Domain Generalization via Multidomain Discriminant Analysis

Shoubo Hu[*], **Kun Zhang**[†], **Zhitang Chen**[‡], **Laiwan Chan**[*]
[*]Department of Computer Science and Engineering, The Chinese University of Hong Kong
[†]Department of Philosophy, Carnegie Mellon University [‡]Huawei Noah's Ark Lab

Abstract

Domain generalization (DG) aims to incorporate knowledge from multiple source domains into a single model that could generalize well on unseen target domains. This problem is ubiquitous in practice since the distributions of the target data may rarely be identical to those of the source data. In this paper, we propose Multidomain Discriminant Analysis (MDA) to address DG of classification tasks in general situations. MDA learns a domain-invariant feature transformation that aims to achieve appealing properties, including a minimal divergence among domains within each class, a maximal separability among classes, and overall maximal compactness of all classes. Furthermore, we provide the bounds on excess risk and generalization error by learning theory analysis. Comprehensive experiments on synthetic and real benchmark datasets demonstrate the effectiveness of MDA.

1 INTRODUCTION

Supervised learning has made considerable progress in tasks such as image classification [Krizhevsky et al., 2012], object recognition [Simonyan and Zisserman, 2014], and object detection [Girshick et al., 2014]. In standard setting, a model is trained on training or source data and then applied on test or target data for prediction, where one implicitly assumes that both source and target data follow the same distribution. However, this assumption is very likely to be violated in real problems. For example, in image classification, images from different sources may be collected under different conditions (e.g., viewpoints, illumination, backgrounds, etc), which makes classifiers trained on one domain perform poorly on instances of *previously unseen* domains. These problems of transferring knowledge to unseen domains are known as domain generalization (DG; [Blanchard et al., 2011]). Note that no data from target domain is available in DG, whereas unlabeled data from the target domain is usually available in domain adaptation, for which a much richer literature exists (e.g., see Patel et al. [2015]).

Denote the space of feature X by $\mathcal{X}$ and the space of label Y by $\mathcal{Y}$. A domain is defined as a joint distribution $\mathbb{P}(X, Y)$ over $\mathcal{X} \times \mathcal{Y}$. In DG of classification tasks, one is given m sample sets, which were generated from m source domains, for model training. The goal is to incorporate the knowledge from source domains to improve the model generalization ability on an *unseen* target domain. An example of DG is shown in Figure 1.

Although various techniques such as kernel methods [Muandet et al., 2013, Ghifary et al., 2017, Li et al., 2018b], support vector machine (SVM) [Khosla et al., 2012, Xu et al., 2014], and deep neural network [Ghifary et al., 2015, Motiian et al., 2017, Li et al., 2017, 2018a,c], have been adopted to solve DG problem, the general idea, which is learning a domain-invariant representation with stable (conditional) distribution in all domains, is shared in most works. Among previous works, kernel-based methods interpret the domain-invariant representation as a feature transformation from the original input space to a transformed space $\mathbb{R}^q$, in which the (conditional) distribution shift across domains is minimized.

Unlike previous kernel-based methods, which assume that $\mathbb{P}(Y|X)$ keeps stable and only $\mathbb{P}(X)$ changes across domains (i.e., the covariate shift situation [Shimodaira, 2000]), the problem of DG or domain adaptation has also been investigated from a causal perspective [Zhang et al., 2015]. In particular, Zhang et al. [2013] pointed out that for many learning problems, especially for classification tasks, Y is usually the cause of X, and proposed the setting of target shift ($\mathbb{P}(Y)$ changes while $\mathbb{P}(X|Y)$ stays

[*]Correspondence: Shoubo Hu <sbhu@cse.cuhk.edu.hk>

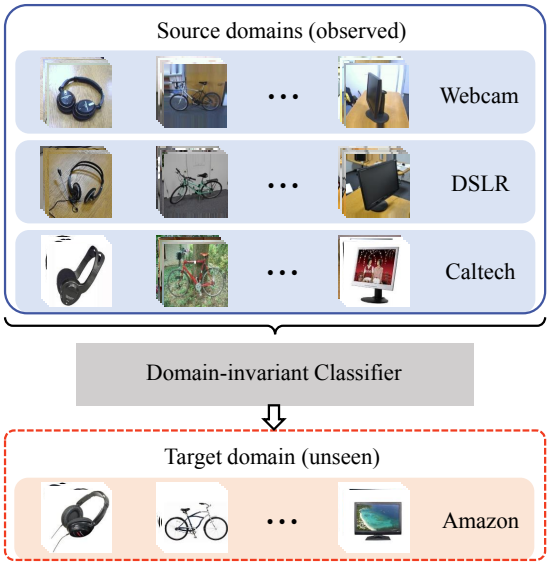


Figure 1: Illustration of DG on Office+Caltech Dataset. One is given source domains: Webcam, DSLR, Caltech, and aims to train a classifier generalizes well on target domain Amazon, which is unavailable in training.

the same across domains), conditional shift ($\mathbb{P}(Y)$ stays the same and $\mathbb{P}(X|Y)$ changes across domains), and their combination accordingly. Gong et al. [2016] proposed to do domain adaptation with conditionally invariant components of X, i.e., the transformations of X that have invariant conditional distribution given Y across domains. Li et al. [2018b] then used this idea for DG, under the assumption of conditional shift. Their assumptions stem from the following postulate of causal independence [Janzing and Scholkopf, 2010, Daniušis et al., 2010]:

Postulate 1 (Independence of cause and mechanism)**.** *If Y causes X ($Y \to X$), then the marginal distribution of the cause, $\mathbb{P}(Y)$, and the conditional distribution of the effect given cause, $\mathbb{P}(X|Y)$, are "independent" in the sense that $\mathbb{P}(X|Y)$ contains no information about $\mathbb{P}(Y)$.*

According to postulate 1, $\mathbb{P}(X|Y)$ and $\mathbb{P}(Y)$ would behave independently across domains. However, this independence typically does not hold in the anti-causal direction [Schölkopf et al., 2012], so $\mathbb{P}(Y|X)$ and $\mathbb{P}(X)$ tends to vary in a coupled manner across domains. Under assumptions that $\mathbb{P}(X|Y)$ changes while $\mathbb{P}(Y)$ keeps stable, generally speaking, both $\mathbb{P}(Y|X)$ and $\mathbb{P}(X)$ change across domains in the anti-causal direction, which is clearly different from the covariate shift situation.

