Supplementary Materials of "Distribution Free Domain Generalization"

Peifeng Tong [1] **Wu Su** [2] **He Li** [3 4] **Jialin Ding** [3] **Haoxiang Zhan** [3] **Song Xi Chen** [3 1]

A. Explicit form of variables defined in Section 3.2

In equation (18), $\boldsymbol{K}_{i,j,s}^{pcd} \in \mathbb{R}^{cm}$ is the i-th row vector of $\boldsymbol{K}^{pcd}$ corresponding to domain s and class j, namely

$$\boldsymbol{K}_{i,j,s}^{pcd} = \Big(\frac{1}{n_1^1}\sum_{l=1}^{n_1^1} k(\boldsymbol{x}_{j,i}^s, \boldsymbol{x}_{1,l}^1), \ldots, \frac{1}{n_c^1}\sum_{l=1}^{n_c^1} k(\boldsymbol{x}_{j,i}^s, \boldsymbol{x}_{c,l}^1), \frac{1}{n_1^2}\sum_{l=1}^{n_1^2} k(\boldsymbol{x}_{j,i}^s, \boldsymbol{x}_{1,l}^2), \ldots, \frac{1}{n_c^2}\sum_{l=1}^{n_c^2} k(\boldsymbol{x}_{j,i}^s, \boldsymbol{x}_{c,l}^2), \ldots,$$
$$\frac{1}{n_1^m}\sum_{l=1}^{n_1^m} k(\boldsymbol{x}_{j,i}^s, \boldsymbol{x}_{1,l}^m), \ldots, \frac{1}{n_c^m}\sum_{l=1}^{n_c^m} k(\boldsymbol{x}_{j,i}^s, \boldsymbol{x}_{c,l}^m)\Big)^T.$$

For Proposition 1, the averaged Gram matrix $\bar{\boldsymbol{K}}$ can be denoted as

$$\bar{\boldsymbol{K}} = \begin{bmatrix} \frac{1}{n_1}\sum_{j=1}^{n_1} k_{1j}^{11} & \cdots & \frac{1}{n_m}\sum_{j=1}^{n_m} k_{1j}^{1m} \\ \vdots & \ddots & \vdots \\ \frac{1}{n_1}\sum_{j=1}^{n_1} k_{n_1 j}^{11} & \cdots & \frac{1}{n_m}\sum_{j=1}^{n_m} k_{n_1 j}^{1m} \\ \frac{1}{n_1}\sum_{j=1}^{n_1} k_{1j}^{21} & \cdots & \frac{1}{n_m}\sum_{j=1}^{n_m} k_{1j}^{2m} \\ \vdots & \ddots & \vdots \\ \frac{1}{n_1}\sum_{j=1}^{n_1} k_{n_2 j}^{21} & \cdots & \frac{1}{n_m}\sum_{j=1}^{n_m} k_{n_2 j}^{2m} \\ \vdots & \ddots & \vdots \\ \frac{1}{n_1}\sum_{j=1}^{n_1} k_{1j}^{m1} & \cdots & \frac{1}{n_m}\sum_{j=1}^{n_m} k_{1j}^{mm} \\ \vdots & \ddots & \vdots \\ \frac{1}{n_1}\sum_{j=1}^{n_1} k_{n_m j}^{m1} & \cdots & \frac{1}{n_m}\sum_{j=1}^{n_m} k_{n_m j}^{mm} \end{bmatrix} \in \mathbb{R}^{n\times m},$$

where $k_{ij}^{ss'} := k(\boldsymbol{x}_i^s, \boldsymbol{x}_j^{s'})$ is a simplified notation for the elements in Gram matrix $\boldsymbol{K}$.

B. Proof of Proposition 1

B.1. Preliminary

For two different variables $\boldsymbol{x}_i^s$ and $\boldsymbol{x}_j^s$ within the same domain, we have

$$\begin{aligned}\tau^s :=& E(||\boldsymbol{x}_i^s - \boldsymbol{x}_j^s||_2^2/h) = E(||\boldsymbol{\Gamma}^s(\boldsymbol{u}_i^s - \boldsymbol{u}_j^s)||_2^2/h) \\ =& \text{tr}(\boldsymbol{\Gamma}^s E((\boldsymbol{u}_i^s - \boldsymbol{u}_j^s)(\boldsymbol{u}_i^s - \boldsymbol{u}_j^s)^T)\boldsymbol{\Gamma}^{s^T}/h) \\ =& 2\text{tr}(\boldsymbol{\Sigma}^s)/h.\end{aligned}$$

---

[1]Guanghua School of Management, Peking University, Beijing 100871, China [2]Center for Big Data Research, Peking University, Beijing 100871, China [3]School of Mathematical Science, Peking University, Beijing 100871, China [4]Pazhou Lab, Guangzhou 510330, China. Correspondence to: Song Xi Chen <csx@gsm.pku.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

While for different domains s and s', we can similarly derive that

$$\begin{aligned}\tau^{(s,s')} :=& E(||\boldsymbol{x}_i^s - \boldsymbol{x}_j^{s'}||_2^2/h)\\ =& \left(\text{tr}(\boldsymbol{\Sigma}^s) + \text{tr}(\boldsymbol{\Sigma}^{s'}) + ||\boldsymbol{\eta}^s - \boldsymbol{\eta}^{s'}||_2^2\right)/h.\end{aligned}$$

Combining the results in (Yan & Zhang, 2022), we have the following second-order Taylor expansions under assumption 1,

$$f(||\boldsymbol{x}_i^s - \boldsymbol{x}_j^s||_2^2/h) = f(\tau^s) + f^{(1)}(\tau^s)\tilde{X}_{i,j}^s + c_{2,\tau^s}(\tilde{X}_{i,j}^s)(\tilde{X}_{i,j}^s)^2, \tag{B.1}$$

$$f(||\boldsymbol{x}_i^s - \boldsymbol{x}_j^{s'}||_2^2/h) = f(\tau^{(s,s')}) + f^{(1)}(\tau^{(s,s')})\tilde{X}_{i,j}^{(s,s')} + c_{2,\tau^{(s,s')}}(\tilde{X}_{i,j}^{(s,s')})(\tilde{X}_{i,j}^{(s,s')})^2, \tag{B.2}$$

where $\tilde{X}_{i,j}^s = ||\boldsymbol{x}_i^s - \boldsymbol{x}_j^s||_2^2/h - \tau^s$, $\tilde{X}_{i,j}^{(s,s')} = ||\boldsymbol{x}_i^s - \boldsymbol{x}_j^{s'}||_2^2/h - \tau^{(s,s')}$, and $c_{2,\tau}(\cdot)$ is a bounded function only depends on τ and function f by Lemma B.1.

Lemma B.1 (Lemma S1 of Yan & Zhang (2022))**.** *Consider a function of the form $h(x) = g((a+x)^{1/2})$ for $a > 0$ and $x \geq -a$, where g is a real-valued function defined on $[0, +\infty)$. Suppose*

$$\sup_{1\leq s\leq l+1} \sup_{x\geq 0} |g^{(s)}(x)| < \infty.$$

Then we can write h as

$$h(x) = \sum_{s=0}^{l} \frac{h^{(s)}(0)x^s}{s!} + c_{l+1,a}(x)x^{l+1}, \quad \sup_{x\geq -a} |c_{l+1,a}(x)| \leq C,$$

for some constant $C > 0$ and any $x \geq -a$. The subscripts of the function $c(x)$ are used to indicate the dependency on $l+1$ and a.

B.2. The proof

Proposition 1. *Given Assumption 1, the mean and variance for the elements in $\bar{\boldsymbol{K}}$ are*

$$E\Big(\frac{1}{n_{s'}}\sum_{j=1}^{n_{s'}} k_{1j}^{ss'}\Big) = f(\tau^{(s,s')}) + \frac{f^{(2)}(\tau^{(s,s')})}{2}(\tilde{X}_{i,j}^{(s,s')})^2 + O(p^{3/2}h^{-3}), \tag{B.3}$$

$$var\Big(\frac{1}{n_{s'}}\sum_{j=1}^{n_{s'}} k_{1j}^{ss'}\Big) = \frac{1}{h^2}\left(f^{(1)}(\tau^{(s,s')})\right)^2\left(2tr(\boldsymbol{\Sigma}^{s^2}) + \sum_{k=1}^{p'}(E(u_i^s(k)^4) - 3)\sigma_{kk}^2 + 4(\boldsymbol{\eta}^s - \boldsymbol{\eta}^{s'})^T\boldsymbol{\Sigma}^s(\boldsymbol{\eta}^s - \boldsymbol{\eta}^{s'})\right) + O(p^2h^{-4}), \tag{B.4}$$

where $O(\cdot)$ denotes the order of high order non-linear terms and $\boldsymbol{\Gamma}^{s^T}\boldsymbol{\Gamma}^s := (\sigma_{ij})_{p'\times p'}$. While for the covariances, if the two elements are in the same row, we have

$$cov\Big(\frac{1}{n_{s'}}\sum_{j=1}^{n_{s'}} k_{ij}^{ss'}, \frac{1}{n_{s''}}\sum_{j=1}^{n_{s''}} k_{ij}^{ss''}\Big) = O(p^2h^{-2}), \tag{B.5}$$

and if they are in the same column,

$$cov\Big(\frac{1}{n_s}\sum_{j=1}^{n_s} k_{ij}^{s's}, \frac{1}{n_s}\sum_{j=1}^{n_s} k_{lj}^{s''s}\Big) = O(p^2n_s^{-1}h^{-2}), \tag{B.6}$$

otherwise their covariance equal 0.

