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]

Abstract

Accurate prediction of the out-of-distribution data is desired for a learning algorithm. In domain generalization, training data from source domains tend to have different distributions from that of the target domain, while the target data are absence in the training process. We propose a Distribution Free Domain Generalization (DFDG) procedure for classification by conducting standardization to avoid the dominance of a few domains in the training process. The essence of the DFDG is its reformulating the cross domain/class discrepancy by pairwise two sample test statistics, and equally weights their importance or the covariance structures to avoid dominant domain/class. A theoretical generalization bound is established for the multi-class classification problem. The DFDG is shown to offer a superior performance in empirical studies with fewer hyperparameters, which means faster and easier implementation.

1. Introduction

Domain generalization (DG) aims at transferring knowledge from the source domains to the target domains without the target data in the training process (Blanchard et al., 2011). A major challenge of DG is that the source and target data are not identically distributed. An algorithm trained from the source domains tends to be less performing in the target domain. DG is designed to attain robust performance in the target domain.

Compared with the domain adaptation where the target data are accessible in training to obtain a target specific predictor (Long et al., 2015; Li et al., 2021), DG is designed for a single global predictor or classifier that performs well in both the source and target domains (Blanchard et al., 2021). Studies have been proposed for the DG (Zhou et al., 2021; Fan et al., 2021; Shu et al., 2021), such as the kernel based domain invariant feature representation (Hu et al., 2020), the meta learning framework (Balaji et al., 2018) and the model selection or model average (Ye et al., 2021). See Wang et al. (2022) and Zhou et al. (2022) for a review.

Among the existed DG methods, we follow the kernel DG methods (Muandet et al., 2013; Ghifary et al., 2017; Li et al., 2018; Hu et al., 2020) for new development. These methods first map data to a high dimensional reproducing kernel Hilbert space (RKHS), and then construct metrics to measure the cross domain and class discrepancy, followed by a low dimensional feature representation that minimizes the cross domain dissimilarity while keeping new features with different classes well separated. The metrics are usually constructed as variants of the maximum mean discrepancy (MMD) (Gretton et al., 2012).

A common challenge with the DG is to counter the different mean levels and the variation among the discrepancy measures of different domains in the training stage. A robust DG procedure has to avoid domains with higher mean levels or variations to dictate the feature selection as features much influenced by the outlaying domains are doom to be weak in domain generalization. Existing kernel DG methods have to use more hyperparameters to balance the between-domain discrepancy measures, which may reduce the generalization ability of the methods.

We propose two standardization procedures which are designed to reduce the heterogeneity in the kernel DG discrepancy statistics among the domains by conducting mean and variance adjustments. These standardizations are based on asymptotic analysis ((12) and Proposition 1) on the pairwise MMD statistics, which reduces the number of hyperparameters and speeds up the training process, and hence allows more computation intensive classifier in the DG procedure.

Specifically, we put forward a distribution-free DG (DFDG) approach that provides a superior performance using fewer hyperparameters, which is well suited for DG. We unify the kernel DG methods as an optimization problem based on pairwise two-sample test statistics with concise matrix form in terms of the sandwich structure. Two distribution-free standardized metrics are proposed, one reweights the

[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. Copyright 2023 by the author(s).

weighting matrix by the means of the null distributions, and the other de-correlates the averaged Gram matrix. A generalization bound for the multi-class classification based on the DFDG is derived, which provides theoretical guarantee for the proposed DFDG approach.

The paper is organized as follows. Section 2 gives the unified framework of the DG problem for classification. Section 3 proposes two distribution free metrics. Section 4 is for the generalization bound. Simulation and case studies are provided in Section 5, followed by a conclusion in Section 6. Some technical and numerical details are relegated to the supplementary material (SM).

2. Unified framework of DG problem

Throughout the paper, we use bold lowercase letters for column vectors, and bold uppercase letters for matrices.

We consider a classification task. Let $\mathcal{X} \subset \mathbb{R}^p$ denote the observation space and $\mathcal{Y} \subset \mathbb{R}$ be the set of class labels. Let $\mathfrak{P}_{\mathcal{X}\times\mathcal{Y}}$ denote the set of joint distributions on $\mathcal{X} \times \mathcal{Y}$. It is assumed that there exists a unimodal super distribution $\mathscr{P}$ with finite variance over $\mathfrak{P}_{\mathcal{X}\times\mathcal{Y}}$, such that $P_{XY}^{(1)}, \dots, P_{XY}^{(m)}$ are independent and identically distributed (IID) realizations from $\mathscr{P}$ in $\mathfrak{P}_{\mathcal{X}\times\mathcal{Y}}$. For a domain s, there is a sample $\{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ of n_s IID realizations of $(\boldsymbol{x}, y)$ according to the distribution $P_{XY}^{(s)}$. In general, for any $s \neq s'$, $P_{XY}^{(s)} \neq P_{XY}^{(s')}$, implying no-identical distribution cross the domains.

Consider a target distribution $P_{XY}^{(t)} \sim \mathscr{P}$ and target sample $\{(\boldsymbol{x}_i^t, y_i^t)\}_{i=1}^{n_t}$, where the class labels $\{y_i^t\}$ are not available, and $\{\boldsymbol{x}_i^t\}$ are not used in the training. This forces us to establish a global model without retraining the model for a specific target domain. Our goal is to extract domain-invariant features that have minimum cross domain discrepancy and maximum cross class discrepancy simultaneously.

The kernel method is founded on a RKHS $\mathcal{H}$ associated with a kernel k and inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ having the reproducing property that for any function $f : \mathcal{X} \to \mathbb{R}$, $\langle f(\cdot), k(\boldsymbol{x}, \cdot) \rangle_{\mathcal{H}} = f(\boldsymbol{x})$. The canonical map $\phi(\boldsymbol{x}) : \mathcal{X} \to \mathcal{H}$ can be denoted as $\phi(\boldsymbol{x}) := k(\boldsymbol{x}, \cdot)$ satisfying $k(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^T \phi(\boldsymbol{x}')$.

To map a probability distribution to the RKHS, we define the kernel mean embedding $\boldsymbol{\mu} : \mathfrak{P}_{\mathcal{X}} \to \mathcal{H}$ induced by k

$$\boldsymbol{\mu}_{P_X} := E_X[\phi(X)] = \int_{\mathcal{X}} \phi(\boldsymbol{x}) dP_X.$$

If k is a bounded and characteristic kernel, the mapping is injective so that $||\boldsymbol{\mu}_{P_X} - \boldsymbol{\mu}_{P_X'}||_{\mathcal{H}} = 0$ if and only if (iff) $P_X = P_X'$. The sample estimator $\hat{\boldsymbol{\mu}}_{P_X} = \frac{1}{n} \sum_{i=1}^n \phi(\boldsymbol{x}_i)$.

Denote the kernel mean embedding of $P_X^{(s)}$and $P_{X|Y=j}^{(s)}$ by $\boldsymbol{\mu}^s$ and $\boldsymbol{\mu}_j^s$, respectively. These mean maps are all high dimensional and we assume that $\boldsymbol{\mu}_P \in \mathbb{R}^N$ for a large integer N, where N can be infinity.

2.1. Cross domain discrepancy

The cross domain discrepancy can be regarded as the sum of pairwise distances at each domain condition over every class, as follows.

Definition 1 (pairwise cross domain discrepancy (PDD))**.** Given the class-conditional distributions $\{P_{X|Y=j}^{(s)}\}$ for $s \in \{1, \dots, m\}$ and $j \in \{1, \dots, c\}$, the PDD

$$\Psi^{pdd} := \frac{1}{c\binom{m}{2}} \sum_{j=1}^{c} \sum_{1 \le s < s' \le m} ||\boldsymbol{\mu}_j^s - \boldsymbol{\mu}_j^{s'}||_{\mathcal{H}}^2, \tag{1}$$

where $\binom{m}{2}$ is the number of combination.

Each term in (1) is a squared MMD , which describes the distance between two distributions. It is also similar to the traditional Hotelling's T-test but without weighting via a covariance matrix.