In this paper, we further relax the causally motivated assumptions in Li et al. [2018b] and propose a novel DG method, which is applicable when both $\mathbb{P}(X|Y)$ and $\mathbb{P}(Y)$ change across domains. Our method focuses on separability between classes and does not enforce the transformed marginal distribution of features to be stable, which allows us to relax the assumption of stable $\mathbb{P}(Y)$. To improve the separability, a novel measure named average class discrepancy, which measures the class discriminative power of source domains, is proposed. Average class discrepancy and other three measures are unified in one objective for feature transformation learning to improve its generalization ability on the target domain. As the second contribution, we derive the bound on excess risk and generalization error[1] for kernel-based domain-invariant feature transformation methods. To the best of our knowledge, this is one of the first works to give theoretical learning guarantees on excess risk of DG. Lastly, experimental results on synthetic and real datasets demonstrate the efficacy of our method in handling varying class prior distributions $\mathbb{P}(Y)$ and complex high-dimensional distributions, respectively.

This paper is organized as follows. Section 2 gives the background on kernel mean embedding. Section 3 introduces our method in detail. Section 4 gives the bounds on excess risk and generalization error for kernel-based methods. Section 5 gives experimental settings and analyzes the results. Section 6 concludes this work.

2 PRELIMINARY ON KERNEL MEAN EMBEDDING

Kernel mean embedding is the main technique to characterize probability distributions in this paper. Kernel mean embedding represents probability distributions as elements in a reproducing kernel Hilbert space (RKHS). More precisely, an RKHS $\mathcal{H}$ over domain $\mathcal{X}$ with a kernel k is a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$. Denoting its inner product by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, RKHS $\mathcal{H}$ fulfills the reproducing property $\langle f(\cdot), k(\boldsymbol{x}, \cdot) \rangle_{\mathcal{H}} = f(\boldsymbol{x})$, where $\phi(\boldsymbol{x}) := k(\boldsymbol{x}, \cdot)$ is the canonical feature map of $\boldsymbol{x}$. The kernel mean embedding of a distribution $\mathbb{P}(X)$ is defined as [Smola et al., 2007, Gretton et al., 2007]:

$$\mu_X := \mathbb{E}_X[\phi(X)] = \int_{\mathcal{X}} \phi(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{x}), \tag{1}$$

where $\mathbb{E}_X[\phi(X)]$ is the expectation of $\phi(X)$ with respect to $\mathbb{P}(X)$. It was shown that μ_X is guaranteed to be an element in the RKHS if $\mathbb{E}_X[k(\boldsymbol{x}, \boldsymbol{x})] < \infty$ is satisfied [Smola et al., 2007]. In practice, given a finite sample of size n, the kernel mean embedding of $\mathbb{P}(X)$ is empirically estimated as $\hat{\mu}_X = \frac{1}{n} \sum_{i=1}^{n} \phi(\boldsymbol{x}_i)$, where $\{\boldsymbol{x}_i\}_{i=1}^{n}$ are independently drawn from $\mathbb{P}(X)$. When k is a char-

[1] Blanchard et al. [2011] proved the generalization error bound of DG in general settings

acteristic kernel [Schölkopf and Smola, 2001], μ_X captures all information about $\mathbb{P}(X)$ [Sriperumbudur et al., 2008], which means that $\|\mu_X - \mu_{X'}\|_{\mathcal{H}} = 0$ if and only if $\mathbb{P}(X)$ and $\mathbb{P}(X')$ are the same distribution.

3 MULTIDOMAIN DISCRIMINANT ANALYSIS

3.1 PROBLEM DEFINITION

DG of classification tasks is studied in this paper. Let $\mathcal{X}$ be the feature space, $\mathcal{Y}$ be the space of class labels, and c be the number of classes. A domain is defined to be a joint distribution $\mathbb{P}(X, Y)$ on $\mathcal{X} \times \mathcal{Y}$. Let $\mathfrak{P}_{\mathcal{X}\times\mathcal{Y}}$ denote the set of domains $\mathbb{P}(X, Y)$ and $\mathfrak{P}_{\mathcal{X}}$ denote the set of distributions $\mathbb{P}(X)$ on $\mathcal{X}$. We assume that there is an underlying finite-variance unimodal distribution over $\mathfrak{P}_{\mathcal{X}\times\mathcal{Y}}$. In practice, domains are not observed directly, but given in the form of finite sample sets.