Proof. To obtain the first two moments of $\bar{\boldsymbol{K}}$, we consider the following two cases:

1. The element $\frac{1}{n_s}\sum_{j=1}^{n_s} k_{ij}^{ss}$ is within the same domain s.

2. The element $\frac{1}{n_{s'}}\sum_{j=1}^{n_{s'}} k_{ij}^{ss'}$ is calculated between domain s and s'.

Case 1:

Now for case 1, using the result in (B.1), we have

$$\begin{aligned}\frac{1}{n_s}\sum_{j=1}^{n_s} k_{ij}^{ss} =&\frac{1}{n_s}\left\{f(0)+(n_s-1)f(\tau^s)+f^{(1)}(\tau^s)\sum_{\substack{j=1\\j\neq i}}^{n_s}\tilde{X}_{i,j}^s+\sum_{\substack{j=1\\j\neq i}}^{n_s}c_{2,\tau^s}(\tilde{X}_{i,j}^s)(\tilde{X}_{i,j}^s)^2\right\}\\ =&f(\tau^s)+\frac{f^{(1)}(\tau^s)}{hn_s}\sum_{\substack{j=1\\j\neq i}}^{n_s}(\boldsymbol{x}_i^s-\boldsymbol{x}_j^s)^T(\boldsymbol{x}_i^s-\boldsymbol{x}_j^s)-\frac{2(n_s-1)f^{(1)}(\tau^s)}{hn_s}\mathrm{tr}(\Sigma^s)+\frac{1}{n_s}\sum_{\substack{j=1\\j\neq i}}^{n_s}c_{2,\tau^s}(\tilde{X}_{i,j}^s)(\tilde{X}_{i,j}^s)^2+O(\frac{1}{n_s})\\ :=&\Delta_0+\Delta_1+\Delta_2+O(\frac{1}{n_s}),\end{aligned}\tag{B.7}$$

where

$$\begin{aligned}\Delta_0 =&f(\tau^s),\\ \Delta_1 =&-\frac{2f^{(1)}(\tau^s)}{hn_s}\sum_{\substack{j=1\\j\neq i}}^{n_s}(\boldsymbol{x}_i^s-\boldsymbol{\eta}^s)^T(\boldsymbol{x}_j^s-\boldsymbol{\eta}^s)+\frac{f^{(1)}(\tau^s)}{hn_s}\sum_{\substack{j=1\\j\neq i}}^{n_s}\left\{||\boldsymbol{x}_i^s-\boldsymbol{\eta}^s||_2^2+||\boldsymbol{x}_j^s-\boldsymbol{\eta}^s||_2^2\right\}-\frac{2(n_s-1)f^{(1)}(\tau^s)}{hn_s}\mathrm{tr}(\boldsymbol{\Sigma}^s),\\ \Delta_2 =&\frac{1}{n_s}\sum_{\substack{j=1\\j\neq i}}^{n_s}c_{2,\tau^s}(\tilde{X}_{i,j}^s)(\tilde{X}_{i,j}^s)^2.\end{aligned}$$

Now we will analyze the asymptotic distribution of Δ_1 first. Let us further decompose Δ_1 as

$$\Delta_{11}=-\frac{2f^{(1)}(\tau^s)}{hn_s}\sum_{\substack{j=1\\j\neq i}}^{n_s}(\boldsymbol{x}_i^s-\boldsymbol{\eta}^s)^T(\boldsymbol{x}_j^s-\boldsymbol{\eta}^s)$$

and

$$\Delta_{12}=\frac{f^{(1)}(\tau^s)}{hn_s}\sum_{\substack{j=1\\j\neq i}}^{n_s}\left\{||\boldsymbol{x}_i^s-\boldsymbol{\eta}^s||_2^2+||\boldsymbol{x}_j^s-\boldsymbol{\eta}^s||_2^2\right\}-\frac{2(n_s-1)f^{(1)}(\tau^s)}{hn_s}\mathrm{tr}(\boldsymbol{\Sigma}^s).$$

Following similar calculations of Chen & Qin (2010), we can show that

$$E(\Delta_{11}) =E(\Delta_{12})=0,\tag{B.8}$$

$$\mathrm{var}(\Delta_{11}) =\frac{4(n_s-1)}{h^2n_s^2}\left(f^{(1)}(\tau^s)\right)^2\mathrm{tr}(\boldsymbol{\Sigma}^{s^2}),\tag{B.9}$$

$$\begin{aligned}\mathrm{var}(\Delta_{12}) =&\frac{n_s-1}{h^2n_s}\left(f^{(1)}(\tau^s)\right)^2\left(\mathrm{var}(\boldsymbol{u}_i^{s^T}\boldsymbol{\Gamma}^{s^T}\boldsymbol{\Gamma}^s\boldsymbol{u}_i^s)\right)\\ =&\frac{n_s-1}{h^2n_s}\left(f^{(1)}(\tau^s)\right)^2\left(2\mathrm{tr}(\boldsymbol{\Sigma}^{s^2})+\sum_{j=1}^{p'}(E(u_i^s(j)^4)-3)\sigma_{jj}^2\right),\end{aligned}\tag{B.10}$$

where $\boldsymbol{\Gamma}^{s^T}\boldsymbol{\Gamma}^s := (\sigma_{ij})_{p'\times p'}$, and the last equality have used Lemma S2 in Yan & Zhang (2022). Note that $\mathrm{var}(\Delta_{12}) = O(1) \gg \mathrm{var}(\Delta_{11}) = O(n_s^{-1})$, thus $\mathrm{var}(\Delta_1) \approx \mathrm{var}(\Delta_{12})$. Asymptotic normality can not hold here since Δ_{12} contains a $||\boldsymbol{x}_i^s-\boldsymbol{\eta}^s||_2^2$ term which is not affected by the change of sample size n_s.

The next goal is to derive the first two moments of Δ_2. To achieve this, let us first consider the variance:

$$
\begin{aligned}
&\mathrm{var}\Big(\frac{1}{n_s}\sum_{\substack{j=1\\ j\neq i}}^{n_s} c_{2,\tau^s}(\tilde{X}^s_{i,j})(\tilde{X}^s_{i,j})^2\Big)\\
=&\frac{1}{n_s^2}\sum_{\substack{j=1\\ j\neq i}}^{n_s}\sum_{\substack{j'=1\\ j'\neq i\\ j\neq j'}}^{n_s}\mathrm{cov}\Big(c_{2,\tau^s}(\tilde{X}^s_{i,j})(\tilde{X}^s_{i,j})^2, c_{2,\tau^s}(\tilde{X}^s_{i,j'})(\tilde{X}^s_{i,j'})^2\Big)+\frac{1}{n_s^2}\sum_{\substack{j=1\\ j\neq i}}^{n_s}\mathrm{var}\Big(c_{2,\tau^s}(\tilde{X}^s_{i,j})(\tilde{X}^s_{i,j})^2\Big)\\
=&\frac{(n_s-1)(n_s-2)}{n_s^2}\mathrm{cov}\Big(c_{2,\tau^s}(\tilde{X}^s_{i,j})(\tilde{X}^s_{i,j})^2, c_{2,\tau^s}(\tilde{X}^s_{i,j'})(\tilde{X}^s_{i,j'})^2\Big)+\frac{(n_s-1)}{n_s^2}\mathrm{var}\Big(c_{2,\tau^s}(\tilde{X}^s_{i,j})(\tilde{X}^s_{i,j})^2\Big)\\
\leq&\frac{(n_s-1)(n_s-2)}{n_s^2}\sqrt{\mathrm{var}\Big(c_{2,\tau^s}(\tilde{X}^s_{i,j})(\tilde{X}^s_{i,j})^2\Big)}\sqrt{\mathrm{var}\Big(c_{2,\tau^s}(\tilde{X}^s_{i,j'})(\tilde{X}^s_{i,j'})^2\Big)}+\frac{(n_s-1)}{n_s^2}\mathrm{var}\Big(c_{2,\tau^s}(\tilde{X}^s_{i,j})(\tilde{X}^s_{i,j})^2\Big)\\
=&\frac{(n_s-1)^2}{n_s^2}\mathrm{var}\Big(c_{2,\tau^s}(\tilde{X}^s_{i,j})(\tilde{X}^s_{i,j})^2\Big)\\
\leq&\frac{(n_s-1)^2}{n_s^2}E\Big(c^2_{2,\tau^s}(\tilde{X}^s_{i,j})(\tilde{X}^s_{i,j})^4\Big)\\
=&O(1)\times E\Big((\tilde{X}^s_{i,j})^4\Big)\\
=&O(1)\times E\Big\{[(\boldsymbol{u}_i^{s^T}\boldsymbol{\Gamma}^{s^T}\boldsymbol{\Gamma}^s\boldsymbol{u}_i^s+\boldsymbol{u}_j^{s^T}\boldsymbol{\Gamma}^{s^T}\boldsymbol{\Gamma}^s\boldsymbol{u}_j^s-2\boldsymbol{u}_i^{s^T}\boldsymbol{\Gamma}^{s^T}\boldsymbol{\Gamma}^s\boldsymbol{u}_j^s-2\mathrm{tr}(\boldsymbol{\Sigma}^s))/h]^4\Big\}.
\end{aligned}
$$

Using Jensen’s inequality, we have

$$
\begin{aligned}
&E\Big\{[(\boldsymbol{u}_i^{s^T}\boldsymbol{\Gamma}^{s^T}\boldsymbol{\Gamma}^s\boldsymbol{u}_i^s+\boldsymbol{u}_j^{s^T}\boldsymbol{\Gamma}^{s^T}\boldsymbol{\Gamma}^s\boldsymbol{u}_j^s-2\boldsymbol{u}_i^{s^T}\boldsymbol{\Gamma}^{s^T}\boldsymbol{\Gamma}^s\boldsymbol{u}_j^s-2\mathrm{tr}(\boldsymbol{\Sigma}^s))/h]^4\Big\}\\
\leq&\frac{C}{h^4}\Big[E((\boldsymbol{u}_i^{s^T}\boldsymbol{\Gamma}^{s^T}\boldsymbol{\Gamma}^s\boldsymbol{u}_i^s-\mathrm{tr}(\boldsymbol{\Sigma}^s))^4)+E((\boldsymbol{u}_i^{s^T}\boldsymbol{\Gamma}^{s^T}\boldsymbol{\Gamma}^s\boldsymbol{u}_j^s)^4)\Big],
\end{aligned}
$$

for some constant $C>0$. By applying Lemma S4 and S5 of Yan & Zhang (2022), we know that

$$
\begin{aligned}
&E((\boldsymbol{u}_i^{s^T}\boldsymbol{\Gamma}^{s^T}\boldsymbol{\Gamma}^s\boldsymbol{u}_i^s-\mathrm{tr}(\boldsymbol{\Sigma}^s))^4)=O(\mathrm{tr}^2(\boldsymbol{\Sigma}^{s^2}))=O(p^2),\\
&E((\boldsymbol{u}_i^{s^T}\boldsymbol{\Gamma}^{s^T}\boldsymbol{\Gamma}^s\boldsymbol{u}_j^s)^4)=O(\mathrm{tr}^2(\boldsymbol{\Sigma}^{s^2}))=O(p^2).
\end{aligned}
$$

Hence

$$
\mathrm{var}\Big(\frac{1}{n_s}\sum_{\substack{j=1\\ j\neq i}}^{n_s} c_{2,\tau^s}(\tilde{X}^s_{i,j})(\tilde{X}^s_{i,j})^2\Big)=O(p^2/h^4). \tag{B.11}
$$

