Supplementary Material for Domain Adaptation by Joint Distribution Invariant Projections

Sentao Chen, Mehrtash Harandi, *Member, IEEE,* Xiaona Jin, and Xiaowei Yang

I. Proofs, Optimization Algorithm, and Statistical test

In this supplementary material, we present the details of the proofs, the optimization algorithm, and the statistical test for the main paper.

1, In Part A, we provide the proof for Proposition 1 in Section III-C.

2, In Part B, we elaborate on the Riemannian Gradient Descent (RGD) technique in Section IV.

3, In Part C, we provide the proof for Theorem 1 in Section IV-A.

4, In Part D, we provide the proof for Lemma 1 in Section V.

5, In Part E, we provide the proof for Theorem 2 in Section V.

6, In Part F, we describe the procedure of the Wilcoxon signed-ranks test conducted in Section VI-C.

This work was supported by the National Natural Science Foundation of China (61906069), Guangdong Basic and Applied Basic Research Foundation (2019A1515011411, 2019A1515011700), Project Funded by China Postdoctoral Science Foundation (2019M662912), Science and Technology Program of Guangzhou (202002030355), and Fundamental Research Funds for the Central Universities (2019MS088). *(Corresponding author: Sentao Chen; Xiaowei Yang.)*

S. Chen, X. Jin and X. Yang are with the School of Software Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: sentaochen@yahoo.com).

M. Harandi is with the Department of Electrical and Computer Systems Engineering, Monash University, Clayton VIC 3800, Australia.

A. Proof of Proposition 1 in Section III-C of the Main Paper

Proof. By Eq. (3) in the main paper, we have

$$\begin{aligned}
&L^2\big(P^s(\boldsymbol{z},y),P^t(\boldsymbol{z},y)\big)\\
&=\int_{(\boldsymbol{z},y)}\big(P^s(\boldsymbol{z},y)-P^t(\boldsymbol{z},y)\big)^2d\boldsymbol{z}dy &(1)\\
&=\int_{(\boldsymbol{z},y)}\big(P^s(\boldsymbol{z},y)^2-2P^s(\boldsymbol{z},y)P^t(\boldsymbol{x},y)+P^t(\boldsymbol{z},y)^2\big)d\boldsymbol{z}dy &(2)\\
&=\int_{(\boldsymbol{z},y)}\big(P^s(\boldsymbol{z},y)^2-P^s(\boldsymbol{z},y)P^t(\boldsymbol{z},y)\big)d\boldsymbol{z}dy\\
&\quad-\int_{(\boldsymbol{z},y)}\big(P^s(\boldsymbol{z},y)P^t(\boldsymbol{z},y)-P^t(\boldsymbol{z},y)^2\big)d\boldsymbol{z}dy &(3)\\
&=\int_{(\boldsymbol{z},y)}\big(P^s(\boldsymbol{z},y)-P^t(\boldsymbol{z},y)\big)P^s(\boldsymbol{z},y)d\boldsymbol{z}dy\\
&\quad-\int_{(\boldsymbol{z},y)}\big(P^s(\boldsymbol{z},y)-P^t(\boldsymbol{z},y)\big)P^t(\boldsymbol{z},y)d\boldsymbol{z}dy. &(4)
\end{aligned}$$

Note that the two terms in the last equation are expectations. According to the law of large numbers [1], given two sets of samples $\mathcal{D}^s_{\varphi_s}$ and $\mathcal{D}^t_{\varphi_t}$, the empirical estimate of the L^2-distance can therefore be obtained as

$$\begin{aligned}
\widehat{L^2}\big(\mathcal{D}^s_{\varphi_s},\mathcal{D}^t_{\varphi_t}\big)&=\frac{1}{m_s}\sum_{i=1}^{m_s}\big(P^s(\boldsymbol{z}^s_i,y^s_i)-P^t(\boldsymbol{z}^s_i,y^s_i)\big)\\
&\quad-\frac{1}{m_t}\sum_{i=1}^{m_t}\big(P^s(\boldsymbol{z}^t_i,y^t_i)-P^t(\boldsymbol{z}^t_i,y^t_i)\big), &(5)
\end{aligned}$$

which concludes the proof. □

B. Optimization Framework for Section IV of the Main paper

Recently, optimization on Riemannian manifolds have become increasingly prevalent in domain adaptation [2], [3], [4], [5]. In the main paper, we will face non-convex optimization problem defined in the product space of two Stiefel manifolds, which can be characterized as