Assumption 1 (Data-generating process)**.** *Each sample set is assumed to be generated in two separate steps: 1) a domain $\mathbb{P}^s(X, Y)$ is sampled from $\mathfrak{P}_{\mathcal{X}\times\mathcal{Y}}$, where s is the domain index; 2) n^s independent and identically distributed (i.i.d.) instances are then drawn from $\mathbb{P}^s(X, Y)$.*

Suppose there are m domains sampled from $\mathfrak{P}_{\mathcal{X}\times\mathcal{Y}}$, the set of m observed sample sets is denoted by $\mathcal{D} = \{\mathcal{D}^s\}_{s=1}^m$, where each $\mathcal{D}^s = \{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{n^s}$ consists of n^s i.i.d. instances from $\mathbb{P}^s(X, Y)$. Since in general $\mathbb{P}^s(X, Y) \neq \mathbb{P}^{s'}(X, Y)$, instances in $\mathcal{D}$ are not i.i.d.

In DG of classification tasks, one aims to incorporate the knowledge in $\mathcal{D}$ into a model which could generalize well on a previously unseen target domain. In this work, features X are first mapped to an RKHS $\mathcal{H}$. Then we resort to learning a transformation from the RKHS $\mathcal{H}$ to a q-dimensional transformed space $\mathbb{R}^q$, in which instances of the same class are close and instances of different classes are distant from each other. 1-nearest neighbor is adopted to conduct classification in $\mathbb{R}^q$.

Table 1: Notations used in the paper

Notation	Description	Notation	Description
X, Y	feature/label variable	$\boldsymbol{x}, y$	feature/label instance
m, c	# domains/classes	s, j	domain/class index
$\mathfrak{P}_{\mathcal{X}\times\mathcal{Y}}, \mathfrak{P}_{\mathcal{X}}$	the set of $\mathbb{P}(X, Y)$ / $\mathbb{P}(X)$	$\mathcal{D}^s$	sample set of domain s
$\mathbb{P}_j^s$	class-conditional distribution	μ_j^s	kernel mean embedding of $\mathbb{P}_j^s$
u_j	mean representation of class j	$\bar{u}$	mean representation of $\mathcal{D}$
k	kernel	$\mathcal{H}_k$	RKHS associated with k

3.2 REGULARIZATION MEASURES

3.2.1 Average Domain Discrepancy

To achieve the goal that instances of the same class are close to each other, we first consider minimizing the discrepancy of the class-conditional distributions, $\mathbb{P}^s(X|Y = j)$, within each class over all source domains.

For ease of notation, the class-conditional distribution of class j in domain s, $\mathbb{P}^s(X|Y = j)$, is denoted by $\mathbb{P}_j^s$. Denoting the kernel mean embedding (1) of $\mathbb{P}_j^s$ by μ_j^s, the average domain discrepancy is defined below.

Definition 1 (Average domain discrepancy)**.** Given the set of all class-conditional distributions $\mathcal{P} = \{\mathbb{P}_j^s\}$ for $s \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, c\}$, the average domain discrepancy, $\Psi^{add}(\mathcal{P})$, is defined as

$$\Psi^{add}(\mathcal{P}) \coloneqq \frac{1}{c\binom{m}{2}} \sum_{j=1}^{c} \sum_{1\leq s<s'\leq m} \|\mu_j^s - \mu_j^{s'}\|_{\mathcal{H}}^2, \quad (2)$$

where $\binom{m}{2}$ is the number of 2-combinations from a set of m elements, $\|\cdot\|_{\mathcal{H}}^2$ denotes the squared norm in RKHS $\mathcal{H}$, and $\|\mu_j^s - \mu_j^{s'}\|_{\mathcal{H}}$ is thus the Maximum Mean Discrepancy (MMD; [Gretton et al., 2007]) between $\mathbb{P}_j^s$ and $\mathbb{P}_j^{s'}$.

The following theorem shows that $\Psi^{add}(\mathcal{P})$ is suitable for measuring the discrepancy between class-conditional distributions of the same class from multiple domains.

Theorem 1. *Let $\mathcal{P}$ denote the set of all class-conditional distributions. If k is a characteristic kernel [Schölkopf and Smola, 2001], $\Psi^{add}(\mathcal{P}) = 0$ if and only if $\mathbb{P}_j^1 = \mathbb{P}_j^2 = \cdots = \mathbb{P}_j^m$, for $j = 1, \ldots, c$.*

Proof. Since k is a characteristic kernel, $\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$ is a metric and attains 0 if and only if $\mathbb{P} = \mathbb{Q}$ for any distributions $\mathbb{P}$ and $\mathbb{Q}$ [Sriperumbudur et al., 2008]. Therefore, $\|\mu_j^s - \mu_j^{s'}\|_{\mathcal{H}} = 0$ if and only if $\mathbb{P}_j^s = \mathbb{P}_j^{s'}$ for all s and s' given j, which means $\mathbb{P}_j^1 = \mathbb{P}_j^2 = \cdots = \mathbb{P}_j^m$ within each class j. Conversely, if $\mathbb{P}_j^1 = \mathbb{P}_j^2 = \cdots = \mathbb{P}_j^m$ for $j = 1, \ldots, c$, then each term $\|\mu_j^s - \mu_j^{s'}\| = 0$ and $\Psi^{\text{add}}(\mathcal{P})$ is thus 0. □

3.2.2 Average Class Discrepancy

Minimizing average domain discrepancy Ψ^{add} (2) would make the means of class-conditional distributions of the same class close in $\mathcal{H}$. However, it is possible that the means of class-conditional distributions of different classes are also close, which is a major source of performance reduction of existing kernel-based DG methods. To this end, average class discrepancy is proposed.

Definition 2 (Average class discrepancy)**.** Let $\mathcal{P}$ denote the set of all class-conditional distributions. The average class discrepancy is defined as

$$\Psi^{acd}(\mathcal{P}) \coloneqq \frac{1}{\binom{c}{2}} \sum_{1\leq j<j'\leq c} \|u_j - u_{j'}\|_{\mathcal{H}}^2, \quad (3)$$

where $u_j = \sum_{s=1}^{m} \mathbb{P}(S = s|Y = j)\mu_j^s$ is the mean representation of class j in RKHS $\mathcal{H}$.