Similarly, we can establish the order for $E(\Delta_2)$. First, using the connection between Taylor expansion, we know that

$$
c_{2,\tau^s}(\tilde{X}^s_{i,j})(\tilde{X}^s_{i,j})^2=\frac{f^{(2)}(\tau^s)}{2}(\tilde{X}^s_{i,j})^2+c_{3,\tau^s}(\tilde{X}^s_{i,j})(\tilde{X}^s_{i,j})^3.
$$

Thus,

$$
\begin{aligned}
&\Big|E(c_{2,\tau^s}(\tilde{X}^s_{i,j})(\tilde{X}^s_{i,j})^2) - \frac{1}{2}f^{(2)}(\tau^s)E((\tilde{X}^s_{i,j})^2)\Big| \\
=&\Big|E\Big(c_{2,\tau^s}(\tilde{X}^s_{i,j})(\tilde{X}^s_{i,j})^2 - \frac{1}{2}f^{(2)}(\tau^s)(\tilde{X}^s_{i,j})^2\Big)\Big| \\
\leq&E\Big(\Big|c_{2,\tau^s}(\tilde{X}^s_{i,j})(\tilde{X}^s_{i,j})^2 - \frac{1}{2}f^{(2)}(\tau^s)(\tilde{X}^s_{i,j})^2\Big|\Big) \\
=&E\Big(\Big|c_{2,\tau^s}(\tilde{X}^s_{i,j}) - \frac{1}{2}f^{(2)}(\tau^s)\Big|(\tilde{X}^s_{i,j})^2\Big) \\
=&E\Big(\Big|c_{3,\tau^s}(\tilde{X}^s_{i,j})(\tilde{X}^s_{i,j})\Big|(\tilde{X}^s_{i,j})^2\Big) \\
=&O(1) \times E\big(\big|\tilde{X}^s_{i,j}\big|^3\big) \\
\leq&O(p^{3/2}/h^3).
\end{aligned} \tag{B.12}
$$

The last piece of the variance is $\text{cov}(\Delta_1, \Delta_2)$. Since $E(\Delta_1) = 0$, we have

$$
\begin{aligned}
&\text{cov}(\Delta_1, \Delta_2) = E(\Delta_1\Delta_2) \\
=&E\Big\{(\frac{1}{n_s}f^{(1)}(\tau^s)\sum_{\substack{j=1\\j\neq i}}^{n_s}\tilde{X}^s_{i,j})(\frac{1}{n_s}\sum_{\substack{j'=1\\j'\neq i}}^{n_s}c_{2,\tau^s}(\tilde{X}^s_{i,j'})(\tilde{X}^s_{i,j'})^2)\Big\} \\
=&E\Big\{\frac{f^{(1)}(\tau^s)}{n_s^2}\sum_j c_{2,\tau^s}(\tilde{X}^s_{i,j'})(\tilde{X}^s_{i,j'})^3 + \frac{f^{(1)}(\tau^s)}{n_s^2}\sum_j\sum_{j'\neq j}(\tilde{X}^s_{i,j})c_{2,\tau^s}(\tilde{X}^s_{i,j'})(\tilde{X}^s_{i,j'})^2)\Big\} \\
\leq&O(p^{3/2}/(n_s h^3)),
\end{aligned} \tag{B.13}
$$

which means the covariance between Δ_1 and Δ_2 is negligible for large sample size.

Combining (B.7) to (B.13), we conclude that for case 1,

$$
E(\frac{1}{n_s}\sum_{j=1}^{n_s}k^{ss}_{ij}) = f(\tau^s) + \frac{1}{2}f^{(2)}(\tau^s)E((\tilde{X}^s_{i,j})^2) + O(p^{3/2}/h^3) \tag{B.14}
$$

$$
\text{var}(\frac{1}{n_s}\sum_{j=1}^{n_s}k^{ss}_{ij}) = \frac{n_s-1}{h^2 n_s}\left(f^{(1)}(\tau^s)\right)^2\left(2\text{tr}(\boldsymbol{\Sigma}^{s^2}) + \sum_{j=1}^{p'}(E(u^s_i(j)^4) - 3)\sigma^2_{jj}\right) + O(p^2/h^4). \tag{B.15}
$$

Case 2:

For case 2, we can similarly obtain that

$$
\frac{1}{n_{s'}}\sum_{j=1}^{n_{s'}}k^{ss'}_{ij} := \Delta_0 + \Delta_{11} + \Delta_{12} + \Delta_{13} + \Delta_2, \tag{B.16}
$$

where

$$\begin{aligned}
\Delta_0 =& f(\tau^{(s,s')}),\\
\Delta_{11} =& -\frac{2f^{(1)}(\tau^{(s,s')})}{hn_{s'}}\sum_{j=1}^{n_{s'}}(\boldsymbol{x}_i^s-\boldsymbol{\eta}^s)^T(\boldsymbol{x}_j^{s'}-\boldsymbol{\eta}^{s'}),\\
\Delta_{12} =& \frac{f^{(1)}(\tau^{(s,s')})}{hn_{s'}}\sum_{j=1}^{n_{s'}}\left(||\boldsymbol{x}_i^s-\boldsymbol{\eta}^s||_2^2+||\boldsymbol{x}_j^{s'}-\boldsymbol{\eta}^{s'}||_2^2\right)-\frac{f^{(1)}(\tau^{(s,s')})}{h}(\text{tr}(\boldsymbol{\Sigma}^s)+\text{tr}(\boldsymbol{\Sigma}^{s'})),\\
\Delta_{13} =& \frac{2f^{(1)}(\tau^{(s,s')})}{h}(\boldsymbol{x}_i^s-\boldsymbol{\eta}^s)^T(\boldsymbol{\eta}^s-\boldsymbol{\eta}^{s'})-\frac{2f^{(1)}(\tau^{(s,s')})}{hn_{s'}}\sum_{j=1}^{n_{s'}}(\boldsymbol{x}_j^{s'}-\boldsymbol{\eta}^{s'})^T(\boldsymbol{\eta}^s-\boldsymbol{\eta}^{s'}),\\
\Delta_2 =& \frac{1}{n_{s'}}\sum_{j=1}^{n_{s'}}c_{2,\tau^{(s,s')}}(\tilde{X}_{i,j}^{(s,s')})(\tilde{X}_{i,j}^{(s,s')})^2.
\end{aligned}$$

Meanwhile, we have

$$E(\Delta_{11}) = E(\Delta_{12}) = E(\Delta_{13}) = 0, \tag{B.17}$$

$$\text{var}(\Delta_{11}) = \frac{4}{h^2n_{s'}}\left(f^{(1)}(\tau^{(s,s')})\right)^2\text{tr}(\boldsymbol{\Sigma}^s\boldsymbol{\Sigma}^{s'}), \tag{B.18}$$

$$\text{var}(\Delta_{12}) = \frac{1}{h^2}\left(f^{(1)}(\tau^{(s,s')})\right)^2\left(2\text{tr}(\boldsymbol{\Sigma}^{s^2})+\sum_{k=1}^{p'}(E(u_i^s(k)^4)-3)\sigma_{kk}^2\right)+O(\frac{1}{n_{s'}}), \tag{B.19}$$

$$\text{var}(\Delta_{13}) = \frac{4}{h^2}\left(f^{(1)}(\tau^{(s,s')})\right)^2(\boldsymbol{\eta}^s-\boldsymbol{\eta}^{s'})^T\boldsymbol{\Sigma}^s(\boldsymbol{\eta}^s-\boldsymbol{\eta}^{s'})+O(\frac{1}{n_{s'}}), \tag{B.20}$$

and

$$\begin{aligned}
\text{var}(\Delta_2) \le& \text{var}\Big(c_{2,\tau^{(s,s')}}(\tilde{X}_{i,j}^{(s,s')})(\tilde{X}_{i,j}^{(s,s')})^2\Big)\\
\le& \frac{1}{h^4}\Big\{E\big((\boldsymbol{u}_i^{s^T}\boldsymbol{\Gamma}^{s^T}\boldsymbol{\Gamma}^s\boldsymbol{u}_i^s-\text{tr}(\boldsymbol{\Sigma}^s))^4\big)+E\big((\boldsymbol{u}_j^{s'^T}\boldsymbol{\Gamma}^{s'^T}\boldsymbol{\Gamma}^{s'}\boldsymbol{u}_i^{s'}-\text{tr}(\boldsymbol{\Sigma}^{s'}))^4\big)+E\big((2\boldsymbol{u}_i^{s^T}\boldsymbol{\Gamma}^{s^T}\boldsymbol{\Gamma}^{s'}\boldsymbol{u}_j^{s'})^4\big)+\\
&E\big((2\boldsymbol{u}_i^{s^T}\boldsymbol{\Gamma}^{s^T}(\boldsymbol{\eta}^s-\boldsymbol{\eta}^{s'}))^4\big)+E\big((2\boldsymbol{u}_j^{s'^T}\boldsymbol{\Gamma}^{s'^T}(\boldsymbol{\eta}^s-\boldsymbol{\eta}^{s'}))^4\big)\Big\}\\
=& \frac{1}{h^4}\Big\{O(\text{tr}^2(\boldsymbol{\Sigma}^{s^2}))+O(\text{tr}^2(\boldsymbol{\Sigma}^{s'^2}))+O(\text{tr}^2(\boldsymbol{\Sigma}^s\boldsymbol{\Sigma}^{s'}))+O(((\boldsymbol{\eta}^s-\boldsymbol{\eta}^{s'})^T\boldsymbol{\Sigma}^s(\boldsymbol{\eta}^s-\boldsymbol{\eta}^{s'}))^2)+\\
&O(((\boldsymbol{\eta}^s-\boldsymbol{\eta}^{s'})^T\boldsymbol{\Sigma}^{s'}(\boldsymbol{\eta}^s-\boldsymbol{\eta}^{s'}))^2)\Big\}\\
=& O(p^2/h^4),
\end{aligned} \tag{B.21}$$

$$E(\Delta_2) = E(c_{2,\tau^{(s,s')}}(\tilde{X}_{i,j}^{(s,s')})(\tilde{X}_{i,j}^{(s,s')})^2) = \frac{f^{(2)}(\tau^{(s,s')})}{2}(\tilde{X}_{i,j}^{(s,s')})^2+O(p^{3/2}/h^3), \tag{B.22}$$

$$\text{cov}(\Delta_1,\Delta_2) \le O(p^{3/2}/(n_{s'}h^3)). \tag{B.23}$$

Thus, for case 2, by (B.16) to (B.23), we have shown that

$$E(\frac{1}{n_{s'}}\sum_{j=1}^{n_{s'}}k_{ij}^{ss'}) = f(\tau^{(s,s')})+\frac{f^{(2)}(\tau^{(s,s')})}{2}(\tilde{X}_{i,j}^{(s,s')})^2+O(p^{3/2}/h^3), \tag{B.24}$$

$$\text{var}(\frac{1}{n_{s'}}\sum_{j=1}^{n_{s'}}k_{ij}^{ss'}) = \frac{1}{h^2}\left(f^{(1)}(\tau^{(s,s')})\right)^2\left(2\text{tr}(\boldsymbol{\Sigma}^{s^2})+\sum_{k=1}^{p'}(E(u_i^s(k)^4)-3)\sigma_{kk}^2+4(\boldsymbol{\eta}^s-\boldsymbol{\eta}^{s'})^T\boldsymbol{\Sigma}^s(\boldsymbol{\eta}^s-\boldsymbol{\eta}^{s'})\right)+O(p^2/h^4). \tag{B.25}$$

Combining (B.14) and (B.15), as well as (B.24) and (B.25), we have proofed the results in (B.3) and (B.4).