$$\begin{aligned}
&\min_{\boldsymbol{W}_s,\boldsymbol{W}_t} f(\boldsymbol{W}_s,\boldsymbol{W}_t)\\
&\text{s.t. } (\boldsymbol{W}_s,\boldsymbol{W}_t)\in \mathrm{St}(d,D)\times\mathrm{St}(d,D), &(6)
\end{aligned}$$

where $f(\boldsymbol{W}_s,\boldsymbol{W}_t)$ is an objective function and $\mathrm{St}(d,D)\times\mathrm{St}(d,D)$ is the product set of two Stiefel manifolds. The Stiefel manifold $\mathrm{St}(d,D)$ is defined as the space of $(D\times d)$-dimensional matrices with orthonormal columns, $d\leq D$ [6], [4]. That is, $\mathrm{St}(d,D)=\{\boldsymbol{W}\in\mathbb{R}^{D\times d}|\boldsymbol{W}^\top\boldsymbol{W}=\mathbf{I}\}$, where $\mathbf{I}$ is the identity matrix. Due to the concept of product topology [6], [4], the product set $\mathrm{St}(d,D)\times\mathrm{St}(d,D)$ can be given

the structure of a Riemannian manifold. In the following, this product manifold is denoted as $\mathcal{M}_{prod.} = \mathrm{St}(d,D)\times\mathrm{St}(d,D)$. In the experiments of the main paper, we will make use of the first-order Riemannian Gradient Descent (RGD) method [6], [4], [7] to optimize over $\mathcal{M}_{prod.}$.

In general, the RGD method involves calculating the Euclidean gradient of the objective function, then projecting it onto the *tangent space*[1] to get the Riemannian gradient, and finally performing a *retraction* that maps the tangent vector back to the manifold. On the product manifold $\mathcal{M}_{prod.}$, the Riemannian gradient of $f(\boldsymbol{W}_s,\boldsymbol{W}_t)$ is expressed as

$$\begin{aligned}\mathrm{grad}f(\boldsymbol{W}_s,\boldsymbol{W}_t) =&\Big(\nabla_{\boldsymbol{W}_s}(f) - \boldsymbol{W}_s\mathrm{sym}\Big(\boldsymbol{W}_s^\top\nabla_{\boldsymbol{W}_s}(f)\Big),\\ &\nabla_{\boldsymbol{W}_t}(f) - \boldsymbol{W}_t\mathrm{sym}\Big(\boldsymbol{W}_t^\top\nabla_{\boldsymbol{W}_t}(f)\Big)\Big),\end{aligned} \tag{7}$$

where $\nabla_{\boldsymbol{W}}(f)$ is the Euclidean gradient at $\boldsymbol{W}$ and $\mathrm{sym}(\boldsymbol{W}^\top\nabla_{\boldsymbol{W}}(f)) = \frac{1}{2}\Big(\boldsymbol{W}^\top\nabla_{\boldsymbol{W}}(f) + \big(\nabla_{\boldsymbol{W}}(f)\big)^\top\boldsymbol{W}\Big)$. Besides, the *retraction* is given by

$$r_{(\boldsymbol{W}_s,\boldsymbol{W}_t)}(\boldsymbol{Z}_s,\boldsymbol{Z}_t) = \Big(\mathrm{uf}(\boldsymbol{W}_s+\boldsymbol{Z}_s),\mathrm{uf}(\boldsymbol{W}_t+\boldsymbol{Z}_t)\Big), \tag{8}$$

where $\mathrm{uf}(\boldsymbol{W}+\boldsymbol{Z}) = (\boldsymbol{W}+\boldsymbol{Z})[(\boldsymbol{W}+\boldsymbol{Z})^\top(\boldsymbol{W}+\boldsymbol{Z})]^{-\frac{1}{2}}$. Finally, RGD on the product manifold $\mathcal{M}_{prod.}$ can be summarized by the following steps:

(i) Compute $\mathrm{grad}f(\boldsymbol{W}_s,\boldsymbol{W}_t)$ at the current solution $(\boldsymbol{W}_s,\boldsymbol{W}_t)$.

(ii) Update the current solution as

$$(\boldsymbol{W}_s,\boldsymbol{W}_t) \leftarrow r_{(\boldsymbol{W}_s,\boldsymbol{W}_t)}\big(-\alpha\mathrm{grad}f(\boldsymbol{W}_s,\boldsymbol{W}_t)\big),$$

where $\alpha>0$ is the step size.