To reformulate (1) as a concise matrix form, for a class j, denote $\boldsymbol{M}_j = [\boldsymbol{\mu}_j^1, \dots, \boldsymbol{\mu}_j^m] \in \mathbb{R}^{N\times m}$ and $\boldsymbol{\Gamma}_1 = m\boldsymbol{I}_m - \boldsymbol{1}_m \boldsymbol{1}_m^T$, we have $\Psi^{pdd} = c^{-1} \binom{m}{2}^{-1} \sum_{j=1}^c \mathrm{tr}(\boldsymbol{M}_j \boldsymbol{\Gamma}_1 \boldsymbol{M}_j^T)$, where $\boldsymbol{I}_m$ is a $m \times m$ identity matrix and $\boldsymbol{1}_m$ is a vector in $\mathbb{R}^m$ whose elements are all ones. Moreover, denote $\boldsymbol{M} = [\boldsymbol{M}_1, \dots, \boldsymbol{M}_c]$ and let $\boldsymbol{\Gamma}^{pdd} = \boldsymbol{I}_c \otimes \boldsymbol{\Gamma}_1$ where "$\otimes$" denotes the Kronecker product. Then, it is readily shown that

$$\Psi^{pdd} = c^{-1} \binom{m}{2}^{-1} \mathrm{tr}(\boldsymbol{M} \boldsymbol{\Gamma}^{pdd} \boldsymbol{M}^T). \tag{2}$$

The above formulation introduces a matrix sandwich form with the block diagonal $\boldsymbol{\Gamma}^{pdd}$ as the weighting matrix.

2.2. Cross class discrepancy

While the PDD metric (1) has been considered by Li et al. (2018) and Hu et al. (2020), to measure the class dissimilarity, now we propose a cross class discrepancy measure.

Definition 2 (pairwise cross class discrepancy (PCD))**.** Domain specified cross class discrepancy is defined as

$$\Psi^{pcd} := \frac{1}{m\binom{c}{2}} \sum_{s=1}^{m} \sum_{1 \le j < j' \le c} ||\boldsymbol{\mu}_j^s - \boldsymbol{\mu}_{j'}^{s}||_{\mathcal{H}}^2, \tag{3}$$

the average class dissimilarity among the domains.

Compared with Ψ^{pdd}, Ψ^{pcd} exchanges the order of the domain and class indexes. Let $\boldsymbol{U}_s = [\boldsymbol{\mu}_1^s, \dots, \boldsymbol{\mu}_c^s] \in \mathbb{R}^{N\times c}$, and $\boldsymbol{\Gamma}_2 = c\boldsymbol{I}_c - \boldsymbol{1}_c \boldsymbol{1}_c^T$, (3) becomes

$$\Psi^{pcd} = m^{-1} \binom{c}{2}^{-1} \mathrm{tr}(\boldsymbol{U} \boldsymbol{\Gamma}^{pcd} \boldsymbol{U}^T), \tag{4}$$

where $\boldsymbol{U} = [\boldsymbol{U}_1, \dots, \boldsymbol{U}_m]$ and $\boldsymbol{\Gamma}^{pcd} = \boldsymbol{I}_m \otimes \boldsymbol{\Gamma}_2$.

It is noted that the existing kernel DG methods, e.g. in Ghifary et al. (2017) and Hu et al. (2020), had two metrics for the domain discrepancy and two metrics for the class mismatch, respectively. Our proposal has only one metric for each measure, which means fewer hyperparameters than the existing DG methods, which leads to more efficient DG methods as shown later.

2.3. Feasible optimization framework

The metrics Ψ^{pdd} and Ψ^{pcd} depend on the distribution P_{XY} and the kernel k. The goal of the DG is to find a q-dimensional invariant feature via a transformation $\boldsymbol{W} \in \mathbb{R}^{N\times q} : \mathcal{H} \to \mathbb{R}^q$, such that the new features $\boldsymbol{M}^T\boldsymbol{W}$ and $\boldsymbol{U}^T\boldsymbol{W}$ have minimum cross domain discrepancy and maximum cross class discrepancy simultaneously.

As the dimension of RKHS $\mathcal{H}$ is high for finite samples, we need a practically calculable form of $\boldsymbol{W}$. Let $n = n_1 + \cdots + n_m$ be the total training data size, and $\boldsymbol{\Phi} = [\phi(\boldsymbol{x}_1) \dots, \phi(\boldsymbol{x}_n)]^T \in \mathbb{R}^{n\times N}$ be the high dimensional feature matrix. Since $\boldsymbol{\Phi}$ has rank no more than n, $\boldsymbol{W}$ can be expressed as a linear combination of $\boldsymbol{\Phi}$ such as

$$\boldsymbol{W} = \boldsymbol{\Phi}^T\boldsymbol{B}, \tag{5}$$

where $\boldsymbol{B} \in \mathbb{R}^{n\times q}$ is a feasible mixing matrix.

We construct PDD and PCD metrics based on lower dimensional features $\boldsymbol{M}^T\boldsymbol{W}$ and $\boldsymbol{U}^T\boldsymbol{W}$. By (2), for the low dimensional feature $\boldsymbol{M}^T\boldsymbol{W}$,

$$\begin{aligned}\Psi^{pdd} =& c^{-1}\binom{m}{2}^{-1}\text{tr}(\boldsymbol{W}^T\boldsymbol{M}\boldsymbol{\Gamma}^{pdd}\boldsymbol{M}^T\boldsymbol{W})\\ =& c^{-1}\binom{m}{2}^{-1}\text{tr}(\boldsymbol{B}^T\boldsymbol{\Phi}\boldsymbol{M}\boldsymbol{\Gamma}^{pdd}\boldsymbol{M}^T\boldsymbol{\Phi}^T\boldsymbol{B})\\ :=& \text{tr}(\boldsymbol{B}^T\boldsymbol{Q}\boldsymbol{B}^T),\end{aligned}$$

where

$$\boldsymbol{Q} = c^{-1}\binom{m}{2}^{-1}\boldsymbol{K}^{pdd}\boldsymbol{\Gamma}^{pdd}\boldsymbol{K}^{pdd^T} \text{ and } \boldsymbol{K}^{pdd} := \boldsymbol{\Phi}\boldsymbol{M}. \tag{6}$$

Similarly, we can update $\Psi^{pcd} = \text{tr}(\boldsymbol{B}^T\boldsymbol{F}\boldsymbol{B}^T)$ with

$$\boldsymbol{F} = m^{-1}\binom{c}{2}^{-1}\boldsymbol{K}^{pcd}\boldsymbol{\Gamma}^{pcd}\boldsymbol{K}^{pcd^T} \text{ and } \boldsymbol{K}^{pcd} := \boldsymbol{\Phi}\boldsymbol{U}. \tag{7}$$

We note that both the $N \times cm$ matrices $\boldsymbol{M}$ and $\boldsymbol{U}$ are consisted with column vectors $\boldsymbol{\mu}_j^s$. After applying the kernel trick, the column vectors of $\boldsymbol{K}^{pdd}$ and $\boldsymbol{K}^{pcd}$ have the form

$$\bar{\boldsymbol{K}}_j^s := \boldsymbol{\Phi}\boldsymbol{\mu}_j^s = \frac{1}{n_j^s}\sum_{i=1}^{n_j^s}\boldsymbol{\Phi}\phi(\boldsymbol{x}_{j,i}^s) = \frac{1}{n_j^s}\sum_{i=1}^{n_j^s}\boldsymbol{K}_{j,i}^s,$$

where $\boldsymbol{K}_{j,i}^s = (k(\boldsymbol{x}_1, \boldsymbol{x}_{j,i}^s), \dots, k(\boldsymbol{x}_n, \boldsymbol{x}_{j,i}^s))^T \in \mathbb{R}^n$ is the i-th column vector of the Gram matrix

$$\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T = [k(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{ij} \in \mathbb{R}^{n\times n}, \tag{8}$$

and we use $\boldsymbol{x}_{j,i}^s$ to emphasize the corresponding class and domain indexes of $\boldsymbol{x}_i$. Since $\boldsymbol{M}$ and $\boldsymbol{Q}$ only differ from the order of columns, $\boldsymbol{K}^{pdd}$ and $\boldsymbol{K}^{pcd}$ share the same property, see Figure S1 for a graphical illustration.