Covariance terms:

Consider the covariance terms, if they are in the same row, one can verify that

$$\text{cov}(\frac{1}{n_{s'}}\sum_{j=1}^{n_{s'}} k_{ij}^{ss'}, \frac{1}{n_{s''}}\sum_{j=1}^{n_{s''}} k_{ij}^{ss''}) = \text{cov}(\Delta_{11}^{ss'} + \Delta_{12}^{ss'} + \Delta_{2}^{ss'}, \Delta_{11}^{ss''} + \Delta_{12}^{ss''} + \Delta_{2}^{ss''}), \tag{B.26}$$

where

$$\begin{aligned}
\Delta_{11}^{ss'} =& \frac{f^{(1)}(\tau^{(s,s')})}{h}\Big\{||\boldsymbol{x}_i^s - \boldsymbol{\eta}^s||_2^2 - \text{tr}(\boldsymbol{\Sigma}^s)\Big\},\\
\Delta_{12}^{ss'} =& \frac{2f^{(1)}(\tau^{(s,s')})}{h}(\boldsymbol{x}_i^s - \boldsymbol{\eta}^s)^T(\boldsymbol{\eta}^s - \boldsymbol{\eta}^{s'}),\\
\Delta_{2}^{ss'} =& \frac{1}{n_{s'}}\sum_{j=1}^{n_{s'}} c_{2,\tau^{(s,s')}}(\tilde{X}_{i,j}^{(s,s')})(\tilde{X}_{i,j}^{(s,s')})^2,
\end{aligned}$$

and $\Delta_{11}^{ss''}$, $\Delta_{12}^{ss''}$, $\Delta_{2}^{ss''}$ are defined analogically. We then handle all the covariance terms as follows,

$$\begin{aligned}
\text{cov}(\Delta_{11}^{ss'}, \Delta_{11}^{ss''}) =& \frac{1}{h^2} f^{(1)}(\tau^{(s,s')}) f^{(1)}(\tau^{(s,s'')}) \text{var}\left(\boldsymbol{u}_i^{s^T}\boldsymbol{\Gamma}^{s^T}\boldsymbol{\Gamma}^s\boldsymbol{u}_i^s)\right)\\
=& \frac{1}{h^2} f^{(1)}(\tau^{(s,s')}) f^{(1)}(\tau^{(s,s'')}) \left(2\text{tr}(\boldsymbol{\Sigma}^{s^2}) + \sum_{j=1}^{p'}(E(u_i^s(j)^4) - 3)\sigma_{jj}^2\right)\\
=& O(p/h^2), && \text{(B.27)}\\
\text{cov}(\Delta_{11}^{ss'}, \Delta_{12}^{ss''}) =& \frac{2}{h^2} f^{(1)}(\tau^{(s,s')}) f^{(1)}(\tau^{(s,s'')}) E\left(||\boldsymbol{x}_i^s - \boldsymbol{\eta}^s||_2^2(\boldsymbol{x}_i^s - \boldsymbol{\eta}^s)^T(\boldsymbol{\eta}^s - \boldsymbol{\eta}^{s''})\right)\\
=& O(p^2/h^2), && \text{(B.28)}\\
\text{cov}(\Delta_{11}^{ss'}, \Delta_{2}^{ss''}) \leq& \frac{C}{h^2} f^{(1)}(\tau^{(s,s')}) E\Big\{(||\boldsymbol{x}_i^s - \boldsymbol{\eta}^s||_2^2 - \text{tr}(\boldsymbol{\Sigma}^s))^2 + 2||\boldsymbol{x}_i^s - \boldsymbol{\eta}^s||_2^2(\boldsymbol{x}_i^s - \boldsymbol{\eta}^s)^T(\boldsymbol{\eta}^s - \boldsymbol{\eta}^{s''})\Big\}\\
=& \frac{C}{h^2} f^{(1)}(\tau^{(s,s')}) \left(2\text{tr}(\boldsymbol{\Sigma}^{s^2}) + \sum_{j=1}^{p'}(E(u_i^s(j)^4) - 3)\sigma_{jj}^2\right) + O(p^2/h^2),\\
=& O(p^2/h^2) && \text{(B.29)}\\
\text{cov}(\Delta_{12}^{ss'}, \Delta_{12}^{ss''}) =& \frac{2}{h^2} f^{(1)}(\tau^{(s,s')}) f^{(1)}(\tau^{(s,s'')})(\boldsymbol{\eta}^s - \boldsymbol{\eta}^{s'})^T\boldsymbol{\Sigma}^s(\boldsymbol{\eta}^s - \boldsymbol{\eta}^{s''})\\
=& O(p^2/h^2), && \text{(B.30)}\\
\text{cov}(\Delta_{12}^{ss'}, \Delta_{2}^{ss''}) \leq& \frac{2C}{h^2} f^{(1)}(\tau^{(s,s')}) E\Big\{||\boldsymbol{x}_i^s - \boldsymbol{\eta}^s||_2^2(\boldsymbol{x}_i^s - \boldsymbol{\eta}^s)^T(\boldsymbol{\eta}^s - \boldsymbol{\eta}^{s'}) + 2(\boldsymbol{x}_i^s - \boldsymbol{\eta}^s)^T(\boldsymbol{\eta}^s - \boldsymbol{\eta}^{s'})(\boldsymbol{x}_i^s - \boldsymbol{\eta}^s)^T(\boldsymbol{\eta}^s - \boldsymbol{\eta}^{s''})\Big\}\\
=& \frac{4C}{h^2} f^{(1)}(\tau^{(s,s')})(\boldsymbol{\eta}^s - \boldsymbol{\eta}^{s'})^T\boldsymbol{\Sigma}(\boldsymbol{\eta}^s - \boldsymbol{\eta}^{s''}) + O(p^2/h^2)\\
=& O(p^2/h^2), && \text{(B.31)}\\
\text{cov}(\Delta_{2}^{ss'}, \Delta_{2}^{ss''}) =& \text{cov}\Big\{\frac{1}{h^2 n_{s'}}\sum_{j=1}^{n_{s'}} c_{2,\tau^{(s,s')}}(\tilde{X}_{i,j}^{(s,s')})(||\boldsymbol{x}_i^s - \boldsymbol{\eta}^s||_2^2 - \text{tr}(\boldsymbol{\Sigma}^s) + 2(\boldsymbol{x}_i^s - \boldsymbol{\eta}^s)^T(\boldsymbol{\eta}^s - \boldsymbol{\eta}^{s'})),\\
& \frac{1}{h^2 n_{s''}}\sum_{j=1}^{n_{s''}} c_{2,\tau^{(s,s'')}}(\tilde{X}_{i,j}^{(s,s'')})(||\boldsymbol{x}_i^s - \boldsymbol{\eta}^s||_2^2 - \text{tr}(\boldsymbol{\Sigma}^s) + 2(\boldsymbol{x}_i^s - \boldsymbol{\eta}^s)^T(\boldsymbol{\eta}^s - \boldsymbol{\eta}^{s''}))\Big\}\\
\leq& \frac{C^2}{h^4}\left((2\text{tr}(\boldsymbol{\Sigma}^{s^2}) + \sum_{j=1}^{p'}(E(u_i^s(j)^4) - 3)\sigma_{jj}^2) + 4(\boldsymbol{\eta}^s - \boldsymbol{\eta}^{s'})^T\boldsymbol{\Sigma}^s(\boldsymbol{\eta}^s - \boldsymbol{\eta}^{s''}) + O(p^2/h^2)\right)\\
=& O(p^2/h^4), && \text{(B.32)}
\end{aligned}$$

recall that $c_{2,\tau}(\cdot)$ is a bounded function such that there exists some $C > 0$ satisfies $c_{2,\tau}(\cdot) \leq C$, and by Assumption 1.2, $O(\boldsymbol{\Sigma}^{s^2})) = O(p)$, $O(\sum_{j=1}^{p'}(E(u_i^s(j)^4) - 3)\sigma_{jj}^2) = O(p)$, $O(E(||\boldsymbol{x}_i^s - \boldsymbol{\eta}^s||_2^2(\boldsymbol{x}_i^s - \boldsymbol{\eta}^s)^T(\boldsymbol{\eta}^s - \boldsymbol{\eta}^{s''}))) = O(p^2)$, $O((\boldsymbol{\eta}^s - \boldsymbol{\eta}^{s'})^T\boldsymbol{\Sigma}^s(\boldsymbol{\eta}^s - \boldsymbol{\eta}^{s''})) = O(p^2)$. By (B.26) to (B.32), we conclude that the covariance of two elements belong to the same row is not affected by the sample size n_s since

$$\begin{aligned}
\text{cov}(\frac{1}{n_{s'}}\sum_{j=1}^{n_{s'}} k_{ij}^{ss'}, \frac{1}{n_{s''}}\sum_{j=1}^{n_{s''}} k_{ij}^{ss''}) =& \text{cov}(\Delta_{11}^{ss'}, \Delta_{11}^{ss''}) + 2\text{cov}(\Delta_{11}^{ss'}, \Delta_{12}^{ss''}) + 2\text{cov}(\Delta_{11}^{ss'}, \Delta_{2}^{ss''}) + \text{cov}(\Delta_{12}^{ss'}, \Delta_{12}^{ss''}) + \\
& 2\text{cov}(\Delta_{12}^{ss'}, \Delta_{2}^{ss''}) + \text{cov}(\Delta_{2}^{ss'}, \Delta_{2}^{ss''}) \\
=& O(p^2/h^2)
\end{aligned}$$