These steps are repeated until convergence to a stationary point, or until a maximum number of iterations is reached.

C. Proof of Theorem 1 in Section IV-A of the Main paper

Proof. In Eq. (11) of the main paper, we denote $\boldsymbol{A} = (\boldsymbol{H}+\lambda\mathbf{I})^{-1}$, then $\boldsymbol{A} = \boldsymbol{A}^\top$ and $\widehat{\boldsymbol{\theta}} = \boldsymbol{A}\widehat{\boldsymbol{b}}$. With these notations,

$$f(\boldsymbol{W}_s,\boldsymbol{W}_t) = \sum_{i=1}^{m_{st}}\widehat{b}_i\widehat{\theta}_i = \sum_{i,j=1}^{m_{st}}\widehat{b}_i\widehat{b}_j a_{ij}, \tag{9}$$

where a_{ij} is the (i,j)-th element of $\boldsymbol{A}$.

$$\begin{aligned}\nabla_{\boldsymbol{W}_s}(f) &= \sum_{i=1}^{m_{st}}\widehat{\theta}_i\nabla_{\boldsymbol{W}_s}(\widehat{b}_i) + \sum_{i=1}^{m_{st}}\widehat{b}_i\nabla_{\boldsymbol{W}_s}(\widehat{\theta}_i)\\ &= \sum_{i=1}^{m_{st}}\widehat{\theta}_i\nabla_{\boldsymbol{W}_s}(\widehat{b}_i) + \sum_{i,j=1}^{m_{st}}\widehat{b}_i a_{ij}\nabla_{\boldsymbol{W}_s}(\widehat{b}_j)\\ &\quad + \sum_{i,j=1}^{m_{st}}\widehat{b}_i\widehat{b}_j\nabla_{\boldsymbol{W}_s}(a_{ij})\\ &= 2\sum_{i=1}^{m_{st}}\widehat{\theta}_i\nabla_{\boldsymbol{W}_s}(\widehat{b}_i) + \sum_{i=1}^{m_{st}}\widehat{b}_i\widehat{b}_j\nabla_{\boldsymbol{W}_s}(a_{ij}).\end{aligned} \tag{10}$$

[1] The tangent space at a point on a manifold is a vector space that consists of the tangent vectors of all possible curves passing through this point [3].

Similarly,

$$\nabla_{\boldsymbol{W}_t}(f) = 2\sum_{i=1}^{m_{st}}\widehat{\theta}_i\nabla_{\boldsymbol{W}_t}(\widehat{b}_i) + \sum_{i=1}^{m_{st}}\widehat{b}_i\widehat{b}_j\nabla_{\boldsymbol{W}_t}(a_{ij}). \tag{11}$$

Since $\frac{\partial\boldsymbol{Y}^{-1}}{\partial x} = -\boldsymbol{Y}^{-1}\frac{\partial\boldsymbol{Y}}{\partial x}\boldsymbol{Y}^{-1}$ [8], it holds that $\frac{\partial\boldsymbol{A}}{\partial w_{uv}} = -\boldsymbol{A}\frac{\partial\boldsymbol{H}}{\partial w_{uv}}\boldsymbol{A}$, where w_{uv} is the (u,v)-th element of $\boldsymbol{W}_s$ or $\boldsymbol{W}_t$. Hence, we have

$$\begin{aligned}\sum_{i=1}^{m_{st}}\widehat{b}_i\widehat{b}_j\frac{\partial a_{ij}}{\partial w_{uv}} &= \sum_{i=1}^{m_{st}}\widehat{b}_i\widehat{b}_j\Big(\frac{\partial\boldsymbol{A}}{\partial w_{uv}}\Big)_{ij}\\ &= -\widehat{\boldsymbol{b}}^\top\boldsymbol{A}^\top\frac{\partial\boldsymbol{H}}{\partial w_{uv}}\boldsymbol{A}\widehat{\boldsymbol{b}}\\ &= -\widehat{\boldsymbol{\theta}}^\top\frac{\partial\boldsymbol{H}}{\partial w_{uv}}\widehat{\boldsymbol{\theta}}\\ &= -\sum_{i=1}^{m_{st}}\widehat{\theta}_i\widehat{\theta}_j\Big(\frac{\partial\boldsymbol{H}}{\partial w_{uv}}\Big)_{ij}\\ &= -\sum_{i=1}^{m_{st}}\widehat{\theta}_i\widehat{\theta}_j\frac{\partial h_{ij}}{\partial w_{uv}},\end{aligned} \tag{12}$$