It was shown in Sriperumbudur et al. [2010] that the MMD between two distributions $\mathbb{P}$ and $\mathbb{Q}$, $\text{MMD}[\mathbb{P}, \mathbb{Q}] \leq \sqrt{C} W_1(\mathbb{P}, \mathbb{Q})$ for some constant C satisfying $\sup_{\boldsymbol{x} \in \mathcal{X}} k(\boldsymbol{x}, \boldsymbol{x}) \leq C < \infty$, where $W_1(\mathbb{P}, \mathbb{Q})$ denotes the first Wasserstein distance [Barrio et al., 1999] between distributions $\mathbb{P}$ and $\mathbb{Q}$. In other words, if $\mathbb{P}$ and $\mathbb{Q}$ are distant in MMD metric, they are also distant in the first Wasserstein distance. Therefore, distributions of different classes tend to be distinguishable by maximizing average class discrepancy, $\Psi^{acd}(\mathcal{P})$.

3.2.3 Incorporating Instance-level Information

Both average domain discrepancy $\Psi^{add}(\mathcal{P})$ (2) and average class discrepancy $\Psi^{acd}(\mathcal{P})$ (3) are defined based on the kernel mean embedding of class-conditional distributions $\mathbb{P}_j^s$. By simultaneously minimizing Ψ^{add} (2) and maximizing Ψ^{acd} (3), one would make class-conditional kernel mean embeddings within each class close and the those of different classes distant in $\mathcal{H}$. However, certain subtle information, such as the compactness of the distribution, is not captured in Ψ^{add} and Ψ^{acd}. As a result, although all mean embeddings satisfy the desired condition, there may still be a high chance of misclassification for some instances. To incorporate such information conveyed in each instance, we propose two extra measures based on kernel Fisher discriminant analysis [Mika et al., 1999]. The first is multidomain between-class scatter.

Definition 3 (Multidomain between-class scatter)**.** Let $\mathcal{D}$ denote the set of n instances from m domains, each of which consists of c classes. The multidomain between-class scatter is

$$\Psi^{mbs}(\mathcal{D}) \coloneqq \frac{1}{n} \sum_{j=1}^{c} n_j \|u_j - \bar{u}\|_{\mathcal{H}}^2, \tag{4}$$

where n_j is the total number of instances in class j, and $\bar{u} = \sum_{j=1}^{c} P(Y = j) u_j$ is the the mean representation of the entire set $\mathcal{D}$ in $\mathcal{H}$.

Both $\Psi^{mbs}(\mathcal{D})$ and $\Psi^{acd}(\mathcal{P})$ measure the discrepancy between the distributions of different classes. The difference stems from the weight n_j in $\Psi^{mbs}(\mathcal{D})$ (4). By adding n_j, each term in $\Psi^{mbs}(\mathcal{D})$ is equivalent to pooling all instances of the same class together and summing up their distance to $\bar{u}$. In other words, $\Psi^{mbs}(\mathcal{D})$ corresponds to a simple pooling scheme. Note that when the proportion of instances of each class is the same across all domains (i.e., $n_j^s/n^s = n_j^{s'}/n^{s'}, \forall s, s'$ for $j = 1, \ldots, c$, where n_j^s is the number of instances of class j in domain s), $\Psi^{mbs}(\mathcal{D})$ is consistent with the between-class scatter in Mika et al. [1999].

Multidomain within-class scatter, as a straightforward counterpart of multidomain between-class scatter (4), is defined as follows.

Definition 4 (Multidomain within-class scatter)**.** Let $\mathcal{D}$ denote the set of n instances from m domains, each of which consists of c classes. The multidomain within-class scatter is

$$\Psi^{mws}(\mathcal{D}) \coloneqq \frac{1}{n} \sum_{j=1}^{c} \sum_{s=1}^{m} \sum_{i=1}^{n_j^s} \|\phi(\boldsymbol{x}_{i \in j}^s) - u_j\|_{\mathcal{H}}^2, \tag{5}$$

where $\boldsymbol{x}_{i \in j}^s$ denotes the feature vector of ith instance of class j in domain s.

The definition above indicates that multidomain within-class scatter measures the sum of the distance between the canonical feature map of each instance and the mean representation in RKHS $\mathcal{H}$ of the class it belongs to. It differs from average domain discrepancy in that the information of every instance is considered in multidomain within-class scatter. As a result, by minimizing $\Psi^{mws}(\mathcal{D})$, one increases the overall compactness of the distributions across classes. Similar to $\Psi^{mbs}(\mathcal{D})$, when the proportion of instances of each class is the same across all domains (i.e., $n_j^s/n^s = n_j^{s'}/n^{s'}, \forall s, s'$ for $j = 1, \ldots, c$), $\Psi^{mws}(\mathcal{D})$ is consistent with the within-class scatter in Mika et al. [1999].

We note that each of the measures has its unique contribution and that ignoring any of them may lead to suboptimal solutions, as demonstrated by the empirical results and illustrated in Appendix A.

3.3 FEATURE TRANSFORMATION

Our method resorts to finding a suitable transformation from RKHS $\mathcal{H}$ to a q-dimensional transformed space $\mathbb{R}^q$, i.e., $\mathbf{W} : \mathcal{H} \mapsto \mathbb{R}^q$. We elaborate how the proposed measures are transformed to $\mathbb{R}^q$ in this section.