Next, for two elements within the same column, we have

$$\text{cov}(\frac{1}{n_s}\sum_{j=1}^{n_s} k_{ij}^{s's}, \frac{1}{n_s}\sum_{j=1}^{n_s} k_{lj}^{s''s}) = \text{cov}(\Delta_{11}^{s's} + \Delta_{12}^{s's} + \Delta_{2}^{s's}, \Delta_{11}^{s''s} + \Delta_{12}^{s''s} + \Delta_{2}^{s''s}),$$

where

$$\begin{aligned}
\Delta_{11}^{s's} =& \frac{f^{(1)}(\tau^{(s,s')})}{hn_s}\sum_{j=1}^{n_s}\Big\{||\boldsymbol{x}_j^s - \boldsymbol{\eta}^s||_2^2 - \text{tr}(\boldsymbol{\Sigma}^s)\Big\}, \\
\Delta_{12}^{s's} =& -\frac{2f^{(1)}(\tau^{(s,s')})}{hn_s}\sum_{j=1}^{n_s}(\boldsymbol{x}_j^s - \boldsymbol{\eta}^s)^T(\boldsymbol{\eta}^{s'} - \boldsymbol{\eta}^s), \\
\Delta_{2}^{s's} =& \frac{1}{n_s}\sum_{j=1}^{n_s} c_{2,\tau^{(s,s')}}(\tilde{X}_{i,j}^{(s',s)})(\tilde{X}_{i,j}^{(s',s)})^2,
\end{aligned}$$

and $\Delta_{11}^{s''s}$, $\Delta_{12}^{s''s}$, $\Delta_{2}^{s''s}$ are defined analogically. Mimic the previous calculation, we have

$$\begin{aligned}
&\text{cov}(\Delta_{11}^{s's}, \Delta_{11}^{s''s}) = \frac{1}{h^2n_s}f^{(1)}(\tau^{(s,s')})f^{(1)}(\tau^{(s,s'')})\left(2\text{tr}(\Sigma^{s^2}) + \sum_{j=1}^{p'}(E(u_i^s(j)^4) - 3)\sigma_{jj}^2\right) = O(p/(h^2n_s)), \\
&\text{cov}(\Delta_{11}^{s's}, \Delta_{12}^{s''s}) = \frac{2}{h^2n_s}f^{(1)}(\tau^{(s,s')})f^{(1)}(\tau^{(s,s'')})E\left(||\boldsymbol{x}_j^s - \boldsymbol{\eta}^s||_2^2(\boldsymbol{x}_j^s - \boldsymbol{\eta}^s)^T(\boldsymbol{\eta}^{s''} - \boldsymbol{\eta}^s)\right) = O(p^2/(h^2n_s)), \\
&\text{cov}(\Delta_{11}^{s's}, \Delta_{2}^{s''s}) \leq \frac{C}{h^2n_s}f^{(1)}(\tau^{(s,s')})\left(2\text{tr}(\boldsymbol{\Sigma}^{s^2}) + \sum_{j=1}^{p'}(E(u_i^s(j)^4) - 3)\sigma_{jj}^2\right) + O(p^2/(h^2n_s)) = O(p^2/(h^2n_s)), \\
&\text{cov}(\Delta_{12}^{s's}, \Delta_{12}^{s''s}) = \frac{2}{h^2n_s}f^{(1)}(\tau^{(s,s')})f^{(1)}(\tau^{(s,s'')})(\boldsymbol{\eta}^{s'} - \boldsymbol{\eta}^s)^T\Sigma^s(\boldsymbol{\eta}^{s''} - \boldsymbol{\eta}^s) = O(p^2/(h^2n_s)), \\
&\text{cov}(\Delta_{12}^{s's}, \Delta_{2}^{s''s}) \leq \frac{4C}{h^2n_s}f^{(1)}(\tau^{(s,s')})(\boldsymbol{\eta}^s - \boldsymbol{\eta}^{s'})^T\Sigma^s(\boldsymbol{\eta}^s - \boldsymbol{\eta}^{s''}) + O(p^2/(h^2n_s)) = O(p^2/(h^2n_s)), \\
&\text{cov}(\Delta_{2}^{s's}, \Delta_{2}^{s''s}) \leq \frac{C^2}{h^4n_s}\Big((2\text{tr}(\boldsymbol{\Sigma}^{s^2}) + \sum_{j=1}^{p'}(E(u_i^s(j)^4) - 3)\sigma_{jj}^2) + 4(\boldsymbol{\eta}^{s'} - \boldsymbol{\eta}^s)^T\boldsymbol{\Sigma}^s(\boldsymbol{\eta}^{s''} - \boldsymbol{\eta}^s) + O(p^2/h^2)\Big) = O(p^2/(h^4n_s)),
\end{aligned}$$

which means $\text{cov}(\frac{1}{n_s}\sum_{j=1}^{n_s} k_{ij}^{s's}, \frac{1}{n_s}\sum_{j=1}^{n_s} k_{lj}^{s''s}) = O(n_s^{-1})$ for fixed h and p. To complete our discussion, if two elements are not within the same row or column, we have their covariance equal 0. □

C. Proof of Theorem 1

C.1. Preliminary

To prove Theorem 1, we first present the classical generalization bound towards IID data. The empirical ρ-margin loss in IID scenario given g and $\rho > 0$ is denoted as

$$\hat{R}_{n,\rho}(g) = \frac{1}{n}\sum_{i=1}^{n} l_\rho(r_g(\tilde{\boldsymbol{x}}_i, y_i)).$$

Consider the set of scoring functions $g \in \mathcal{G}$, we define $\Pi(\mathcal{G})$ by

$$\Pi(\mathcal{G}) = \{\tilde{\boldsymbol{x}} \mapsto g(\tilde{\boldsymbol{x}}, y) : y \in \mathcal{Y}, g \in \mathcal{G}\},$$

and the Rademacher complexity

$$\mathfrak{R}_n(\Pi(\mathcal{G})) = E_{(\tilde{x},y),\sigma}\Big\{\sup_{g\in\mathcal{G}}\sum_{i=1}^{n}\sigma_i g(\tilde{\boldsymbol{x}}_i, y)\Big\},$$

where $\sigma_i \in \{-1, 1\}$ are independent Rademacher random variables with equally probabilities.

The following theorem gives a generalization bound for multi-class classification.

Lemma C.1 (Theorem 9.2 of Mohri et al. (2018))**.** *Let $\mathcal{G} \subset \mathbb{R}^{\mathcal{X}\times\mathcal{Y}}$ be a set of scoring functions with $\mathcal{Y} = \{1, \ldots, c\}$. Fix $\rho > 0$, for any $\delta > 0$, with probability at least $1 - \delta$, the following multi-class classification generalization bound holds for all $g \in \mathcal{G}$:*

$$R(g) \leq \hat{R}_{n,\rho}(g) + \frac{4c}{\rho}\mathfrak{R}_n(\Pi(\mathcal{G})) + \sqrt{\frac{\log\delta^{-1}}{2n}}.$$

Combining our kernel based distribution free domain generalization algorithm and a linear classifier, we can further upper bound the Rademacher complexity $\mathfrak{R}_n(\Pi(\mathcal{G}))$ as follows.

Lemma C.2 (Modified from Proposition 9.3 of Mohri et al. (2018))**.** *Let $\bar{k} : \tilde{\mathcal{X}} \times \tilde{\mathcal{X}} \to \mathbb{R}$ be a positive definite symmetric kernel and let $\phi_{\bar{k}} : \tilde{\mathcal{X}} \to \mathcal{H}_{\bar{k}}$ be a feature mapping associated to $\bar{k}$. Assume that there exists $r > 0$ such that $\bar{k}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}}) \leq r^2$ for all $\tilde{\boldsymbol{x}} \in \tilde{\mathcal{X}}$. Then, for any $n \geq 1$, $\mathfrak{R}_n(\Pi(\mathcal{G}_{\bar{k}}))$ can be bounded as follows:*

$$\mathfrak{R}_n(\Pi(\mathcal{G}_{\bar{k}})) \leq \sqrt{\frac{r^2q^2\Lambda^2}{n}}.$$

We note that by Assumption 2, $\bar{k}$ is a universal kernel (Blanchard et al., 2011), thus is also a positive definite symmetric kernel (Sriperumbudur et al., 2011). The following Lemma gives the upper bound of $\bar{k}$.

Lemma C.3. *Assume Assumption 2 holds, the kernel $\bar{k}$ is bounded using Cauchy-Schwarz inequality on the equation $\bar{k}(\tilde{x}, \tilde{x}) = \langle\bar{k}(\tilde{x}, \cdot), \bar{k}(\tilde{x}, \cdot)\rangle$, say*

$$||\bar{k}(\tilde{x}, \cdot)|| = ||\mathfrak{K}(\mu(P_X), \cdot) \otimes k_1(\boldsymbol{x}, \cdot)|| \leq U_1||\mathfrak{K}(\mu(P_X), \cdot)|| \leq L_{\mathfrak{K}}U_1U_2. \tag{C.33}$$

Since all the conditions required by Lemma C.2 are satisfied. We finally give the generalization bound combining the results in Lemmas C.1 and C.2.

Theorem C.1. *Let $\bar{k} : \tilde{\mathcal{X}} \times \tilde{\mathcal{X}} \to \mathbb{R}$ be a positive definite symmetric kernel and let $\phi_{\bar{k}} : \tilde{\mathcal{X}} \to \mathcal{H}_{\bar{k}}$ be a feature mapping associated to $\bar{k}$. Assume that there exists $r > 0$ such that $\bar{k}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}}) \leq r^2$ for all $\tilde{\boldsymbol{x}} \in \tilde{\mathcal{X}}$. Fix $\rho > 0$, for any $\delta > 0$, with probability at least $1 - \delta$, the following multi-class classification generalization bound holds for all $g \in \mathcal{G}_{\bar{k}}$:*

$$R(g) \leq \hat{R}_{n,\rho}(g) + \frac{4cq}{\rho}\sqrt{\frac{r^2\Lambda^2}{n}} + \sqrt{\frac{\log\delta^{-1}}{2n}}. \tag{C.34}$$

We note that (C.34) is only applicable if we treat $(\tilde{\boldsymbol{x}}, y)$ IID in training and test domains. However, $(\tilde{\boldsymbol{x}}, y)$ is not IID even within a given class and domain, and we need a new generalization bound as stated in Theorem 1.