where $\Big(\frac{\partial\boldsymbol{A}}{\partial w_{uv}}\Big)_{ij}$ denotes the (i,j)-th element of the matrix $\frac{\partial\boldsymbol{A}}{\partial w_{uv}}$. This immediately leads to the following equations:

$$\sum_{i=1}^{m_{st}}\widehat{b}_i\widehat{b}_j\nabla_{\boldsymbol{W}_s}(a_{ij}) = -\sum_{i=1}^{m_{st}}\widehat{\theta}_i\widehat{\theta}_j\nabla_{\boldsymbol{W}_s}(h_{ij}), \tag{13}$$

$$\sum_{i=1}^{m_{st}}\widehat{b}_i\widehat{b}_j\nabla_{\boldsymbol{W}_t}(a_{ij}) = -\sum_{i=1}^{m_{st}}\widehat{\theta}_i\widehat{\theta}_j\nabla_{\boldsymbol{W}_t}(h_{ij}). \tag{14}$$

Therefore, we have

$$\nabla_{\boldsymbol{W}_s}(f) = 2\sum_{i=1}^{m_{st}}\widehat{\theta}_i\nabla_{\boldsymbol{W}_s}(\widehat{b}_i) - \sum_{i,j=1}^{m_{st}}\widehat{\theta}_i\widehat{\theta}_j\nabla_{\boldsymbol{W}_s}(h_{ij}), \tag{15}$$

$$\nabla_{\boldsymbol{W}_t}(f) = 2\sum_{i=1}^{m_{st}}\widehat{\theta}_i\nabla_{\boldsymbol{W}_t}(\widehat{b}_i) - \sum_{i,j=1}^{m_{st}}\widehat{\theta}_i\widehat{\theta}_j\nabla_{\boldsymbol{W}_t}(h_{ij}). \tag{16}$$

We now elaborate on deriving the gradient $\nabla_{\boldsymbol{W}_s}(\widehat{b}_i)$. The remaining ones $\nabla_{\boldsymbol{W}_t}(\widehat{b}_i), \nabla_{\boldsymbol{W}_s}(h_{ij}), \nabla_{\boldsymbol{W}_t}(h_{ij})$ can also be obtained in a similar way. In particular, we first derive the gradient for the Gaussian kernel function, which is the building block of $\nabla_{\boldsymbol{W}_s}(\widehat{b}_i)$:

$$\begin{aligned}&\nabla_{\boldsymbol{W}}\exp\Big(-\frac{\|\boldsymbol{W}^\top\boldsymbol{x}_i - \boldsymbol{W}^\top\boldsymbol{x}_j\|^2}{2\sigma^2}\Big)\\ &= \frac{-1}{\sigma^2}\exp\Big(-\frac{\|\boldsymbol{W}^\top\boldsymbol{x}_i - \boldsymbol{W}^\top\boldsymbol{x}_j\|^2}{2\sigma^2}\Big)(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^\top\boldsymbol{W},\end{aligned} \tag{17}$$

$$\begin{aligned}&\nabla_{\boldsymbol{W}}\exp\Big(-\frac{\|\boldsymbol{W}^\top\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\Big)\\ &= \frac{-1}{\sigma^2}\exp\Big(-\frac{\|\boldsymbol{W}^\top\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\Big)(\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{W} - \boldsymbol{x}_i\boldsymbol{x}_j^\top).\end{aligned} \tag{18}$$