According to the property of norm in RKHS, $\Psi^{add}(\mathcal{P})$ can be equivalently computed as

$$\text{tr}\left(\frac{1}{c\binom{m}{2}} \sum_{j=1}^{c} \sum_{1 \leq s < s' \leq m} (\mu_j^s - \mu_j^{s'})(\mu_j^s - \mu_j^{s'})^T \right), \tag{6}$$

where $\text{tr}(\cdot)$ denotes the trace operator.

Let the data matrix $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T \in \mathbb{R}^{n \times d}$, where d is the dimension of input features X and $n = \sum_{s=1}^{m} n^s$, and the feature matrix $\mathbf{\Phi} = [\phi(\boldsymbol{x}_1), \ldots, \phi(\boldsymbol{x}_n)]^T$, where $\phi : \mathbb{R}^d \mapsto \mathcal{H}$ denotes the canonical feature map. Then $\mathbf{W}$ can be expressed as a linear combination of all canonical feature maps in $\mathbf{\Phi}$ [Schölkopf et al., 1998], i.e.,

$\mathbf{W} = \mathbf{\Phi}^T\mathbf{B}$, where $\mathbf{B}$ is a matrix collecting coefficients of canonical feature maps. Then by applying the transformation $\mathbf{W}$, $\Psi^{add}(\mathcal{P})$ in trace formulation (6) becomes

$$\Psi^{add}_{\mathbf{B}} = \text{tr}\left(\mathbf{B}^T\mathbf{G}\mathbf{B}\right), \tag{7}$$

where

$$\mathbf{G} = \frac{1}{c\binom{m}{2}}\sum_{j=1}^{c}\sum_{1\leq s<s'\leq m}\mathbf{\Phi}(\mu_j^s - \mu_j^{s'})(\mu_j^s - \mu_j^{s'})^T\mathbf{\Phi}^T. \tag{8}$$

Similarly, after applying the transformation $\mathbf{W}$, average class discrepancy $\Psi^{acd}(\mathcal{P})$ (3), multidomain between-class scatter $\Psi^{mbs}(\mathcal{D})$ (4), and multidomain within-class scatter $\Psi^{mws}(\mathcal{D})$ (5) are given by:

$$\Psi^{acd}_{\mathbf{B}} = \text{tr}\left(\mathbf{B}^T\mathbf{F}\mathbf{B}\right), \Psi^{mbs}_{\mathbf{B}} = \text{tr}\left(\mathbf{B}^T\mathbf{P}\mathbf{B}\right),$$
$$\Psi^{mws}_{\mathbf{B}} = \text{tr}\left(\mathbf{B}^T\mathbf{Q}\mathbf{B}\right), \tag{9}$$

where

$$\mathbf{F} = \frac{1}{\binom{c}{2}}\sum_{1\leq j<j'\leq c}\mathbf{\Phi}(u_j - u_{j'})(u_j - u_{j'})^T\mathbf{\Phi}^T, \tag{10}$$

$$\mathbf{P} = \frac{1}{n}\sum_{j=1}^{c} n_j\mathbf{\Phi}(u_j - \bar{u})(u_j - \bar{u})^T\mathbf{\Phi}^T, \tag{11}$$

$$\mathbf{Q} = \frac{1}{n}\sum_{j=1}^{c}\sum_{s=1}^{m}\sum_{i=1}^{n_j^s}\mathbf{\Phi}(\phi(\boldsymbol{x}^s_{i\in j}) - u_j)(\phi(\boldsymbol{x}^s_{i\in j}) - u_j)^T\mathbf{\Phi}^T. \tag{12}$$

3.4 EMPIRICAL ESTIMATION

In practice, one exploits a finite number of instances from m source domains to estimate the transformed measures in $\mathbb{R}^q$. Since all measures depend on μ_j^s and u_j, the estimation of measures reduces to the estimation of μ_j^s and u_j ($s = 1, \ldots, m, j = 1, \ldots, c$) using the source data. Let $\boldsymbol{x}^s_{i\in j}$ denote the feature vector of ith instance of class j in domain s and n_j^s denote the total number of instances of class j in domain s, each μ_j^s can be empirically estimated as

$$\hat{\mu}_j^s = \frac{1}{n_j^s}\sum_{i=1}^{n_j^s}\phi(\boldsymbol{x}^s_{i\in j}). \tag{13}$$

The empirical estimation of u_j requires $\mathbb{P}(S = s|Y = j)$, which can be estimated using Bayes rule as $\mathbb{P}(S = s|Y = j) = \frac{\Pr(Y=j|S=s)\Pr(S=s)}{\Pr(Y=j)}$. Since it is usually hard to model the underlying distribution over $\mathfrak{P}_{\mathcal{X}\times\mathcal{Y}}$, we assume that the probabilities of sampling all source domains are equal, i.e., $\Pr(S = s) = \frac{1}{m}$ for $s = 1, \ldots, m$ given $\mathcal{D}$. As a result, $\mathbb{P}(S = s|Y = j) = \frac{n_j^s/n^s}{\sum_{s'=1}^{m}(n_j^{s'}/n^{s'})}$. Then the empirical estimation of the mean representation of class j in RKHS $\mathcal{H}$ is given by

$$\hat{u}_j = \sum_{s=1}^{m}\frac{n_j^s/n^s}{\sum_{s'=1}^{m}(n_j^{s'}/n^{s'})}\hat{\mu}_j^s. \tag{14}$$

By substituting the empirical class-conditional kernel mean embedding (13) and empirical mean representation of each class (14) into formulation (8), (10), (11), and (12), these matrices can be estimated from m observed sample sets using the *kernel trick* [Theodoridis and Koutroumbas, 2008].