As one cannot optimize $\Psi^{pcd} = \text{tr}(\boldsymbol{B}^T\boldsymbol{F}\boldsymbol{B})$ and $\Psi^{pdd} = \text{tr}(\boldsymbol{B}^T\boldsymbol{Q}\boldsymbol{B})$ simultaneously, we maximize Ψ^{pcd} while keeping Ψ^{pdd} fixed, namely

$$\arg\max_{\boldsymbol{B}} \text{tr}(\boldsymbol{B}^T\boldsymbol{F}\boldsymbol{B}) \text{ s.t. } \text{tr}(\boldsymbol{B}^T(\boldsymbol{Q}+\gamma\boldsymbol{K})\boldsymbol{B}) = 1, \tag{9}$$

where the first term in the trace constraint is to limit Ψ^{pdd} while the second term $\boldsymbol{B}^T\boldsymbol{K}\boldsymbol{B} = \boldsymbol{B}^T\boldsymbol{\Phi}\boldsymbol{\Phi}^T\boldsymbol{B} = \boldsymbol{W}^T\boldsymbol{W}$ is to control the magnitude of $\boldsymbol{W}$. The first order condition yields the generalized eigenvalue problem

$$\boldsymbol{F}\boldsymbol{B} = (\boldsymbol{Q}+\gamma\boldsymbol{K})\boldsymbol{B}\boldsymbol{\Gamma},$$

where $\boldsymbol{\Gamma} = \text{diag}(\lambda_1, \dots, \lambda_q)$ is a diagonal matrix collecting q leading eigenvalues, $\boldsymbol{B}$ is the corresponding eigenvectors. In practice, one may add $\varepsilon\boldsymbol{I}$ with $\varepsilon = 10^{-5}$ for numerically stable performance so that

$$\boldsymbol{F}\boldsymbol{B} = (\boldsymbol{Q}+\gamma\boldsymbol{K}+\varepsilon\boldsymbol{I})\boldsymbol{B}\boldsymbol{\Gamma}. \tag{10}$$

Compared with the existing kernel DG methods which involve more metrics and more hyperparameters as shown in Table 2, there is only one hyperparameter γ in (10).

3. Distribution free metrics

In this section, we propose two sets of empirical estimates of matrices $\boldsymbol{F}$ and $\boldsymbol{Q}$ used in the generalized eigenvalue problem (10) based on two ways of standardization that reweight the domain/class discrepancy measures. The first approach mainly focuses on the first moment difference of the features, while the second approach focuses on the second moment and adjusts for the empirical covariance matrix of the averaged Gram matrix.

3.1. Eigenvalue adjustment

The first set of $\boldsymbol{F}$ and $\boldsymbol{Q}$ estimates comes from standardizing the maximum mean discrepancy (MMD) statistic (Gretton et al., 2012) based on an asymptotic analysis. The MMD used in both (1) and (3) is a distance between two domain distributions,

$$\begin{aligned}&\text{MMD}^2(P_X^{(s)}, P_X^{(s')}) := \|\boldsymbol{\mu}^s - \boldsymbol{\mu}^{s'}\|_{\mathcal{H}}^2 = E_{x_i^s, x_j^s}[k(\boldsymbol{x}_i^s, \boldsymbol{x}_j^s)] -\\ &2E_{x_i^s, x_j^{s'}}[k(\boldsymbol{x}_i^s, \boldsymbol{x}_j^{s'})] + E_{x_i^{s'}, x_j^{s'}}[k(\boldsymbol{x}_i^{s'}, \boldsymbol{x}_j^{s'})].\end{aligned} \tag{11}$$

Estimates of MMD^2 can be made via the U- or the V-statistics. We consider the V-statistic formulation since it leads to positive semidefinite estimates for $\boldsymbol{F}$ and $\boldsymbol{Q}$:

$$\widehat{\text{MMD}}_b^2(P_X^{(s)}, P_X^{(s')}) = \frac{1}{(n_s)^2}\sum_{i,j=1}^{n_s} k(\boldsymbol{x}_i^s, \boldsymbol{x}_j^s) - \frac{2}{n_s n_{s'}}\sum_{i=1}^{n_s}\sum_{j=1}^{n_{s'}} k(\boldsymbol{x}_i^s, \boldsymbol{x}_j^{s'}) + \frac{1}{(n_{s'})^2}\sum_{i,j=1}^{n_{s'}} k(\boldsymbol{x}_i^{s'}, \boldsymbol{x}_j^{s'}).$$

Under the null hypothesis that $P_X^{(s)} = P_X^{(s')}$, the MMD statistic is equivalent to the one based on a centered kernel k' (Sejdinovic et al., 2013)

$$k'(\boldsymbol{x}_i, \boldsymbol{x}_j) = k(\boldsymbol{x}_i, \boldsymbol{x}_j) - E_x k(\boldsymbol{x}_i, \boldsymbol{x}) - E_x k(\boldsymbol{x}, \boldsymbol{x}_i) + E_{x,x'} k(\boldsymbol{x}, \boldsymbol{x}').$$

The null distribution of $\widehat{\text{MMD}}_b^2$ (Gretton et al., 2012) under $\lim_{n_s, n_{s'}\to\infty} \frac{n_s}{n_s + n_{s'}} = \rho^{s,s'}$ is

$$\frac{\widehat{\text{MMD}}_b^2(P_X^{(s)}, P_X^{(s)})}{n_s + n_{s'}} \xrightarrow{d} \frac{1}{\rho^{s,s'}(1-\rho^{s,s'})}\sum_{l=1}^{\infty} \lambda_l^s z_l^2, \quad (12)$$

where z_l^2 are IID χ_1^2 distributed, and $\{\lambda_l^s\}$ are the solutions to the eigenvalue equations

$$\int_X k'(\boldsymbol{x}, \boldsymbol{x}_j)\phi_l(\boldsymbol{x})dP_X^{(s)}(\boldsymbol{x}) = \lambda_l^s \phi_l(\boldsymbol{x}_j).$$

Note that the expectation of the limiting distribution in (12) is $\frac{1}{\rho^{s,s'}(1-\rho^{s,s'})}\sum_{l=1}^{\infty}\lambda_l^s$, which can be estimated by $\text{tr}(\boldsymbol{K}')$ (Shawe-Taylor et al., 2005) or the nuclear norm $||\boldsymbol{K}'||_*$, where $\boldsymbol{K}' = \boldsymbol{K} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\boldsymbol{K} - \frac{1}{n}\boldsymbol{K}\mathbf{1}_n\mathbf{1}_n^T + \frac{1}{n^2}\mathbf{1}_n\mathbf{1}_n^T\boldsymbol{K}\mathbf{1}_n\mathbf{1}_n^T$. This leads to an eigenvalue adjusted Ψ^{pdd} and Ψ^{pcd} by dividing each pair of MMD^2 by its expectation. Such that for each domain and class, the expectation of the scaled MMD^2 are asymptotically equal to one under the null hypothesis.