Therefore, when $1 \le i \le m_s$,

$$\begin{aligned}
&\nabla_{\boldsymbol{W}_s}(\widehat{b_i}) \\
&= \nabla_{\boldsymbol{W}_s}\Big(\frac{1}{m_s}\sum_{n=1}^{m_s}\exp\Big(-\frac{\|\boldsymbol{W}_s^\top \boldsymbol{x}_n^s - \boldsymbol{W}_s^\top \boldsymbol{x}_i\|^2}{2\sigma^2}\Big)\delta(y_n^s = y_i) \\
&\quad - \frac{1}{m_t}\sum_{n=1}^{m_t}\exp\Big(-\frac{\|\boldsymbol{W}_t^\top \boldsymbol{x}_n^t - \boldsymbol{W}_s^\top \boldsymbol{x}_i\|^2}{2\sigma^2}\Big)\delta(y_n^t = y_i)\Big) \\
&= \frac{-1}{m_s\sigma^2}\sum_{n=1}^{m_s}\exp\Big(-\frac{\|\boldsymbol{W}_s^\top \boldsymbol{x}_n^s - \boldsymbol{W}_s^\top \boldsymbol{x}_i\|^2}{2\sigma^2}\Big) \\
&\quad \times \delta(y_n^s = y_i)(\boldsymbol{x}_n^s - \boldsymbol{x}_i)(\boldsymbol{x}_n^s - \boldsymbol{x}_i)^\top \boldsymbol{W}_s \\
&\quad + \frac{1}{m_t\sigma^2}\sum_{n=1}^{m_t}\exp\Big(-\frac{\|\boldsymbol{W}_t^\top \boldsymbol{x}_n^t - \boldsymbol{W}_s^\top \boldsymbol{x}_i\|^2}{2\sigma^2}\Big) \\
&\quad \times \delta(y_n^t = y_i)\big(\boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{W}_s - \boldsymbol{x}_i(\boldsymbol{x}_n^t)^\top \boldsymbol{W}_t\big). \qquad (19)
\end{aligned}$$

When $m_s + 1 \le i \le m_{st}$,

$$\begin{aligned}
&\nabla_{\boldsymbol{W}_s}(\widehat{b_i}) \\
&= \nabla_{\boldsymbol{W}_s}\Big(\frac{1}{m_s}\sum_{n=1}^{m_s}\exp\Big(-\frac{\|\boldsymbol{W}_s^\top \boldsymbol{x}_n^s - \boldsymbol{W}_t^\top \boldsymbol{x}_i\|^2}{2\sigma^2}\Big)\delta(y_n^s = y_i) \\
&\quad - \frac{1}{m_t}\sum_{n=1}^{m_t}\exp\Big(-\frac{\|\boldsymbol{W}_t^\top \boldsymbol{x}_n^t - \boldsymbol{W}_t^\top \boldsymbol{x}_i\|^2}{2\sigma^2}\Big)\delta(y_n^t = y_i)\Big) \\
&= \frac{1}{m_s\sigma^2}\sum_{n=1}^{m_s}\exp\Big(-\frac{\|\boldsymbol{W}_s^\top \boldsymbol{x}_n^s - \boldsymbol{W}_t^\top \boldsymbol{x}_i\|^2}{2\sigma^2}\Big) \\
&\quad \times \delta(y_n^s = y_i)\big(\boldsymbol{x}_n^s(\boldsymbol{x}_n^s)^\top \boldsymbol{W}_s - \boldsymbol{x}_n^s\boldsymbol{x}_i^\top \boldsymbol{W}_t\big). \qquad (20)
\end{aligned}$$

□

D. Proof of Lemma 1 in Section V of the Main Paper

Proof. For any hypothesis $h \in \mathcal{H}$,

$$\begin{aligned}
&\big|\mathbb{E}_{(\boldsymbol{z},y)\sim P^t(\boldsymbol{z},y)}[\ell(h(\boldsymbol{z}),y)] - \mathbb{E}_{(\boldsymbol{z},y)\sim P^s(\boldsymbol{z},y)}[\ell(h(\boldsymbol{z}),y)]\big| \\
&= \Big|\int_{(\boldsymbol{z},y)} \ell(h(\boldsymbol{z}),y)\big(P^t(\boldsymbol{z},y) - P^s(\boldsymbol{z},y)\big)d\boldsymbol{z}dy\Big| \qquad (21) \\
&\le \int_{(\boldsymbol{z},y)} \Big|P^t(\boldsymbol{z},y) - P^s(\boldsymbol{z},y)\Big|d\boldsymbol{z}dy \qquad (22) \\
&\le \frac{1}{M}\int_{(\boldsymbol{z},y)} \Big(P^t(\boldsymbol{z},y) - P^s(\boldsymbol{z},y)\Big)^2 d\boldsymbol{z}dy. \qquad (23)
\end{aligned}$$