3.5 THE OPTIMIZATION PROBLEM

Following the solution in Ghifary et al. [2017] and in the spirit of Fisher's discriminant analysis [Mika et al., 1999], we unify measures introduced in previous sections and solve the matrix $\mathbf{B}$ as

$$\arg\max_{\mathbf{B}}\frac{\Psi^{acd}_{\mathbf{B}} + \Psi^{mbs}_{\mathbf{B}}}{\Psi^{add}_{\mathbf{B}} + \Psi^{mws}_{\mathbf{B}}}. \tag{15}$$

It can be seen that through maximizing the numerator, the objective (15) preserves the separability among different classes. Through minimizing the denominator, (15) tries to find a domain-invariant transformation which improves the overall compactness of distributions of all classes and make the class-conditional distributions of the same class as close as possible.

By substituting the transformed average domain discrepancy (7), average class discrepancy, multidomain between-class scatter, and multidomain within-class scatter (9), adding $\mathbf{W}^T\mathbf{W} = \mathbf{B}^T\mathbf{K}\mathbf{B}$ for regularization, and introducing a trade-off between the measures for further flexibility into the objective (15), we aim to achieve

$$\arg\max_{\mathbf{B}} = \frac{\text{tr}\left(\mathbf{B}^T\left(\beta\mathbf{F} + (1-\beta)\mathbf{P}\right)\mathbf{B}\right)}{\text{tr}\left(\mathbf{B}^T(\gamma\mathbf{G} + \alpha\mathbf{Q} + \mathbf{K})\mathbf{B}\right)}, \tag{16}$$

where α, β, and γ are trade-off parameters controlling the significance of corresponding measures. Since the objective (16) is invariant to re-scaling of $\mathbf{B}$, rewriting (16) as a constrained optimization problem and setting the derivative of its Lagrangian to zero (see Appendix B) yields the following generalized eigenvalue problem:

$$\left(\beta\mathbf{F} + (1-\beta)\mathbf{P}\right)\mathbf{B} = \left(\gamma\mathbf{G} + \alpha\mathbf{Q} + \mathbf{K}\right)\mathbf{B}\mathbf{\Gamma}, \tag{17}$$

where $\mathbf{\Gamma} = \text{diag}(\lambda_1, \ldots, \lambda_q)$ is the diagonal matrix collecting q leading eigenvalues, $\mathbf{B}$ is the matrix collecting corresponding eigenvectors.[2]

[2]In practice, $\gamma\mathbf{G}+\alpha\mathbf{Q}+\mathbf{K}$ is replaced by $\gamma\mathbf{G}+\alpha\mathbf{Q}+\mathbf{K}+\epsilon\mathbf{I}$ for numerical stability, where ϵ is a small constant and set to be $1\text{e}{-5}$ for kernel-based DG methods in all experiments.

Computing the matrices $\mathbf{G}$, $\mathbf{F}$, $\mathbf{P}$, and $\mathbf{Q}$ takes $\mathcal{O}(n^2)$. Solving the generalized eigenvalue problem (17) takes $\mathcal{O}(qn^2)$. In sum, the overall computational complexity is $\mathcal{O}(n^2+qn^2)$, which is the same as existing kernel-based methods. After the transformation learning, unseen target instances can then be transformed into $\mathbb{R}^q$ using $\mathbf{B}$ and $\mathbf{\Gamma}$. We term the proposed method Multidomain Discriminant Analysis (MDA) and summarize the algorithm in Algorithm 1.

Algorithm 1: Multidomain discriminant analysis

input : $\mathcal{D} = \{\mathcal{D}^s\}_{s=1}^m$ - the set of instances from m domains;
α, β, γ - trade-off parameters.

output: Optimal projection $\mathbf{B}_{n\times q}$;
corresponding eigenvalues $\mathbf{\Gamma}$.

Construct kernel matrix $\mathbf{K}$ from $\mathcal{D}$, whose entry on ith row and i'th column $[\mathbf{K}]_{ii'} = k(\boldsymbol{x}_i, \boldsymbol{x}_{i'})$;

Compute matrices $\mathbf{G}$, $\mathbf{F}$, $\mathbf{P}$, $\mathbf{Q}$ from (8), (10), (11), (12), respectively;

Center the kernel matrix as $\mathbf{K} \leftarrow \mathbf{K} - \mathbf{1}_n\mathbf{K} - \mathbf{K}\mathbf{1}_n + \mathbf{1}_n\mathbf{K}\mathbf{1}_n$, where $\mathbf{1}_n \in \mathbb{R}^{n\times n}$ denotes a matrix with all entries equal to $\frac{1}{n}$;

Solve (17) for the projection $\mathbf{B}$ and corresponding eigenvalues $\mathbf{\Gamma}$, then select q leading components.

Target domain transformation

Denote the set of instances from the target domain by $\mathcal{D}^t$, one first constructs the kernel matrix $\mathbf{K}^t$, where $[\mathbf{K}^t]_{i'i} = k(\boldsymbol{x}^t_{i'}, \boldsymbol{x}_i)$, $\forall \boldsymbol{x}^t_{i'} \in \mathcal{D}^t$, $\forall \boldsymbol{x}_i \in \mathcal{D}$;

Center the kernel matrix as $\mathbf{K}^t \leftarrow \mathbf{K}^t - \mathbf{1}_{n^t}\mathbf{K}^t - \mathbf{K}^t\mathbf{1}_n + \mathbf{1}_{n^t}\mathbf{K}^t\mathbf{1}_n$, where n^t is the number of instances in $\mathcal{D}^t$;

Then the transformed features of the target domain are given by $\mathbf{X}^t = \mathbf{K}^t\mathbf{B}\mathbf{\Gamma}^{-\frac{1}{2}}$.