C.2. The proof

To begin our proof, we first decompose $R(g) - \hat{R}_{n,\rho}(g)$ into two parts.

$$\begin{aligned}
R(g) - \hat{R}_{n,\rho}(g) =& R(g) - \frac{1}{cm}\sum_{s=1}^{m}\sum_{j=1}^{c}\frac{1}{\bar{n}}\sum_{i=1}^{\bar{n}} l_\rho(r_g(\tilde{\boldsymbol{x}}^s_{j,i}, j)) \\
\leq& \Big|\frac{1}{cm}\sum_{s=1}^{m}\sum_{j=1}^{c}\frac{1}{\bar{n}}\sum_{i=1}^{\bar{n}} \Big[l_\rho(r_g(\hat{P}^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, j) - l_\rho(r_g(P^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, j)\Big]\Big| + \\
& \Big\{R(g) - \frac{1}{cm}\sum_{s=1}^{m}\sum_{j=1}^{c}\frac{1}{\bar{n}}\sum_{i=1}^{\bar{n}} l_\rho(r_g(P^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, j)\Big\} \\
:=& (I) + (II)
\end{aligned}$$

To control the first term, we have

$$\begin{aligned}
(I) \leq& \frac{1}{cm}\sum_{s=1}^{m}\sum_{j=1}^{c}\frac{1}{\bar{n}}\sum_{i=1}^{\bar{n}} \Big|l_\rho(r_g(\hat{P}^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, j)) - l_\rho(r_g(P^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, j))\Big| \\
\leq& \frac{1}{\rho cm}\sum_{s=1}^{m}\sum_{j=1}^{c}\frac{1}{\bar{n}}\sum_{i=1}^{\bar{n}} \Big|r_g(\hat{P}^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, j) - r_g(P^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, j)\Big| \\
\leq& \frac{1}{\rho cm}\sum_{s=1}^{m}\sum_{j=1}^{c}\frac{1}{\bar{n}}\sum_{i=1}^{\bar{n}} \Big\{\Big|g(\hat{P}^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, j) - g(P^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, j)\Big| + \\
& \Big|\max_{y'\neq j}(g(\hat{P}^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y')) - \max_{y'\neq j}(g(P^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y'))\Big|\Big\} \\
\leq& \frac{1}{\rho cm}\sum_{s=1}^{m}\sum_{j=1}^{c} \Big\|\Big\{\Big|g(\hat{P}^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, j) - g(P^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, j)\Big| + \\
& \Big|\max_{y'\neq j}(g(\hat{P}^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y')) - \max_{y'\neq j}(g(P^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y'))\Big|\Big\}\Big\|_\infty.
\end{aligned} \tag{C.35}$$

If $\max_{y'\neq j}(g(\hat{P}^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y')) - \max_{y'\neq j}(g(P^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y')) \geq 0$, let $y_{max} = \arg\max_{y'\neq j}(g(\hat{P}^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y'))$, then

$$\max_{y'\neq j}(g(\hat{P}^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y')) - \max_{y'\neq j}(g(P^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y')) \leq g(\hat{P}^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y_{max}) - g(P^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y_{max}).$$

Similarly, if $\max_{y'\neq j}(g(\hat{P}^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y')) - \max_{y'\neq j}(g(P^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y')) < 0$, let $y_{max} = \arg\max_{y'\neq j}(g(P^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y'))$, we have

$$\max_{y'\neq j}(g(\hat{P}^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y')) - \max_{y'\neq j}(g(P^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y')) \geq g(\hat{P}^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y_{max}) - g(P^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y_{max}).$$

Hence,

$$\Big|\max_{y'\neq j}(g(\hat{P}^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y')) - \max_{y'\neq j}(g(P^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y'))\Big| \leq \Big|g(\hat{P}^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y_{max}) - g(P^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y_{max})\Big|. \tag{C.36}$$

For a given $y \in \{1, \dots, c\}$, note that

$$\begin{aligned}
&\Big|g(\hat{P}^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y) - g(P^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, y)\Big| \\
\leq &||\boldsymbol{a}_y^T \boldsymbol{W}^T|| \times ||\phi_{\bar{k}}(\hat{P}^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}) - \phi_{\bar{k}}(P^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i})|| \\
\leq &q\Lambda (k_1(\boldsymbol{x}^s_{j,i}, \boldsymbol{x}^s_{j,i}))^{1/2} ||\phi_{\mathfrak{K}}(\mu_{\hat{P}^{(s)}_{X|Y=j}}) - \phi_{\mathfrak{K}}(\mu_{P^{(s)}_{X|Y=j}})|| \\
\leq &q\Lambda U_1 L_{\mathfrak{K}} ||\mu_{\hat{P}^{(s)}_{X|Y=j}} - \mu_{P^{(s)}_{X|Y=j}}|| \\
= &q\Lambda U_1 L_{\mathfrak{K}} ||\frac{1}{\bar{n}} \sum_{i=1}^{\bar{n}} \phi_{k_2}(\boldsymbol{x}^s_{j,i}) - \mu_{P^{(s)}_{X|Y=j}}||.
\end{aligned} \tag{C.37}$$

By applying the Hoeffding's inequality in a Hilbert space, with probability $1 - \delta$,

$$||\frac{1}{\bar{n}} \sum_{i=1}^{\bar{n}} \phi_{k_2}(\boldsymbol{x}^s_{j,i}) - \mu_{P^{(s)}_{X|Y=j}}|| \leq 3U_2 \sqrt{\frac{\log 2\delta^{-1}}{\bar{n}}} \tag{C.38}$$

Combining (C.35) to (C.38), with the union bound over all domain and classes, with probability $1 - \delta$, we have

$$(I) \leq \frac{6}{\rho} q\Lambda U_1 U_2 L_{\mathfrak{K}} \sqrt{\frac{\log 2cm\delta^{-1}}{\bar{n}}}, \tag{C.39}$$

which gives the upper bound towards (I).

Next, we will turn to control the second term (II). To achieve this, we first define the expected ρ-margin loss condition on $P^{(s)}_{X|Y=j}$ as

$$R(g|P^{(s)}_{X|Y=j}) = E_{\tilde{x} \sim P^{(s)}_{X|Y=j}} I(r_g(\tilde{\boldsymbol{x}}_i, j) \leq 0),$$

and further decompose (II) as

$$\begin{aligned}
(II) = &\frac{1}{cm} \sum_{s=1}^{m} \sum_{j=1}^{c} \frac{1}{\bar{n}} \sum_{i=1}^{\bar{n}} \Big\{R(g|P^{(s)}_{X|Y=j}) - l_\rho(r_g(P^{(s)}_{X|Y=j}, \boldsymbol{x}^s_{j,i}, j)\Big\} + \frac{1}{cm} \sum_{s=1}^{m} \sum_{j=1}^{c} \Big\{R(g) - R(g|P^{(s)}_{X|Y=j})\Big\} \\
:= &(IIa) + (IIb)
\end{aligned}$$

Now for (IIa), noticed that given $P^{(s)}_{X|Y=j}$, $\boldsymbol{x}^s_{j,i}$ are IID generated within domain s and class j. While for (IIb), $P^{(s)}_{X|Y=j}$ are IID generated since we have assumed $n^s_j = \bar{n}$. Thus, we can apply Theorem C.1 to upper bound (IIa) and (IIb), as stated below.

$$\begin{aligned}
II(a) \leq &\frac{4cq\Lambda U_1 U_2 L_{\mathfrak{K}}}{\rho cm} \sqrt{\sum_{s,j} \frac{1}{\bar{n}}} + \frac{1}{cm} \sqrt{\sum_{s,j} \frac{\log \delta^{-1}}{2\bar{n}}} \\
= &\frac{4}{\rho} q\Lambda U_1 U_2 L_{\mathfrak{K}} \sqrt{\frac{c}{m\bar{n}}} + \sqrt{\frac{\log \delta^{-1}}{2cm\bar{n}}},
\end{aligned} \tag{C.40}$$

$$II(b) \leq \frac{4}{\rho} q\Lambda U_1 U_2 L_{\mathfrak{K}} \sqrt{\frac{c}{m}} + \sqrt{\frac{\log \delta^{-1}}{2cm}}, \tag{C.41}$$

where $\bar{k}(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{x}})$ is bounded via Lemma C.3.

Combining (C.39), (C.40) and (C.41), the multi-class generalization bound after considering the heterogeneity is

$$R(g) \leq \hat{R}_{n,\rho}(g) + \frac{1}{\rho} q\Lambda U_1 U_2 L_{\mathfrak{K}} \Big(6\sqrt{\frac{\log 2cm\delta^{-1}}{\bar{n}}} + 4\sqrt{\frac{c}{m\bar{n}}} + 4\sqrt{\frac{c}{m}}\Big) + \sqrt{\frac{\log \delta^{-1}}{2cm\bar{n}}} + \sqrt{\frac{\log \delta^{-1}}{2cm}}.$$

D. Experimental Configurations

The hyperparameters settings for the different methods are as follows:

- k-NN: the number of the nearest neighbours $k \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ were validated.
- SVM: the regularization coefficient $C \in \{0.1, 0.5, 1, 2, 5, 10, 20, 50\}$ and the kernel bandwidth $h \in \{0.1d_M, 0.5d_M, 1d_M, 5d_M, 10d_M, 50d_M, 100d_M\}$, where $d_M = \text{median}\left(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2\right), \forall \boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$ were validated.
- DICA: Two parameters (λ, δ) require tuning. $\lambda \in \left\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\right\}$ and $\delta \in \left\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4, 10^5, 10^6\right\}$ were validated.
- SCA: Two parameters (β, δ) require tuning. $\beta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $\delta \in \left\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4, 10^5, 10^6\right\}$ were validated.
- CIDG: Four hyper-parameters $(\beta, \delta, \sigma, \gamma)$ require tuning. $\beta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $\delta \in \left\{1, 10, 10^2, 10^3, 10^4, 10^5, 10^6\right\}$, $\sigma \in \left\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\right\}$ and $\gamma \in \left\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\right\}$ were validated.
- MDA: Three hyper-parameters (β, α, γ) require tuning. $\beta \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$, $\alpha \in \left\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4, 10^5, 10^6\right\}$, and $\gamma \in \left\{1, 10, 10^2, 10^3, 10^4, 10^5, 10^6\right\}$ were validated.
- DFDG: For 1-NN as classifier, only one hyper-parameter γ requires tuning. $\gamma \in \{0.01, 0.1, 0.5, 1, 2, 3, 5, 10, 20, 50, 100\}$ were validated.
  For SVM as classifier, two extra hyper-parameters were also considered, the kernel bandwidth h for transfer kernel and regularization coefficient C in SVM. $C \in \{0.1, 0.5, 1, 2, 5, 10, 20, 50\}$ and $h \in \{0.1d_M, 0.5d_M, 1d_M, 5d_M, 10d_M, 50d_M, 100d_M\}$, where $d_M = \text{median}\left(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2\right), \forall \boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$, were validated.