Definition 3 (scaled pairwise cross domain discrepancy (SPDD))**.** Given the set of class-conditional distributions $\{P_{X|Y=j}^{(s)}\}$, the empirical SPDD measure is

$$\hat{\Psi}^{spdd} := \frac{1}{c\binom{m}{2}}\sum_{j=1}^{c}\sum_{1\le s<s'\le m}\{\frac{n_j^s n_j^{s'}}{n_j^s + n_j^{s'}}||\boldsymbol{K}'^{j,s,s'}||_*^{-1}\times \widehat{\text{MMD}}_b^2(P_{X|Y=j}^{(s)}, P_{X|Y=j}^{(s')})\}. \quad (13)$$

Definition 4 (scaled pairwise cross class discrepancy (SPCD))**.** Given the set of domain-conditional distributions $\{P_{X|Y=j}^{(s)}\}$, the empirical SPCD metric

$$\hat{\Psi}^{spcd} := \frac{1}{m\binom{c}{2}}\sum_{s=1}^{m}\sum_{1\le j<j'\le c}\{\frac{n_j^s n_{j'}^s}{n_j^s + n_{j'}^s}||\boldsymbol{K}'^{s,j,j'}||_*^{-1}\times \widehat{\text{MMD}}_b^2(P_{X|Y=j}^{(s)}, P_{X|Y=j'}^{(s)})\}. \quad (14)$$

In (13) and (14), $\boldsymbol{K}'^{s,j,j'} \in \mathbb{R}^{n_j^s \times n_{j'}^s}$ is a submatrix of $\boldsymbol{K}'$, whose (i,l)-th element is $k'(\boldsymbol{x}_{j,i}^s, \boldsymbol{x}_{j',l}^s)$.

Mimic a similar dimension reduction as in Section 2.3, we work on the optimization problem (9) leading to the generalized eigenvalue problem (10) by replacing $\boldsymbol{F}$ and $\boldsymbol{Q}$ with their empirical estimates

$$\hat{\boldsymbol{F}} = \frac{1}{m\binom{c}{2}}\sum_{s=1}^{m}\sum_{1\le j<j'\le c}\{\frac{n_j^s n_{j'}^s}{n_j^s + n_{j'}^s}||\boldsymbol{K}'^{s,j,j'}||_*^{-1}\times (\bar{\boldsymbol{K}}'^s_j - \bar{\boldsymbol{K}}'^s_{j'})(\bar{\boldsymbol{K}}'^s_j - \bar{\boldsymbol{K}}'^s_{j'})^T\}, \quad (15)$$

$$\hat{\boldsymbol{Q}} = \frac{1}{c\binom{m}{2}}\sum_{j=1}^{c}\sum_{1\le s<s'\le m}\{\frac{n_j^s n_j^{s'}}{n_j^s + n_j^{s'}}||\boldsymbol{K}'^{j,s,s'}||_*^{-1}\times (\bar{\boldsymbol{K}}'^s_j - \bar{\boldsymbol{K}}'^{s'}_j)(\bar{\boldsymbol{K}}'^s_j - \bar{\boldsymbol{K}}'^{s'}_j)^T\}. \quad (16)$$

Solving (10) with the $\hat{\boldsymbol{F}}$ and $\hat{\boldsymbol{Q}}$ leads to the estimated eigenvectors $\hat{\boldsymbol{B}}$ whose i-th column $\hat{\boldsymbol{B}}_i \in \mathbb{R}^n$ associated with the nonzero eigenvalues, which needs to be standardized so that $||\hat{\boldsymbol{W}}_i||_{\mathcal{H}} = \hat{\boldsymbol{B}}_i^T\boldsymbol{K}\hat{\boldsymbol{B}}_i = 1$, which means we let

$$\hat{\boldsymbol{B}}_i \leftarrow \hat{\boldsymbol{B}}_i / \sqrt{\hat{\boldsymbol{B}}_i^T\boldsymbol{K}\hat{\boldsymbol{B}}_i}. \quad (17)$$

Compared with the existed kernel DG methods that standardizes $\hat{\boldsymbol{B}}$ by $\hat{\boldsymbol{B}}\hat{\boldsymbol{\Gamma}}^{-\frac{1}{2}}$ where $\hat{\boldsymbol{\Gamma}}$ is the estimated eigenvalue matrix in (10), (17) is more robust for a large feature dimension q by avoiding dividing near zero eigenvalues.

The scaling conducted in (13) and (14) is designed to remove the mean differences among the pairwise MMD-statistics by reweighting the statistics by their asymptotic means according to (12). The scaling allows the pairwise discrepancy measures between domains and classes being treated more equally. Thus, for the extracted invariant features in the DG, all the domains have a similar and balanced contribution, so that the features reflect the collective information of all participants, avoiding a few domains or classes dominate the selected features.

3.2. One side covariance filter

This subsection considers another standardization on the $\boldsymbol{F}$ and $\boldsymbol{Q}$ estimates in (6) and (7) by rotating $\boldsymbol{K}^{pcd}$ and $\boldsymbol{K}^{pdd}$ via their covariance matrices, namely

$$\tilde{\boldsymbol{K}}^{pcd} = \boldsymbol{K}^{pcd}(\tilde{\boldsymbol{\Gamma}}^{pcd})^{-\frac{1}{2}}, \quad \text{where} \quad (18)$$

$$\tilde{\boldsymbol{\Gamma}}^{pcd} = \frac{1}{cm}\sum_{s=1}^{m}\sum_{j=1}^{c}\frac{1}{n_j^s}\sum_{i=1}^{n_j^s}(\boldsymbol{K}_{i,j,s}^{pcd} - \bar{\boldsymbol{K}}_{j,s}^{pcd})(\boldsymbol{K}_{i,j,s}^{pcd} - \bar{\boldsymbol{K}}_{j,s}^{pcd})^T,$$

$$\bar{\boldsymbol{K}}_{j,s}^{pcd} = \frac{1}{n_j^s}\sum_{i=1}^{n_j^s}\boldsymbol{K}_{i,j,s}^{pcd}.$$

The standardization of $\boldsymbol{K}^{pdd}$, denoted by $\tilde{\boldsymbol{K}}^{pdd}$, can be obtained similarly by replacing $\boldsymbol{K}^{pcd}$ with $\boldsymbol{K}^{pdd}$ in the above formulation. In the above equations, $(\boldsymbol{K}^{pcd}_{i,j,s})^T \in \mathbb{R}^{1\times cm}$ is the i-th row vector of $\boldsymbol{K}^{pcd}$ corresponding to domain s and class j, whose explicit form is left in Supplementary Materials (SM).

One may wonder why we can treat $\boldsymbol{K}^{pcd}$ like the data matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T$, and the covariance matrix is calculated like $\boldsymbol{K}^{pcd}$ contains n independent observations whose underlying distribution are the same under the same domain and class. Proposition 1 shows that although there exist correlations among the rows and columns of $\boldsymbol{K}^{pcd}$, the column-wise covariances are of the order $O(1)$ and the row-wise covariances are $O((n_j^s)^{-1})$. The latter means that the column-wise covariances can be ignored in large samples. We use the averaged covariance matrix $\tilde{\boldsymbol{\Gamma}}^{pcd}$ in the Euclidean space for the standardization (rotation), and one may consider the averaged version in Riemannian space (Barachant et al., 2012).

We call the rotation in $\tilde{\boldsymbol{K}}^{pcd}$ and $\tilde{\boldsymbol{K}}^{pdd}$ the one side covariance filter and redefine $\boldsymbol{F}$ and $\boldsymbol{Q}$ in (10) by the rotated $\tilde{\boldsymbol{K}}^{pcd}$ and $\tilde{\boldsymbol{K}}^{pdd}$ as

$$\hat{\boldsymbol{F}} = m^{-1}\binom{c}{2}^{-1}\tilde{\boldsymbol{K}}^{pcd}\boldsymbol{\Gamma}^{pcd}\tilde{\boldsymbol{K}}^{pcd^T}, \tag{19}$$

$$\hat{\boldsymbol{Q}} = c^{-1}\binom{m}{2}^{-1}\tilde{\boldsymbol{K}}^{pdd}\boldsymbol{\Gamma}^{pdd}\tilde{\boldsymbol{K}}^{pdd^T}, \tag{20}$$

where $\boldsymbol{\Gamma}^{pcd}$ and $\boldsymbol{\Gamma}^{pdd}$ are similarly defined as those in (2) and (4). The algorithm is summarized in Algorithm 1.