The first inequality uses the assumption $\ell(\cdot,\cdot) \le 1$ and the property of integral. The second inequality exploits the assumption $M = \inf_{(\boldsymbol{z},y)\in\mathcal{Z}\times\mathcal{Y}}|P^s(\boldsymbol{z},y) - P^t(\boldsymbol{z},y)| > 0$, which means $\frac{1}{M}|P^s(\boldsymbol{z},y) - P^t(\boldsymbol{z},y)| \ge 1$ and thus $|P^s(\boldsymbol{z},y) - P^t(\boldsymbol{z},y)| \le \frac{1}{M}\big(P^s(\boldsymbol{z},y) - P^t(\boldsymbol{z},y)\big)^2, \forall(\boldsymbol{z},y) \in \mathcal{Z}\times\mathcal{Y}$ follows. Therefore,

$$\begin{aligned}
\mathbb{E}_{(\boldsymbol{z},y)\sim P^t(\boldsymbol{z},y)}[\ell(h(\boldsymbol{z}),y)] &\le \mathbb{E}_{(\boldsymbol{z},y)\sim P^s(\boldsymbol{z},y)}[\ell(h(\boldsymbol{z}),y)] \\
&+ \frac{1}{M}\int_{(\boldsymbol{z},y)} \Big(P^t(\boldsymbol{z},y) - P^s(\boldsymbol{z},y)\Big)^2 d\boldsymbol{z}dy, \qquad (24)
\end{aligned}$$

which concludes the proof. □

E. Proof of Theorem 2 in Section V of the Main Paper

Proof. We first prove a fact, which will be used to obtain our final result.

Fact 1. Let $\mathcal{A}$ and $\mathcal{B}$ be two random events. For any $\delta \in (0,1)$, if $P(\mathcal{A}) \ge 1 - \delta/2$ and $P(\mathcal{B}) \ge 1 - \delta/2$, then the random event $\mathcal{A}\cap\mathcal{B}$ holds with probability at least $1-\delta$.

The proof is straightforward: $P(\mathcal{A}\cap\mathcal{B}) = P(\mathcal{A}) + P(\mathcal{B}) - P(\mathcal{A}\cup\mathcal{B}) \ge 2 - \delta - P(\mathcal{A}\cup\mathcal{B}) \ge 1 - \delta$.

We then bound the source domain generalization error and the L^2-distance respectively by their empirical estimates. For bounding the source error, we apply the Vapnik-Chervonenkis inequality [9]. That is, with probability exceeding $1-\delta/2$, for any hypothesis $h \in \mathcal{H}$, the following inequality holds:

$$\begin{aligned}
&E_{(\boldsymbol{z},y)\sim P^s(\boldsymbol{z},y)}[\ell(h(\boldsymbol{z}),y)] \\
&\le \frac{1}{m}\sum_{i=1}^{m}\ell(h(\boldsymbol{z}_i^s),y_i^s) + \sqrt{\frac{4}{m}\Big(p\log\frac{2em}{p} + \log\frac{8}{\delta}\Big)}. \qquad (25)
\end{aligned}$$

Since the empirical estimate $\widehat{L^2}\big(\mathcal{D}_{\varphi_s}^s, \mathcal{D}_{\varphi_t}^t\big)$ converges in probability to its true value $L^2\big(P^s(\boldsymbol{z},y), P^t(\boldsymbol{z},y)\big)$, for a given δ, there exists a positive integer N_δ, such that when $m > N_\delta$, with probability at least $1-\delta/2$, the following inequality holds:

$$L^2(P^s(\boldsymbol{z},y), P^t(\boldsymbol{z},y)) \le \widehat{L^2}\big(\mathcal{D}_{\varphi_s}^s, \mathcal{D}_{\varphi_t}^t\big) + 1. \qquad (26)$$

Now regard inequalities (25) and (26) respectively as random events $\mathcal{A}$ and $\mathcal{B}$. Based on Lemma 1 in the main paper and Fact 1, we have that with probability at least $1-\delta$, for $m > N_\delta$ and any hypothesis $h \in \mathcal{H}$,