For kernel-based DG methods (DICA, SCA, CIDG, MDA, DFDG), different number of extracted features q was also validated. For synthetic data, $q \in \{2, 3, 4, 5\}$ were tested. For real data, different number of extracted features q (i.e., the number of leading eigenvectors) that contribute to certain proportions ({0.5, 0.8, 0.9, 0.95, 0.98}) of the sum of all eigenvalues were validated.

E. Sensitivity analysis

We have performed a sensitivity analysis to evaluate the impact on the hyperparameter on classification performance of the proposed method and the other kernel DG methods with results reported in Table S1. We have also conducted an ablation analysis with results in Tables S2-S3.

The sensitivity analysis used the same hyperparameter settings as those in the Section 5 of the manuscript. The results in Table S1 demonstrate that the proposed method exhibits a much less sensitivity in 3 of the 4 existing kernel methods. While having comparable sensitivity with the SCA method, the proposed methods have better maximum, minimum and median accuracy than those of the SCA method.

The ablation analysis was conducted by setting $\gamma = 0$ for the proposed DFDG with the results shown in Tables S2 and S3. There is a significant reduction in model performance in quite many cases in the synthetic experiments in Table S2. The most profound reduction in the classification accuracy happen for the real data case studies reported in Table S3, where we see that out of a total of 64 trials (16 tasks with 4 algorithms) in the case studies, the ablation test was found to perform worst in 43.5 trials, while it faired between the worst and median accuracy in 12.5 trials with only 3.5 trails being the best performer. These suggest the important role played by the hyperparameter.

It is also noted that the generalized eigenvalue problem requires the right matrix to be positive definite, and with high-dimensional RKHS spaces but only n observations, a γI term is necessary to ensure positive definiteness, where I is the identity matrix. Therefore, γ should not be omitted (set to zero) for numerical reasons too.

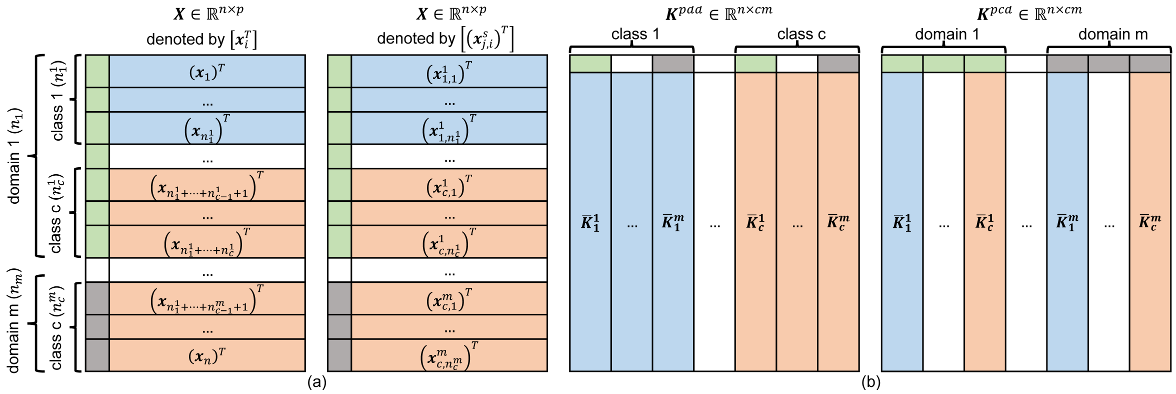

Figure S1. (a). Data $\boldsymbol{X}$ is organized by double indexing, where we first loop the domain index s and then the class index j. We use $\boldsymbol{X} = [\boldsymbol{x}_i^T]$ and $\boldsymbol{X} = [(\boldsymbol{x}_{j,i}^s)^T]$ interchangeably, while the latter one is used to emphasize the class and domain of $\boldsymbol{x}_i$; (b). Both $\boldsymbol{K}^{pcd}$ and $\boldsymbol{K}^{pdd}$ has the same row order as what in $\boldsymbol{X}$, while they differ only in the order of columns.

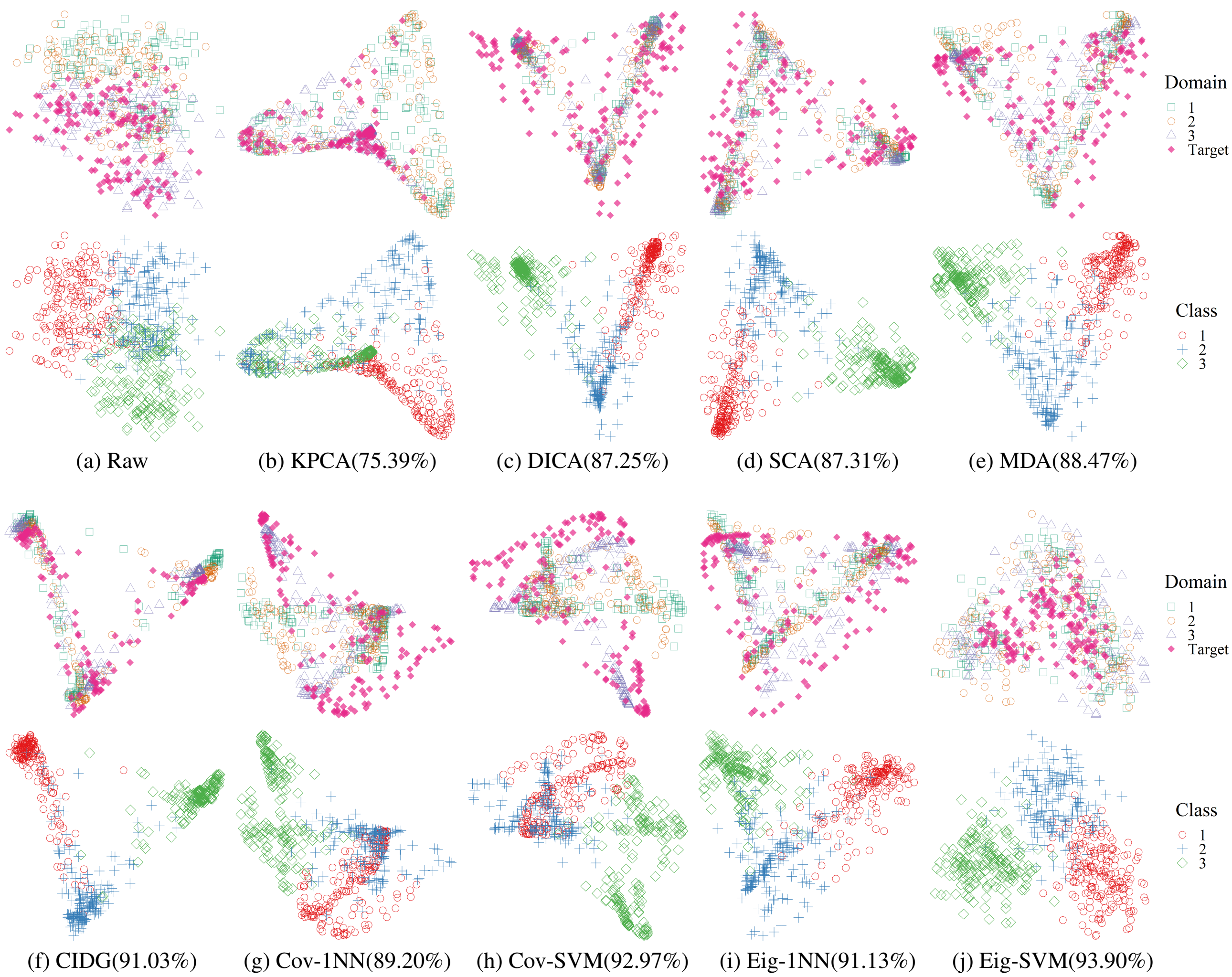

Figure S2. Extracted features of synthetic data corresponding to the first two eigenvectors of different methods on Case 1. The numbers in brackets show the accuracy of each method on the target domain. The top and bottom rows show the domains and classes of data respectively. Only 50 of the 200 samples were displayed for each class of each domain.

Table S1. Maximum, median and minimum accuracy with respective to the range of parameters. The sensitivity is defined as the difference between the maximum and minimum accuracy.

method		parameter	max	median	min	sensitivity
DFDG-Eig	SVM	γ	94.00	92.67	88.50	5.50
DFDG-Eig	1NN	γ	91.33	89.17	86.33	5.00
DFDG-Cov	SVM	γ	92.50	91.67	90.83	1.67
DFDG-Cov	1NN	γ	90.83	89.33	86.67	4.16
CIDG	1NN	β	90.17	89.83	89.67	0.50
CIDG	1NN	δ	90.33	88.83	85.33	5.00
CIDG	1NN	σ	90.50	88.67	31.67	58.83
CIDG	1NN	γ	90.33	89.67	87.83	2.50
CIDG	1NN	All	92.83	80.00	33.33	59.50
MDA	1NN	β	90.00	89.17	86.50	3.50
MDA	1NN	α	89.67	88.17	87.00	2.67
MDA	1NN	γ	89.17	88.25	86.50	2.67
MDA	1NN	All	90.00	87.83	72.83	17.17
DICA	1NN	λ	88.67	87.33	76.67	12.00
DICA	1NN	δ	88.67	86.75	81.17	7.50
DICA	1NN	All	89.17	88.17	78.33	10.84
SCA	1NN	β	89.50	87.67	86.83	2.67
SCA	1NN	δ	88.33	87.25	86.50	1.83
SCA	1NN	All	89.17	88.17	85.50	3.67

Table S2. Sensitivity and ablation analyses of the proposed methods in the synthetic experiments.

		case1	case2	case3	case4	case5	case6
DFDG-Eig SVM	max	94.00	91.50	93.17	92.83	88.50	83.50
	median	92.67	86.83	91.50	92.00	86.33	79.17
	min	88.50	82.83	89.50	86.17	78.50	77.00
	ablation	94.17	88.00	88.83	91.67	85.17	71.50
DFDG-Eig 1NN	max	91.33	90.83	89.17	90.00	83.67	76.33
	median	89.17	88.33	87.17	88.83	82.67	75.50
	min	86.33	82.33	83.00	87.67	81.50	75.00
	ablation	90.33	88.83	82.50	85.83	84.00	72.00
DFDG-Cov SVM	max	92.50	93.17	92.33	92.83	88.17	78.33
	median	91.67	92.33	91.50	92.50	86.00	76.00
	min	90.83	87.67	84.17	90.67	83.67	71.00
	ablation	87.17	90.17	85.67	92.83	82.67	80.33
DFDG-Cov 1NN	max	90.83	91.00	89.00	90.00	83.50	76.00
	median	89.33	89.33	85.50	87.83	82.17	75.17
	min	86.67	83.33	83.17	84.83	79.00	71.17
	ablation	84.17	86.50	80.50	86.50	69.00	70.17