The rest of the subsection provides the theoretical justification for the rotations, which is based on the correlation structures of $\boldsymbol{K}^{pcd}$ and $\boldsymbol{K}^{pdd}$. For notation simplicity, we first omit the class index j and consider a generic $\bar{\boldsymbol{K}} = [\frac{1}{n_{s'}}\sum_{l=1}^{n_{s'}} k_{il}^{ss'}]_{is'} \in \mathbb{R}^{n\times m}$ for either $\boldsymbol{K}^{pcd}$ and $\boldsymbol{K}^{pdd}$ for $i = 1, \ldots, n$ and $s' = 1, \ldots, m$, and $k_{ij}^{ss'} := k(\boldsymbol{x}_i^s, \boldsymbol{x}_j^{s'})$ is a simplified notation for elements of the Gram matrix $\boldsymbol{K}$. We note that $\bar{\boldsymbol{K}}$ can be obtained by merging $\boldsymbol{K}^{pdd}$ or $\boldsymbol{K}^{pcd}$ over the class lever. The results provided in Proposition 1 can be easily extended to cover both domain s and class j by merging s and j in a new defined single index $s^* : (s, j) \mapsto \{1, \ldots, cm\}$.

For a general kernel function $k(\boldsymbol{x}, \boldsymbol{y}) = f(||\boldsymbol{x} - \boldsymbol{y}||_2^2/h)$, where h is the bandwidth, we want to derive its first two moments. We begin with the following assumptions.

Assumption 1.

1. For each domain s, the covariates $\{\boldsymbol{x}_1^s, \ldots, \boldsymbol{x}_{n_s}^s\}$ are generated according to
$$\boldsymbol{x}_i^s = \boldsymbol{\Gamma}^s\boldsymbol{u}_i^s + \boldsymbol{\eta}^s, \quad i = 1, \ldots, n_s, \tag{21}$$
where $\boldsymbol{u}_i^s \in \mathbb{R}^{p'}$ are IID random variables satisfying $E(\boldsymbol{u}_i^s) = 0$, var$(\boldsymbol{u}_i^s) = \boldsymbol{I}_q$. For the j-th element $u_i^s(j)$ of u_i^s, $E(u_i^s(j)^8) < \infty$. The parameters $\boldsymbol{\Gamma}^s \in \mathbb{R}^{p\times p'}$, the mean $\boldsymbol{\eta}^s \in \mathbb{R}^p$, and $\boldsymbol{\Gamma}^s\boldsymbol{\Gamma}^{s^T} = \boldsymbol{\Sigma}^s$.
2. For each domain s, tr$(\boldsymbol{\Sigma}^s) = O(p)$, the operator norms of $\boldsymbol{\Sigma}^s$ are bounded from above and $||\boldsymbol{\eta}^s||_2^2 = O(p)$.
3. The domain sample sizes are balanced so that $\lim_{n\to\infty} n_s/n = \kappa_s \in (0, 1)$ where $n = \sum_{s=1}^m n_s$. Let $g(x) = f(x^2)$. Then, $g \in C^3[0, \infty)$ and $\sup_{1\le s\le 3}\sup_{x\ge 0} |g^{(s)}(x)| < \infty$.

Algorithm 1: Distribution free domain generalization

```
Input: Source data: $\{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{n_s}$, $s = 1, \ldots, m$;
       Hyperparameter $\gamma$ and kernel function $k(\cdot, \cdot)$;
       Number of subspace features $q$.
Output: Mixing matrix $\boldsymbol{B}$, Gram matrix $\boldsymbol{K}$, new
        features $\boldsymbol{Z}$ and $\boldsymbol{Z}^t$.
Calculate Gram matrix $\boldsymbol{K}$ via (8);
if use eigenvalue adjustment then
    Obtain the centered $\boldsymbol{K}' =$
      $\boldsymbol{K} - \frac{1}{n}\boldsymbol{1}_n\boldsymbol{1}_n^T\boldsymbol{K} - \frac{1}{n}\boldsymbol{K}\boldsymbol{1}_n\boldsymbol{1}_n^T + \frac{1}{n^2}\boldsymbol{1}_n\boldsymbol{1}_n^T\boldsymbol{K}\boldsymbol{1}_n\boldsymbol{1}_n^T$ ;
    Calculate $\hat{\boldsymbol{F}}$ and $\hat{\boldsymbol{Q}}$ via (15) and (16);
else if use one side covariance filter then
    Calculate $\hat{\boldsymbol{F}}$ and $\hat{\boldsymbol{Q}}$ via (19) and (20);
Solve eigenvalues $\boldsymbol{\Gamma}$ and corresponding eigenvectors $\boldsymbol{B}$
  from the generalized eigenvalue problem (10), select the
  $q$ leading eigenvectors;
Standardize $\boldsymbol{B}$ by letting $\boldsymbol{B}_i \leftarrow \boldsymbol{B}_i/\sqrt{\boldsymbol{B}_i^T\boldsymbol{K}\boldsymbol{B}_i}$;
Construct Gram matrix at test set as $[\boldsymbol{K}^t]_{ij} = k(\boldsymbol{x}_i^t, \boldsymbol{x}_j)$.
  The extracted features at training/testing set are
  $\boldsymbol{Z} = \boldsymbol{K}\boldsymbol{B}$ and $\boldsymbol{Z}^t = \boldsymbol{K}^t\boldsymbol{B}$, respectively.
```

The following proposition gives the covariance structures of $\bar{\boldsymbol{K}}$, whose proof follows the Taylor expansions in Yan & Zhang (2022) as showed in the SM.

Proposition 1. *Given Assumption 1,*

$$E\Big(\frac{1}{n_{s'}}\sum_{j=1}^{n_{s'}} k_{ij}^{ss'}\Big) = \mu^{(s,s')} + O(p^{3/2}h^{-3}),$$

$$var\Big(\frac{1}{n_{s'}}\sum_{j=1}^{n_{s'}} k_{ij}^{ss'}\Big) = \sigma^{(s,s')} + O(p^2h^{-4}),$$

where the specific forms of $\mu^{(s,s')}$ and $\sigma^{(s,s')}$ are given in (B.3) and (B.4) of the SM. For the covariances, if there is a common row

$$cov\Big(\frac{1}{n_{s'}}\sum_{j=1}^{n_{s'}} k_{ij}^{ss'}, \frac{1}{n_{s''}}\sum_{j=1}^{n_{s''}} k_{ij}^{ss''}\Big) = O(p^2h^{-2}), \tag{22}$$

and if they is a common column domain

$$cov\Big(\frac{1}{n_s}\sum_{j=1}^{n_s} k_{ij}^{s's}, \frac{1}{n_s}\sum_{j=1}^{n_s} k_{lj}^{s''s}\Big) = O(p^2n_s^{-1}h^{-2}) \tag{23}$$

and the covariance is 0 if there is no common row or column domain.

Proposition 1 suggests that the leading variance of $\frac{1}{n_{s'}}\sum_{l=1}^{n_{s'}} k_{il}^{ss'} = \frac{1}{n_{s'}}\sum_{l=1}^{n_{s'}} k(x_i^s, x_l^{s'})$ is $\sigma^{(s,s')}$, which is $O(p^2h^{-2})$ as shown in the SM. This is due to all the $n_{s'}$ terms in the summation have a common x_i^s, which leads to all $\mathrm{cov}(k_{il}^{ss'}, k_{il'}^{ss'})$ being a constant. If the two elements of $\bar{\boldsymbol{K}}$ are in the same row, their covariance is also $O(p^2h^{-2})$. But if they are in the same column, the covariance is $O(p^2n_s^{-1}h^{-2})$, which is a smaller order of $O(p^2h^{-2})$. Moreover, row vectors of $\bar{\boldsymbol{K}}$ belong to the same domain have the same mean and covariance structures. When the sample size goes to infinity, the correlation between different rows vanishes, and we treat them like independent variables. The results of Proposition 1 justifies the form of the $\tilde{\boldsymbol{\Gamma}}$ used in the rotation after (18).

Since the distribution of $\boldsymbol{X}$ and the kernel f in Assumption 1 are very general without much restriction, the one side covariance filter is generally applicable, for instance for non-Gaussian data and a general kernel.

