text{otherwise.} \end{cases} \quad (37)$$

Therefore, \mathbf{G}^4 is a semi-positive matrix and \mathbf{K}^4 is the reproducing kernel.

□

6 Experiments

We run the JDA+JMMD/HSIC/Our(JMMD-HSIC) and adopt the classifiers of 1-nearest neighbor (1-NN), support vector machines (SVM^{*}), label propagation (LP[†]) [11] and nearest class prototype

^{*}<https://www.csie.ntu.edu.tw/~cjlin/libsvm>

[†]<https://www.escience.cn/people/fpnle/index.html>

Table 1: Ablation study using different classifiers/labels on the Office10-Caltech10 dataset with SURF features.

Classifier	1-NN	SVM		LP		NCP	
Original Features	40.9	47.7		48.3		45.7	
Label	Hard	Hard	Soft	Hard	Soft	Hard	Soft
JDA+JMMD	47.7	50.2	49.2	54.2	52.8	47.8	41.4
JDA+HSIC	47.9	49.0	48.5	54.0	52.8	48.8	46.8
JDA+Our	49.6	50.9	49.9	55.4	54.3	49.6	47.1

Table 2: Comparison average results of our proposed SPL+JMMD-HSIC with state-of-the-art DA methods on Office10-Caltech10 dataset with DECAF-6 features. A, C, D, W in the second row denotes domains of Amazon, Caltech, Dslr, and Webcam, respectively.

Source	Venue	Amazon			Caltech			Dslr			Webcam			Avg.
Target		C	D	W	A	D	W	A	C	W	A	C	D	
PGCD [6]	TIP'23	86.5	90.4	84.1	92.5	92.4	91.2	92.5	87.6	100.0	91.6	85.3	100.0	91.2
RMMD-II [7]	TNNLS'23	88.4	91.7	92.9	93.4	96.8	95.9	93.6	88.9	100.0	92.2	88.9	100.0	93.6
SPL [8]	AAAI'20	87.4	89.2	95.3	92.7	98.7	93.2	92.9	88.6	98.6	92.0	87.0	100.0	93.0
SPL+Our	-	90.0	96.8	93.9	93.7	99.4	93.9	93.8	90.3	100.0	93.3	89.4	99.4	94.5
OGL ² P [9]	TIP'25	89.7	97.5	91.9	94.3	98.7	95.9	94.2	90.2	99.3	94.6	89.5	100.0	94.6
OGL ² P+Our	-	90.2	97.8	93.2	95.3	99.2	95.7	94.6	90.5	100.0	94.2	89.3	100.0	95.0

(NCP[‡]) [8] on the Office-Caltech10 dataset with SURF features (average classification results on 12 DA tasks). As can be seen from Tab. 1, JMMD and HSIC perform better than the original features as JMMD matches the distributions of the source domain and target domain, and HSIC enhances domain-specific discriminative structures. The proposed JMMD-HSIC could achieve the best results no matter what classifiers or labels are, which shows the effectiveness of JMMD-HSIC and indicates that it is necessary to jointly consider JMMD and HSIC for a better DA capacity. Here, the symbol ‘Soft’ denotes the probability soft label and the symbol ‘Hard’ is the hard (one-hot) label. Notably, 1-NN could not produce a soft label thus only ‘Hard’ is reported. It can be seen that the performance of the ‘Soft’ label is even worse than that of the ‘Hard’ label, and it may be because the performance heavily depends on the quality of predicted soft labels of the target domain [12, 13].

We compare our proposed approach with existing state-of-the-art shallow (SPL [8], PGCD [6], RMMD [7]) and deep DA approaches (RSDA-MSTN [10], OGL²P [9]) on D¹ and D². As can be seen from Tabs. 2 and 3, our proposed approach is better than the baseline methods SPL and OGL²P on average, and has achieved 1.5%/0.4% and 0.8/0.7% improvements on the two datasets, respectively. Besides, OGL²P+JMMD-HSIC could achieve the best average results among all compared approaches, which has achieved 0.4% and 0.7% improvements compared with the second-best methods, *i.e.*, OGL²P. Generally speaking, these results can show the effectiveness and competitiveness of our proposed JMMD-HSIC.

[‡]<https://github.com/hellowangqian/domainadaptation-capls>

Table 3: Comparison of average results of our proposed SPL+JMMD-HSIC with state-of-the-art DA methods on ImageCLEF-DA dataset with ResNet-50 features. C, I, P in the second row denotes domains of Caltech-256, ImageNet ILSVRC, and Pascal VOC, respectively.

Source	Venue	Caltech -256		ImageNet ILSVRC		Pascal VOC		Avg.
Target		I	P	C	P	C	I	
RMMD-I [7]	TNNLS'23	93.2	78.3	95.7	79.5	95.5	92.0	89.0
RSDA-MSTN [10]	TPAMI'24	93.3	79.3	97.8	80.5	96.8	94.2	90.3
SPL [8]	AAAI'20	95.7	80.5	96.7	78.3	96.3	94.5	90.3
SPL+Our	-	96.3	81.4	96.7	80.5	96.7	95.0	91.1
OGL ² P [9]	TIP'25	95.8	81.2	96.8	82.2	97.2	95.7	91.5
OGL ² P+Our	-	96.5	81.8	97.4	83.5	97.8	96.0	92.2

References

- [1] Xianda Zhang. *Matrix Analysis and Applications*. Tsinghua University Press, 2013.
- [2] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.
- [3] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 22(2):199–210, 2011.
- [4] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jianguang Sun, and Philip S. Yu. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2200–2207, 2013.
- [5] Jindong Wang, Yiqiang Chen, Shuji Hao, Wenjie Feng, and Zhiqi Shen. Balanced distribution adaptation for transfer learning. In *Proceedings of the IEEE International Conference on Data Mining*, pages 1129–1134, 2017.
- [6] Wenxu Wang, Zhencai Shen, Daoliang Li, Ping Zhong, and Yingyi Chen. Probability-based graph embedding cross-domain and class discriminative feature learning for domain adaptation. *IEEE Transactions on Image Processing*, 32:72–87, 2023.
- [7] Wei Wang, Haojie Li, Zhengming Ding, Feiping Nie, Junyang Chen, Xiao Dong, and Zhihui Wang. Rethinking maximum mean discrepancy for visual domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 34(1):264–277, 2023.
- [8] Qian Wang and Toby P. Breckon. Unsupervised domain adaptation via structured prediction based selective pseudo-labeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6243–6250, 2020.
- [9] Wei Wang, Mengzhu Wang, Chao Huang, Cong Wang, Jie Mu, Feiping Nie, and Xiaochun Cao. Optimal graph learning-based label propagation for cross-domain image classification. *IEEE Transactions on Image Processing*, 34:1529–1544, 2025.
- [10] Xiang Gu, Jian Sun, and Zongben Xu. Unsupervised and semi-supervised robust spherical space domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3):1757–1774, 2024.
- [11] Feiping Nie, Shiming Xiang, Yun Liu, and Changshui Zhang. A general graph-based semi-supervised learning with novel class discovery. *Neural Computing and Applications*, 19(4):549–555, 2010.
- [12] Wei Wang, Baopu Li, Mengzhu Wang, Feiping Nie, Zhihui Wang, and Haojie Li. Confidence regularized label propagation based domain adaptation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6):3319–3333, 2022.

- [13] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3940–3949, 2020.