Table S3. Sensitivity and ablation analyses of the proposed methods in the case studies (Office&Caltech and VLCS datasets)

		Office+Caltech						**VLCS**									
Source		C,D,W	A,D,W	D,W	A,C	A,D	A,W	L,C,S	V,C,S	V,L,S	V,L,C	C,S	L,S	L,C	V,S	V,C	V,L
Target		A	C	A,C	W,D	W,C	D,C	V	L	C	S	V,L	V,C	V,S	L,C	L,S	C,S
DFDG-Eig SVM	max	92.8	84.1	71.6	82.1	84.9	84.8	63.2	58.9	92.5	65.1	58.9	63.2	58.8	70.8	65.2	70.7
	median	92.1	82.6	70.5	80.5	83.9	83.9	60.2	55.5	87.4	64.5	57.5	61.1	56.2	69.4	64.1	69.6
	min	90.9	81.4	67.5	79.0	81.2	83.0	57.2	51.6	80.8	63.2	56.2	56.1	51.1	66.6	62.6	66.5
	ablation	88.8	81.5	61.8	78.8	80.9	83.3	63.2	59.0	80.8	63.1	57.4	60.4	52.4	70.4	61.0	70.8
DFDG-Eig 1NN	max	92.2	83.4	68.4	84.1	83.4	85.2	61.6	57.2	91.5	64.2	57.8	64.0	53.0	69.4	62.5	69.6
	median	92.1	83.0	65.3	82.3	82.5	84.6	59.8	55.0	90.1	62.7	55.2	62.1	50.0	68.2	62.0	66.8
	min	91.4	80.8	64.2	81.9	81.7	83.8	58.4	53.4	87.6	61.9	54.8	58.4	48.2	67.2	61.3	63.6
	ablation	89.9	79.3	61.1	80.8	80.3	83.1	58.3	54.9	80.5	59.3	53.7	57.8	44.4	67.4	57.2	66.6
DFDG-Cov SVM	max	93.1	84.5	72.2	81.9	86.0	86.2	65.3	59.3	91.9	66.4	57.9	66.1	56.7	70.8	64.8	73.7
	median	91.8	83.3	70.8	80.5	84.4	84.9	62.2	58.4	89.1	64.3	56.1	62.9	55.7	69.3	63.9	72.1
	min	90.9	82.1	67.1	79.6	81.6	84.2	60.7	57.0	86.4	63.4	55.2	61.6	53.5	66.5	62.9	70.2
	ablation	90.8	80.1	66.7	79.0	80.8	83.6	62.1	61.7	86.0	61.2	53.1	64.3	52.4	68.5	63.4	65.4
DFDG-Cov 1NN	max	92.2	84.4	71.6	80.3	84.1	86.6	61.1	56.2	93.4	63.9	56.3	60.1	51.3	68.7	63.1	71.2
	median	91.1	84.0	69.7	79.8	82.4	85.1	59.4	55.3	89.3	63.2	55.3	57.0	50.5	67.2	62.5	70.0
	min	90.8	82.7	65.7	79.0	81.4	83.4	58.0	54.4	83.6	62.7	54.0	53.0	47.9	66.2	60.5	68.4
	ablation	85.5	77.6	57.9	76.3	78.6	79.7	56.2	55.9	82.5	59.3	53.9	56.9	44.2	67.2	55.3	65.6

Table S4. The average rank (Rank) of different methods for all 16 missions on Office+Caltech and VLCS dataset and the results of Nemenyi's Paired Test, which is used to analyze whether the performances of the methods are statistically different, where the p-value that smaller than 0.05 is highlighted.

Rank	**Methods**		DFDG-Eig SVM	SVM	SCA 1-NN	DFDG-Eig 1-NN	DFDG-Cov 1-NN	DICA 1-NN	MDA 1-NN	CIDG 1-NN	k-NN
2.38	DFDG-Cov	SVM	0.887	**9.5e-07**	**1.0e-05**	**6.3e-06**	**9.5e-10**	**5.6e-10**	**1.3e-13**	**1.1e-13**	**<2e-16**
2.94	DFDG-Eig	SVM	-	**0.002**	**0.010**	**0.007**	**9.7e-06**	**6.3e-06**	**1.1e-09**	**6.0e-13**	**<2e-16**
4.25	SVM			-	1.000	1.000	0.985	0.976	0.225	**0.011**	**8.8e-14**
4.69	SCA	1-NN			-	1.000	0.881	0.846	0.077	**0.002**	**9.1e-14**
5.06	DFDG-Eig	1-NN				-	0.916	0.887	0.099	**0.003**	**9.5e-14**
6	DFDG-Cov	1-NN					-	1.000	0.897	0.251	**1.3e-13**
6	DICA	1-NN						-	0.925	0.294	**6.6e-14**
6.75	MDA	1-NN							-	0.990	**1.3e-12**
6.94	CIDG	1-NN								-	**2.3e-09**
10	k-NN										-

Table S5. Average accuracy and standard deviation on Office+Caltech dataset.

		Office+Caltech					
Target		A	C	A,C	W,D	W,C	D,C
k-NN		79.7±0.78	68.6±0.00	48.8±0.00	61.2±1.75	71.5±0.00	70.6±0.66
SVM		92.2±0.09	82.8±0.42	68.7±0.09	80.5±0.19	84.9±0.42	84.4±0.08
DICA	1-NN	91.8±0.77	83.2±2.26	61.7±7.10	80.2±0.78	84.9±2.32	85.4±2.38
SCA	1-NN	92.2±0.78	82.3±1.76	65.0±2.73	81.2±0.00	85.2±1.12	83.8±2.17
MDA	1-NN	90.3±1.21	75.1±1.30	56.7±2.92	75.9±0.40	80.9±2.16	78.5±1.68
CIDG	1-NN	92.5±0.69	82.4±0.44	68.6±3.45	79.5±0.90	82.0±2.59	83.4±0.42
DFDG-Eig	SVM	92.3±0.39	83.2±0.49	72.3±1.41	81.2±1.77	83.8±0.65	85.0±0.83
DFDG-Eig	1-NN	91.9±0.48	82.6±0.40	66.2±1.24	82.7±0.55	82.3±0.48	84.9±0.13
DFDG-Cov	SVM	92.5±0.67	83.9±0.72	73.1±0.87	81.6±0.88	83.8±0.88	84.9±1.06
DFDG-Cov	1-NN	90.5±0.75	82.3±0.44	68.2±0.15	81.2±0.40	81.5±0.66	84.3±0.79

Table S6. Average accuracy and standard deviation on VLCS dataset.

Target		VLCS									
		V	L	C	S	V,L	V,C	V,S	L,C	L,S	C,S
k-NN		46.8±0.20	49.5±0.67	72.9±1.25	48.9±0.96	52.5±0.21	50.7±0.54	42.1±0.64	57.5±0.21	49.6±0.32	56.3±1.15
SVM		64.7±0.99	58.6±2.02	84.9±3.27	63.9±0.85	59.5±1.40	63.3±0.80	53.6±0.38	66.8±1.20	64.9±1.31	70.3±0.79
DICA	1-NN	61.7±0.98	56.8±0.91	87.5±1.60	58.7±1.07	57.3±1.32	55.1±1.59	53.7±0.83	68.8±0.63	60.0±0.51	70.0±0.25
SCA	1-NN	65.3±0.37	58.0±0.97	89.4±2.21	60.7±0.39	58.4±1.31	56.8±1.37	54.8±0.24	69.8±0.48	61.1±0.73	70.9±0.24
MDA	1-NN	64.4±0.20	57.8±0.67	90.1±1.25	61.0±0.96	57.1±0.21	61.6±0.54	54.4±0.64	70.6±0.91	59.1±0.32	69.3±1.15
CIDG	1-NN	59.6±1.84	55.3±1.49	88.9±2.21	59.5±1.07	56.4±1.42	56.7±1.98	52.0±0.94	68.7±1.08	58.3±1.54	70.4±1.48
DFDG-Eig	SVM	60.8±1.30	58.4±1.10	90.2±1.14	66.2±0.73	58.4±0.32	64.2±1.60	56.4±2.14	70.8±0.60	63.4±0.99	71.2±0.52
DFDG-Eig	1-NN	61.4±0.74	57.2±1.01	91.6±1.70	64.5±0.26	57.0±0.78	63.8±0.57	51.2±1.28	68.8±0.60	63.7±0.45	68.9±0.84
DFDG-Cov	SVM	64.6±0.69	59.5±0.90	91.4±1.12	65.0±0.42	57.6±0.61	63.4±0.89	56.5±1.57	70.2±1.02	64.5±0.54	72.4±0.52
DFDG-Cov	1-NN	62.6±0.66	56.0±0.97	93.0±0.95	62.9±1.02	56.1±0.24	62.0±0.96	51.5±1.28	68.3±0.79	61.6±1.13	72.0±0.72

Table S7. Average accuracy and standard deviation on Terra Incognita dataset.

method		L100	L38	L43	L46
ERM baseline		53.1	41.1	54.7	36.1
DICA	1-NN	43.8±2.73	32.8±2.00	48.9±0.97	32.5±0.86
SCA	1-NN	44.6±1.68	39.2±0.93	49.0±1.13	30.1±0.90
MDA	1-NN	39.7±1.02	35.4±1.45	47.8±2.04	26.0±1.14
CIDG	1-NN	45.9±0.31	38.0±1.00	50.4±0.83	33.8±1.12
DFDG-Eig	SVM	55.3±2.41	42.7±0.98	56.6±1.01	38.3±1.03
DFDG-Eig	1-NN	53.5±0.65	41.6±0.81	55.7±0.79	36.9±0.80
DFDG-Cov	SVM	55.4±1.05	41.6±2.11	55.9±0.37	37.7±0.67
DFDG-Cov	1-NN	53.7±1.47	41.6±1.85	55.0±0.28	38.4±1.21

References

Blanchard, G., Lee, G., and Scott, C. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Chen, S. X. and Qin, Y.-L. A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.*, 38(2):808–835, 2010. ISSN 0090-5364.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. MIT Press, 2018. ISBN 9780262039406.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. G. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(7):2389–2410, 2011. ISSN 1532-4435. 805vu Times Cited:137 Cited References Count:32.

Yan, J. and Zhang, X. Kernel two-sample tests in high dimensions: interplay between moment discrepancy and dimension-and-sample orders. *Biometrika*, 2022. ISSN 1464-3510.