4. Generalization bound

In this section, we analyze the generalization bound of the multi-class classification problem after applying the proposed DFDG algorithm (10). After providing the classifier, the loss function and the kernel. the generalization bound for the DFDG based classification is established.

The classifier f is of the form $f : \mathfrak{P}_{\mathcal{X}} \times \mathcal{X} \mapsto \mathbb{R}$. Let $\tilde{X} = (P_X, X)$ be the extended covariate. In the training procedure, one has n_s labeled data $\{(\hat{P}_X^{(s)}, \boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{n_s} = \{(\tilde{\boldsymbol{x}}_i^s, y_i^s)\}_{i=1}^{n_s} \in (\tilde{\mathcal{X}} \times \mathcal{Y})^{n_s}$ for each domain s, where $\mathcal{Y} = \{1, \ldots, c\}$ is the set of c classes. For the multi-class classification, a classifier f is defined via a scoring function $g : \tilde{\mathcal{X}} \times \mathcal{Y} \to \mathbb{R}$ as

$$f : \tilde{\boldsymbol{x}} \mapsto \arg\max_{y \in \mathcal{Y}} g(\tilde{\boldsymbol{x}}, y),$$

where we consider a linear scoring function g such that for class j, $g(\tilde{\boldsymbol{x}}, j) = \boldsymbol{a}_j^T \boldsymbol{W}^T \phi_k(\tilde{\boldsymbol{x}})$, and $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_c]^T \in \mathbb{R}^{c \times q}$ is bounded, namely $||\boldsymbol{A}|| \leq \Lambda$. Since $\boldsymbol{W}$ has been standardized to have column norm one,

$$||\boldsymbol{A}\boldsymbol{W}^T||_{\mathcal{H}_k} \leq ||\boldsymbol{A}|| \times ||\boldsymbol{W}||_{\mathcal{H}_k} \leq q\Lambda,$$

where q is the feature dimension of $\boldsymbol{W} = \boldsymbol{\Phi}^T\boldsymbol{B}$. We have reused the notation f and g, different from those in Assumption 1.

The loss function is established by the margin theory. A margin $r_g(\tilde{\boldsymbol{x}}, y)$ of the function g at a labeled observation $(\tilde{\boldsymbol{x}}, y)$ can be defined as

$$r_g(\tilde{\boldsymbol{x}}, y) = g(\tilde{\boldsymbol{x}}, y) - \max_{y' \neq y} g(\tilde{\boldsymbol{x}}, y').$$

Hence, f gives the wrong classification iff $r_g(\tilde{\boldsymbol{x}}, y) \leq 0$. The empirical ρ-margin loss given g and $\rho > 0$ is

$$\hat{R}_{n,\rho}(g) = \frac{1}{cm}\sum_{s=1}^{m}\sum_{j=1}^{c}\frac{1}{n_j^s}\sum_{i=1}^{n_j^s} l_\rho(r_g(\tilde{\boldsymbol{x}}_{j,i}^s, j)),$$

where $\tilde{\boldsymbol{x}}_{j,i}^s = (\hat{P}_{X|Y=j}^{(s)}, \boldsymbol{x}_{j,i}^s)$, $l_\rho(x) = \min(1, \max(0, 1 - x/\rho))$ is a ρ-margin loss function, ρ^{-1}-Lipschitz. The expected loss (risk) of the classification

$$R(g) = E_{(\tilde{x},y)} I(r_g(\tilde{\boldsymbol{x}}_i, y_i) \leq 0),$$

where $I(\cdot)$ is the indicator function. Since $I(x \leq 0) \leq l_\rho(x)$, the expected loss $R(g) \leq E_{(\tilde{x},y)}\hat{R}_{n,\rho}(g)$ for any g.

For the DG problem, the widely used product kernel $\bar{k}$ is

$$\bar{k}((P_X^{(s)}, \boldsymbol{x}_i^s), (P_X^{(s')}, \boldsymbol{x}_j^{s'})) = k_P(P_X^{(s)}, P_X^{(s')}) k_1(\boldsymbol{x}_i^s, \boldsymbol{x}_j^{s'}) \tag{24}$$

with a RKHS $\mathcal{H}_{\bar{k}}$ (Blanchard et al., 2011). For the choice of k_P, let k_2 denote a kernel on $\mathcal{X}$ with RKHS $\mathcal{H}_{k_2}$ and feature map ϕ_{k_2}, we define the k_2 induced kernel mean embedding $\mu : \mathfrak{P}_{\mathcal{X}} \to \mathcal{H}_{k_2}$ as $\boldsymbol{\mu}_{P_X} := \int_{\mathcal{X}} \phi_{k_2}(\boldsymbol{x})dP_X(\boldsymbol{x})$, and introduce another kernel $\mathfrak{K}$ on $\mathcal{H}_{k_2}$ such that

$$k_P\left(P_X^{(s)}, P_X^{(s')}\right) = \mathfrak{K}\left(\boldsymbol{\mu}_{P_X^{(s)}}, \boldsymbol{\mu}_{P_X^{(s')}}\right).$$

Combining the classifier and the kernel $\bar{k}$, a family of the DG based score functions can be denoted as

$$\mathcal{G}_{\bar{k}} = \{(\tilde{\boldsymbol{x}}, y) \in \tilde{\mathcal{X}} \times \{1, \ldots, c\} \mapsto \boldsymbol{a}_y^T \boldsymbol{W}^T \phi_{\bar{k}}(\tilde{\boldsymbol{x}}) : \\ \boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_c)^T, ||\boldsymbol{A}\boldsymbol{W}^T||_{\mathcal{H}_{\bar{k}}} \leq q\Lambda\}.$$

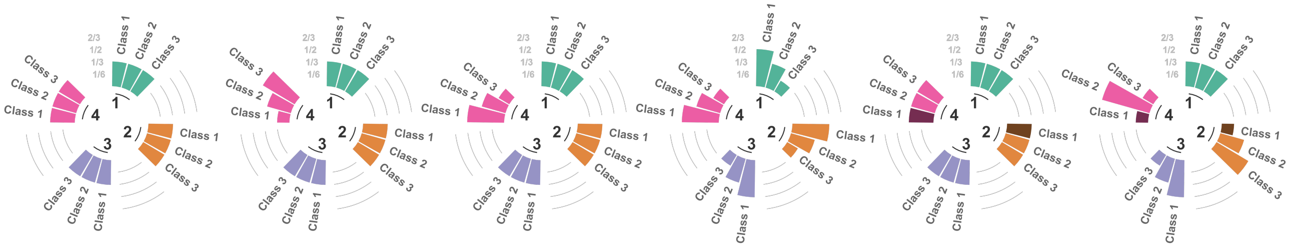
The following assumption makes $\bar{k}$ a bounded universal kernel.

Assumption 2. (i) The kernel k_1 is universal on $\mathcal{X}$, and k_2 is universal and continuous on $\mathcal{X}$, $\mathfrak{K}$ is universal on any compact subset of $\mathcal{H}_{k_2}$. The kernels k_1, k_2 and $\mathfrak{K}$ are bounded by U_1^2, U_2^2 and $U_{\mathfrak{K}}^2$, respectively. (ii) The canonical feature map $\phi_{\mathfrak{K}}$ associated with $\mathfrak{K}$ is $L_{\mathfrak{K}}$-Lipschitz. The observation space $\mathcal{X}$ is a compact metric space.

We have the following theorem regarding the multi-class generalization bound.