4 LEARNING THEORY ANALYSIS

We analyze the the excess risk and generalization error bound after applying feature transformation $\mathbf{W}$.

In standard setting of learning theory analysis, the decision functions of interest are $f : \mathcal{X} \mapsto \mathcal{Y}$. However, our DG problem setting is much more general in the sense that not only $\mathbb{P}(X)$ changes (as in the covariate shift setting), but $\mathbb{P}(Y|X)$, which corresponds to f in learning theory, also changes across domains. As a result, the decision functions of interest in our analysis are $f : \mathfrak{P}_\mathcal{X} \times \mathcal{X} \mapsto \mathcal{Y}$. $\mathbb{P}^s$ and $\mathbb{P}^s_X$ are used interchangeably to denote the marginal distribution of X in domain s.

Let $\bar{k}$ be a kernel on $\mathfrak{P}_\mathcal{X} \times \mathcal{X}$ and $\mathcal{H}_{\bar{k}}$ be the associated RKHS. As in Blanchard et al. [2011], we consider kernel $\bar{k} = k_\mathbb{P}(\mathbb{P}^1, \mathbb{P}^2)k_x(x_1, x_2)$, where $k_\mathbb{P}$ and k_x are kernels on $\mathfrak{P}_\mathcal{X}$ and $\mathcal{X}$, respectively. To ensure that $\bar{k}$ is universal, we consider a particular form for $k_\mathbb{P}$. Let k'_x be another kernel on $\mathcal{X}$ and $\mathcal{H}_{k'_x}$ be its associated RKHS, γ be a mapping $\gamma : \mathfrak{P}_\mathcal{X} \mapsto \mathcal{H}_{k'_x}$. Then $k_\mathbb{P}$ defined as a kernel k_γ on $\mathcal{H}_{k'_x}$, i.e. $k_\mathbb{P}(\mathbb{P}^1, \mathbb{P}^2) = k_\gamma(\gamma(\mathbb{P}^1), \gamma(\mathbb{P}^2))$ would lead $\bar{k}$ to be universal [Blanchard et al., 2011]. We consider following assumptions regarding the kernels and loss function in our analysis:

Assumption 2. *The kernels k_x, k'_x and k_γ are bounded respectively by $U^2_{k_x}$, $U^2_{k'_x}$ and $U^2_{k_\gamma}$.*

Assumption 3. *The canonical feature map $\gamma_{k_\gamma} : \mathcal{H}_{k'_x} \mapsto \mathcal{H}_{k_\gamma}$, where $\mathcal{H}_{k_\gamma}$ is the RKHS associated with k_γ, fulfills that $\forall v, w \in \mathcal{H}_{k'_x}$, there is a constant L_{k_γ} satisfying*

$$\|\gamma_{k_\gamma}(v) - \gamma_{k_\gamma}(w)\| \le L_{k_\gamma}\|v - w\|.$$

Assumption 4. *The loss function $\ell : \mathbb{R} \times \mathcal{Y} \mapsto \mathbb{R}_+$ is L_ℓ-Lipschitz in its first variable and bounded by U_ℓ.*

Assumption 2 and 3 are satisfied when the kernels are bounded. An example of widely adopted bounded kernel is the Gaussian kernel. As a result, we also adopt Gaussian kernel throughout our algorithm.

Let $\tilde{X}^t = (\mathbb{P}^t_X, X^t)$ and Y^t denote the extended input and output pattern of decision function f over target domain, respectively. The quantity of interest is the excess risk, which is the difference between expected test loss of empirical loss minimizer and expected loss minimizer. For functions in the unit ball centered at the origin in the RKHS of $\phi(\tilde{X}^t)$, the control of the excess risk is given in the following theorem.

Theorem 2. *Under assumptions 2 – 4, and further assuming that $\|\hat{f}\|_{\mathcal{H}_{\bar{k}}} \le 1$ and $\|f^*\|_{\mathcal{H}_{\bar{k}}} \le 1$, where $\hat{f}$ denotes the empirical risk minimizer and f^* denotes the expected risk minimizer, then with probability at least $1-\delta$ there is*

$$\begin{aligned}&\mathbb{E}[\ell(\hat{f}(\tilde{X}^t\mathbf{W}), Y^t)] - \mathbb{E}[\ell(f^*(\tilde{X}^t\mathbf{W}), Y^t)] \\ &\le 4L_\ell L_{k_\gamma} U_{k'_x} U_{k_x}\sqrt{\frac{\mathrm{tr}(\mathbf{B}^T\mathbf{K}\mathbf{B})}{n}} + \sqrt{\frac{2\log 2\delta^{-1}}{n}},\end{aligned} \tag{18}$$

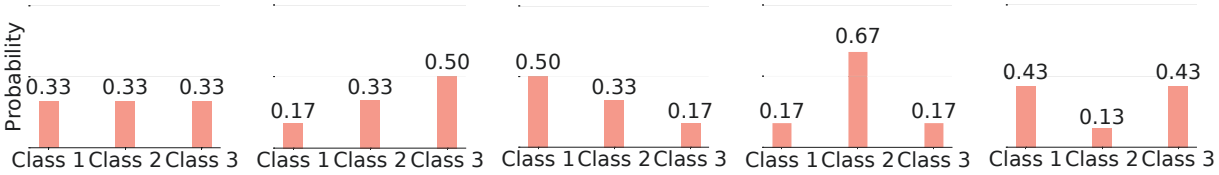
where the expectations are taken over the joint distribution of the test domain $\mathbb{P}^t(X^t, Y^t)$, n is the number of training samples, and $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^T$.