$$\begin{aligned}
&E_{(\boldsymbol{z},y)\sim P^t(\boldsymbol{z},y)}[\ell(h(\boldsymbol{z}),y)] \\
&\le E_{(\boldsymbol{z},y)\sim P^s(\boldsymbol{z},y)}[\ell(h(\boldsymbol{z}),y)] + \frac{1}{M}L^2(P^s(\boldsymbol{z},y), P^t(\boldsymbol{z},y)) \\
&\le \frac{1}{m}\sum_{i=1}^{m}\ell(h(\boldsymbol{z}_i^s),y_i^s) + \sqrt{\frac{4}{m}\Big(p\log\frac{2em}{p} + \log\frac{8}{\delta}\Big)} \\
&\quad + \frac{1}{M}L^2(P^s(\boldsymbol{z},y), P^t(\boldsymbol{z},y)) \qquad (27) \\
&\le \frac{1}{m}\sum_{i=1}^{m}\ell(h(\boldsymbol{z}_i^s),y_i^s) + \sqrt{\frac{4}{m}\Big(p\log\frac{2em}{p} + \log\frac{8}{\delta}\Big)} \\
&\quad + \frac{1}{M}\big(\widehat{L^2}\big(\mathcal{D}_{\varphi_s}^s, \mathcal{D}_{\varphi_t}^t\big) + 1\big), \qquad (28)
\end{aligned}$$

which concludes the proof. □

F. Statistical Test in Section VI-C of the Main Paper

We describe the procedure of the Wilcoxon signed-ranks test [10], [11] on the tasks from Table I in the main paper. The test compares the performance of two algorithms over multiple tasks. Specifically, in each task the mean classification accuracy is adopted as the performance measure of the algorithms. We fix JDIP as a control algorithm, and conduct 13 pairs of tests: SVM versus JDIP, ..., and JDIP-rbf versus JDIP. To run the test, we rank the differences in performance of two algorithms for each task out of N tasks. The differences are ranked according to their absolute values. The smallest absolute value gets the rank of 1, the second smallest gets the

rank of 2, and so on. In case of equality, average ranks are assigned. The statistic of the Wilcoxon signed-ranks test is:

$$z(a,b) = \frac{T(a,b) - N(N+1)/4}{\sqrt{N(N+1)(2N+1)/24}}, \tag{29}$$

where $T(a,b) = \min\{R^+(a,b), R^-(a,b)\}$. $R^+(a,b)$ is the sum of ranks for the tasks on which algorithm b outperforms algorithm a and $R^-(a,b)$ is the sum of ranks for the opposite. They are defined as follows:

$$R^+(a,b) = \sum_{\mathrm{diff}_i>0} \mathrm{rank}(\mathrm{diff}_i) + \frac{1}{2}\sum_{\mathrm{diff}_i=0} \mathrm{rank}(\mathrm{diff}_i), \tag{30}$$

$$R^-(a,b) = \sum_{\mathrm{diff}_i<0} \mathrm{rank}(\mathrm{diff}_i) + \frac{1}{2}\sum_{\mathrm{diff}_i=0} \mathrm{rank}(\mathrm{diff}_i), \tag{31}$$

where diff_i is the difference between the accuracy of two algorithms on the i-th task out of N tasks, and $\mathrm{rank}(\mathrm{diff}_i)$ is the rank of $|\mathrm{diff}_i|$. We fix b as JDIP, and let a vary from SVM to JDIP-rbf in turn. Based on formulas (29)-(31), we can compute $z(a,b)$ for the 13 pairs of tests.

REFERENCES

[1] L. Wasserman, *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[2] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf, "Domain adaptation with conditional transferable components," in *International conference on machine learning*, 2016, pp. 2839–2848.

[3] M. Baktashmotlagh, M. Harandi, and M. Salzmann, "Distribution-matching embedding for visual domain adaptation," *Journal of Machine Learning Research*, vol. 17, no. 108, pp. 1–30, 2016.

[4] S. Herath, M. Harandi, and F. Porikli, "Learning an invariant hilbert space for domain adaptation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 3956–3965.

[5] Y. Liu, W. Tu, B. Du, L. Zhang, and D. Tao, "Homologous component analysis for domain adaptation," *IEEE Transactions on Image Processing*, vol. 29, pp. 1074–1089, 2020.

[6] P.-A. Absil, R.Mahony, and R. Sepulchre, *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[7] T. Zhang and Y. Yang, "Robust pca by manifold optimization," *Journal of Machine Learning Research*, vol. 19, no. 80, pp. 1–39, 2018.

[8] K. B. Petersen and M. S. Pedersen, *The Matrix CookBook*, 2012.

[9] V. N. Vapnik, *Statistical Learning Theory*. John Wiley & Sons, 1998.

[10] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[11] S. Chen, L. Han, X. Liu, Z. He, and X. Yang, "Subspace distribution adaptation frameworks for domain adaptation," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.