Theorem 1. *Given Assumption 2, and assume that $n_j^s = \bar{n}$ for balanced sample size. Then, for a $\rho > 0$ and any $\delta > 0$, with probability at least $1 - \delta$, the following multi-class classification generalization bound holds for all $g \in \mathcal{G}_{\bar{k}}$:*

$$R(g) \leq \hat{R}_{n,\rho}(g) + \frac{1}{\rho}q\Lambda U_1 U_2 L_{\mathfrak{K}}\Big(6\sqrt{\frac{\log 2cm\delta^{-1}}{\bar{n}}} + \\ 4\sqrt{\frac{c}{m\bar{n}}} + 4\sqrt{\frac{c}{m}}\Big) + \sqrt{\frac{\log\delta^{-1}}{2cm\bar{n}}} + \sqrt{\frac{\log\delta^{-1}}{2cm}}. \tag{25}$$



(a) Case 1 (b) Case 2 (c) Case 3 (d) Case 4 (e) Case 5 (f) Case 6

Figure 1. The prior distributions and the variances of the 6 data generalization cases. The bars show the prior probabilities of the different classes within each domain, where the center indexes indicate the domains. The light color indicates that the data are generated with variance one while the darker color (see Cases 5 and 6) means the variance is four.

Table 1. Center points and sample sizes for the synthetic data.

Domain	Domain 1			Domain 2			Domain 3			Domain 4		
Class	1	2	3	1	2	3	1	2	3	1	2	3
X_1	1	4	4	0.5	3.5	3.5	1	4	4	0.5	3.5	3.5
X_2	2	2	-2	1.5	1.5	-2.5	-1.5	-1.5	-5.5	-1.5	-1.5	-5.5
instances	600			600			600			600		

Theorem 1 generalizes the results in Hu et al. (2020) by quantifying the effects of class number c and the feature dimension q introduced by the proposed standardization methods. Indeed, it shows that a larger c or q leads to a weaker guarantee. Given the confidence level $1-\delta$, the excess risk converges to zero if $\frac{\bar{n}}{\log cm}$ and $\frac{m}{c} \to \infty$.

5. Empirical results

We compare the proposed DFDG with the existing DG methods on a synthetic dataset and two real image classification tasks. The two proposed DFDG metrics DFDG-Eig (Section 3.1) and DFDG-Cov (Section 3.2) associated with two classifiers, the 1-nearest neighbor (1-NN) and the support vector machine (SVM), are used for comparison.

The proposed DFDG is compared with the conventional k-NN and SVM without dimension reduction, the Kernel DG methods, namely the domain invariant component analysis (DICA, Muandet et al. 2013), the scatter component analysis (SCA, Ghifary et al. 2017), the conditional invariant DG (CIDG, Li et al. 2018) and the multi-domain discriminant analysis (MDA, Hu et al. 2020), where 1-NN was used for these kernel DG methods. The product kernel (24) was used for all the kernel-based DG methods, where k_1, k_2 and $\mathfrak{K}$ are Gaussian kernels with bandwidth h, h and one, respectively. The bandwidth h is chosen by the median heuristic unless specified otherwise.

Even with the 1-NN classifier, the existing kernel based DG methods typically have three hyperparameters as listed in Table 2. In contrast, the proposed DFDG with 1-NN classifier has one hyperparameter while those with SVM have 3 hyperparameters including a penalty parameter and the kernel bandwidth. The tuning for the kernel bandwidths has been ignored in the existing DG methods (Ramdas et al., 2015). For both the existing and the proposed methods, the hyperparameters were selected by the grid search in the validation set, where 30% of each source domain was chosen as the validation set in the training, the so-called the training-domain validation method (Gulrajani & Lopez-Paz, 2021). The candidate hyperparameters are listed in the SM. After selecting the best hyperparameters in the validation set, the classification accuracy was calculated on the target. We randomly split the source domains as training and validation sets 5 times to calculate the mean and standard deviation of classification accuracy in the target domain.

5.1. Synthetic Data

A two-dimensional dataset with 4 domains and 3 classes was drawn from different Gaussian distributions $\mathcal{N}(\mu, \sigma^2)$ with mean μ (Table 1) and variance σ^2. To investigate the influence of the prior distribution on the classes on different DG methods, the class size may be imbalanced as displayed in Figure 1, while the sample size of each domain was kept 600. The first three domains were the source domains, while the last one was the target domain. All the data were fed into the DG methods without any data preprocessing.

As shown in Table 2, the proposed DFDG outperformed all the kernel DG methods even using only one hyperparameter with the 1-NN classifier. The performance was further lifted by using the SVM classifier with more hyperparameters for the kernel bandwidths and the SVM penalty. See Figure S2 in the SM for the Extracted features by the proposed DFDG methods. The sensitivity analysis provided in SM demonstrated a superior sensitivity performance.

5.2. Case study

We considered three datasets, the Office+Caltech, VLCS and Terra Incognita in case study. The Office+Caltech dataset

Table 2. Mean and standard deviation of the classification accuracy of the synthetic experiments on 6 cases for different methods, where **bold red** and **bold black** indicate the best and second best respectively. And #hp denotes the number of hyperparameters.

Method		#hp	Case1	Case2	Case3	Case4	Case5	Case6
k-NN		1	77.31±0.55	78.14±0.64	76.17±0.46	83.42±1.30	71.17±0.49	51.44±1.48
SVM		2	73.86±1.27	74.86±0.99	73.11±0.89	84.56±0.80	67.28±0.87	44.83±1.04
DICA	1-NN	2	87.25±2.05	84.67±3.36	84.08±1.39	87.03±1.31	78.53±5.18	66.28±1.11
SCA	1-NN	2	87.31±1.17	83.61±0.89	84.69±1.18	86.81±1.12	80.89±1.12	66.58±1.74
MDA	1-NN	3	88.47±1.01	81.00±1.41	82.00±0.51	87.64±1.25	81.14±0.82	64.89±1.41
CIDG	1-NN	4	91.03±0.52	86.58±0.69	84.56±0.81	90.36±0.68	84.52±1.71	69.06±5.79
DFDG-Eig	SVM	3	**93.90±0.48**	**87.57±1.73**	90.03±1.85	**93.40±0.63**	**87.53±0.77**	**79.30±1.84**
	1-NN	1	91.13±0.83	86.87±1.83	**90.23±0.30**	90.57±1.22	84.57±1.63	**75.77±0.67**
DFDG-Cov	SVM	3	**92.97±0.61**	**89.43±1.18**	**92.50±0.35**	**93.57±0.52**	**86.37±0.84**	71.23±1.45
	1-NN	1	89.20±1.20	85.83±1.46	88.83±2.00	90.83±1.13	82.33±0.73	69.60±3.39

Table 3. Accuracy in Office+Caltech and VLCS datasets where **bold red** and **bold black** indicate the best and second best, respectively.

		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
k-NN		79.7	68.6	48.8	61.2	71.5	70.6	46.8	49.5	72.9	48.9	52.5	50.7	42.1	57.5	49.6	56.3
SVM		92.2	82.8	68.7	80.5	**84.9**	84.4	**64.7**	**58.6**	84.9	63.9	**59.5**	63.3	53.6	66.8	**64.9**	70.3
DICA	1-NN	91.8	**83.2**	61.7	80.2	**84.9**	**85.4**	61.7	56.8	87.5	58.7	57.3	55.1	53.7	68.8	60.0	70.0
SCA	1-NN	92.2	82.3	65.0	81.2	**85.2**	83.8	65.3	58.0	89.4	60.7	**58.4**	56.8	54.8	69.8	61.1	70.9
MDA	1-NN	90.3	75.1	56.7	75.9	80.9	78.5	64.4	57.8	90.1	61.0	57.1	61.6	54.4	**70.6**	59.1	69.3
CIDG	1-NN	**92.5**	82.4	68.6	79.5	82.0	83.4	59.6	55.3	88.9	59.5	56.4	56.7	52.0	68.7	58.3	70.4
DFDG-Eig	SVM	92.3	**83.2**	**72.3**	81.2	83.8	**85.0**	60.8	58.4	90.2	**66.2**	**58.4**	**64.2**	**56.4**	**70.8**	63.4	71.2
	1-NN	91.9	82.6	66.2	**82.7**	82.3	84.9	61.4	57.2	**91.6**	64.5	57.0	**63.8**	51.2	68.8	63.7	68.9
DFDG-Cov	SVM	**92.5**	**83.9**	**73.1**	**81.6**	83.8	84.9	**64.6**	**59.5**	91.4	**65.0**	57.6	63.4	**56.5**	70.2	**64.5**	**72.4**
	1-NN	90.5	82.3	68.2	81.2	81.5	84.3	62.6	56.0	**93.0**	62.9	56.1	62.0	51.5	68.3	61.6	**72.0**

(Gong et al., 2012) consists of 2533 images from ten classes over four domains: AMAZON (A), Caltech-256 (C), DSLR (D), and WEBCAM (W). The VLCS dataset (Fang et al., 2013) consists of four domains: PASCAL VOC (V), LabelMe (L), Caltech101 (C) and SUN09 (S), and has 10729 images and five categories. The Terra Incognita data (Beery et al., 2018) were acquired from the DomainBed dataset (Gulrajani & Lopez-Paz, 2021), which contains four locations (domains), 24788 examples and 10 classes. All the images from Office+Caltech and VLCS were preprocessed by feeding into the DeCAF network to extract 4096 dimensional DeCAF features (Donahue et al., 2014). We obtained features for Terra Incognita by training the Empirical Risk Minimization (ERM, Vapnik 1998)-adjusted ResNet 50 and extracting 2048-dimensional features from the last hidden layer. Six cases (domains or combinations of domains) were considered as the target domains for Office+Caltech data, and ten cases were considered for the VLCS data as the target domains as shown in Table 3. To be consistent with the existing studies, we did not consider the four target domains of D, W, A&D and A&W for Office+Caltech, since they all had more than 80% accuracy with the k-NN classifier. For the Terra Incognita dataset, we only considered four single target cases to make them comparable with the results in Gulrajani & Lopez-Paz (2021).

As shown in Table 3, the DFDG with 1-NN classifier achieved a similar performance as the other DG methods but with fewer hyperparameters. While the DFDG with the SVM classifier outperformed others in 9 of the 16 cases. Collectively, the proposed DFDG methods achieved the best performance in 11 out of 16 cases, and the second best in 12 out of 16 cases. The DFDG with SVM classifier significantly outperformed others with a p-value less than 0.002 as shown in Table S4 of the SM. The full results with mean and standard deviation of classification accuracy were given in Tables S5 and S6. We note that since the SVM classifier requires two more hyperparameters, it is hard to implement the SVM for the existing kernel DG methods as the time complexity is exponential with respect to the number of hyperparameters. In contrast, the proposed method can handle the extra computing need with the SVM, as there is only one hyperparameter in the feature selection.

Table 4 demonstrated quite outstanding performance using the proposed methods compared with the ERM baseline and the existing kernel DG methods. This lends support for the suitability of the proposed approach, and provides a way to couple with any deep learning based DG method. Our results showed that the DFDG method with a 1-NN classifier achieved approximately 0.8% performance gain

Table 4. Accuracy in the Terra Incognita dataset, where **bold red** and **bold black** indicate the best and second best, respectively.

method		L100	L38	L43	L46
ERM baseline		53.12	41.07	54.66	36.13
DICA	1-NN	43.81	32.76	48.88	32.51
SCA	1-NN	44.57	39.21	49.00	30.14
MDA	1-NN	39.74	35.44	47.77	26.04
CIDG	1-NN	45.88	38.04	50.43	33.83
DFDG-Eig	SVM	**55.28**	**42.71**	**56.60**	**38.31**
DFDG-Eig	1-NN	53.49	**41.59**	55.68	36.88
DFDG-Cov	SVM	**55.45**	41.58	**55.92**	37.66
DFDG-Cov	1-NN	53.66	**41.59**	54.97	**38.36**

compared to the ERM baseline, while equipping the DFDG method with the SVM classifier increased the classification accuracy by 1.7%. Notably, all the best performances were achieved by the DFDG-based methods. In contrast, the existing kernel DG methods failed to outperform the ERM baseline. A possible reason for this outcome could be the highly imbalanced classes in the Terra Incognita dataset. The class with the smallest number of instances in L38 had only three observations, while the one with the largest number of instances contained 4,485 examples. In such situation, standardization is crucial in handling domain/class dominance issues.

6. Conclusion

This paper proposes a kernel DG algorithm that addresses the fundamental problem of universal generality of a learning approach by proposing two standardization procedures in a unified DG problem framework, which contains fewer hyperparameters. The standardized distribution free metrics can balance the importance of each domain, equally treat each domain and class, and thus is applicable to imbalanced data. We also derive a generalization bound on the multi-class classification problem for the kernel DG methods, and show that the proposed DFDG algorithm produces superior performance in synthetic data and two real image classification experiments.

The proposed framework can be extended to incorporate weighted coefficients towards domains and classes, which enables us to assign a higher weight to the interested domain or the minor class. By reducing the number of hyperparameters, one attains a more efficient invariant feature extraction procedure, that allows for more powerful classifiers with increased generalization ability. One limitation of our work is lack of connections between the number of hyperparameters and the generalization bound, as fewer hyperparameters would reduce the model complexity and tighten the generalization bound. We leave this to future work.

Supplementary Materials

Further technical details, proofs and the example codes are available with this paper at `https://github.com/tongpf/Distribution-Free-Domain-Generalization`.

Acknowledgements

This research was supported by National Natural Science Foundation of China Grant 12026607.

References

Balaji, Y., Sankaranarayanan, S., and Chellappa, R. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems*, 31, 2018.

Barachant, A., Bonnet, S., Congedo, M., and Jutten, C. Multiclass brain–computer interface classification by riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 2012.

Beery, S., Van Horn, G., and Perona, P. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

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.

Blanchard, G., Deshmukh, A. A., Dogan, U., Lee, G., and Scott, C. Domain generalization by marginal transfer learning. *The Journal of Machine Learning Research*, 22 (1):46–100, 2021. ISSN 1532-4435.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. Decaf: A deep convolutional activation feature for generic visual recognition. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 647–655, Bejing, China, 22–24 Jun 2014. PMLR.

Fan, X., Wang, Q., Ke, J., Yang, F., Gong, B., and Zhou, M. Adversarially adaptive normalization for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8208–8217, June 2021.

Fang, C., Xu, Y., and Rockmore, D. N. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1657–1664, 2013.

Ghifary, M., Balduzzi, D., Kleijn, W. B., and Zhang, M. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39 (7):1414–1430, 2017. ISSN 0162-8828.

Gong, B., Shi, Y., Sha, F., and Grauman, K. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2066–2073, 2012.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. ISSN 1532-4435.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Hu, S., Zhang, K., Chen, Z., and Chan, L. Domain generalization via multidomain discriminant analysis. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 292–302. PMLR, 22–25 Jul 2020.

Li, B., Wang, Y., Zhang, S., Li, D., Keutzer, K., Darrell, T., and Zhao, H. Learning invariant representations and risks for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1104–1113, 2021.

Li, Y., Gong, M., Tian, X., Liu, T., and Tao, D. Domain generalization via conditional invariant representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.

Long, M., Cao, Y., Wang, J., and Jordan, M. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105. PMLR, 2015.

Muandet, K., Balduzzi, D., and Schölkopf, B. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18. PMLR, 2013.

Ramdas, A., Jakkam Reddi, S., Poczos, B., Singh, A., and Wasserman, L. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), 2015.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263 – 2291, 2013.

Shawe-Taylor, J., Williams, C. K. I., Cristianini, N., and Kandola, J. On the eigenspectrum of the gram matrix and the generalization error of kernel-pca. *IEEE Transactions on Information Theory*, 51(7):2510–2522, 2005. ISSN 1557-9654.

Shu, Y., Cao, Z., Wang, C., Wang, J., and Long, M. Open domain generalization with domain-augmented meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9624–9633, June 2021.

Vapnik, V. N. *Statistical Learning Theory*. Wiley, 1998.

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., and Yu, P. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2022. ISSN 1041-4347.

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.

Ye, H., Xie, C., Cai, T., Li, R., Li, Z., and Wang, L. Towards a theoretical framework of out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:23519–23531, 2021.

Zhou, K., Yang, Y., Qiao, Y., and Xiang, T. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2021.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Loy, C. C. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2022. ISSN 0162-8828.