See Appendix C for proof. The first term in the bound above involves the size of the distortion $\mathrm{tr}(\mathbf{B}^T\mathbf{K}\mathbf{B})$ introduced by $\mathbf{B}$. Therefore, a poor choice of $\mathbf{B}$ would loose the guarantee. The second term is of order $O(n^{-1/2})$ so it would converge to zero as n tends to infinity given δ.

Another quantity of interest is the generalization error bound, which is the difference between the expected test

Table 2: Generating Distributions of Synthetic Data

Domain	Domain 1			Domain 2			Domain 3		
Class	1	2	3	1	2	3	1	2	3
X_1	(1, 0.3)	(2, 0.3)	(3, 0.3)	(3.5, 0.3)	(4.5, 0.3)	(5.5, 0.3)	(8, 0.3)	(9.5, 0.3)	(10, 0.3)
X_2	(2, 0.3)	(1, 0.3)	(2, 0.3)	(2.5, 0.3)	(1.5, 0.3)	(2.5, 0.3)	(2.5, 0.3)	(1.5, 0.3)	(2.5, 0.3)
# instances	50	50	50	50	50	50	50	50	50



(a) (b) (c) (d) (e)

Figure 2: Class Prior Distributions $\mathbb{P}(Y)$ in Synthetic Experiments.

loss and empirical training loss of the empirical loss minimizer. The generalization error bound of DG in a general setting is given in Blanchard et al. [2011]. Therefore, we derive it for the case where one applies feature transformation involving $\mathbf{B}$. Let $\hat{\tilde{X}}_i^s$ denote the input pattern $(\hat{\mathbb{P}}^s, x_i^s)$, where $\hat{\mathbb{P}}^s$ is the empirical distribution over features in domain s, and x_i^s is the ith observed feature in domain s. Similarly, y_i^s is the ith label in domain s. With $\mathcal{E}(f, \infty)$ being the expected test loss, the generalization bound involving $\mathbf{B}$ is given in the following theorem.

Theorem 3. *Under assumptions 2 – 4, and assuming that all source sample sets are of the same size, i.e.* $n^s = \bar{n}$ *for* $s = 1, \ldots, m$*, then with probability at least* $1 - \delta$ *there is*

$$\sup_{\|f\|_{\mathcal{H}_{\tilde{k}}} \leq 1} \left| \frac{1}{m} \sum_{s=1}^{m} \frac{1}{n^s} \sum_{i=1}^{n^s} \ell\left(f(\hat{\tilde{X}}_i^s \mathbf{W}), y_i^s\right) - \mathcal{E}(f, \infty) \right|$$
$$\leq U_\ell \left(\sqrt{\frac{\log 2\delta^{-1}}{2m\bar{n}}} + \sqrt{\frac{\log \delta^{-1}}{2m}} \right) + \sqrt{\mathrm{tr}(\mathbf{B}^T \mathbf{K} \mathbf{B})}$$
$$\left(c_1 \sqrt{\frac{\log 2\delta^{-1} m}{\bar{n}}} + c_2 \left(\sqrt{\frac{1}{m\bar{n}}} + \sqrt{\frac{1}{m}} \right) \right), \quad (19)$$

where $c_1 = 2\sqrt{2} L_\ell U_{k_x} L_{k_\gamma} U_{k'_x}$, $c_2 = 2 L_\ell U_{k_x} U_{k_\gamma}$.

See Appendix D for proof. The first term is of order $O(m^{-1/2})$ and converges to zero as $m \to \infty$. The second term, involving $\mathrm{tr}(\mathbf{B}^T \mathbf{K} \mathbf{B})$, again depends on the choice of $\mathbf{B}$. The remaining part would converge to zero only if both m and $\bar{n}$ tend to infinity and $\log m / \bar{n} = o(1)$. In a general perspective, our method, as well as existing ones relying on feature extraction can all be viewed as ways of finding transformation $\mathbf{B}$, which could minimize the generalization bound on the test domain, under different understandings of the DG problem.

5 EXPERIMENTS

5.1 EXPERIMENTAL CONFIGURATION

We compare MDA with the following 9 methods:

- Baselines: 1-nearest neighbor (1NN) and support vector machine (SVM) with RBF kernel.
- Feature extraction methods: kernel principal component analysis (KPCA; [Schölkopf et al., 1998]) and kernel Fisher discriminant analysis (KFD; [Mika et al., 1999]). 1NN is applied on the transformed features for classification.
- SVM-based DG method: low-rank exemplar-SVMs (L-SVM; [Xu et al., 2014]).
- Neural network-based DG method: CCSA [Motiian et al., 2017]. The network setting follows [Motiian et al., 2017].
- Kernel-based DG methods: domain invariant component analysis (DICA; [Muandet et al., 2013]), scatter component analysis (SCA; [Ghifary et al., 2017]), and conditional invariant DG (CIDG; [Li et al., 2018b]). 1NN is applied on the domain-invariant representations for classification.

For 1NN and SVM baselines, instances in source domains are directly combined for training in both synthetic and real data experiments. For other methods, in experiments with synthetic data, the models are trained on two

Table 3: Accuracy (%) of Synthetic Experiments (***bold italic*** and **bold** indicate the best and second best).

$\mathbb{P}^1(Y)$	2a	2b	2c	2d	2e	2a	2a	2a	2a
$\mathbb{P}^2(Y)$	2a	2a	2a	2a	2a	2b	2c	2d	2e
SVM	56.00	34.00	33.33	33.33	33.33	33.33	40.00	36.00	60.00
KPCA	66.00	62.00	66.67	33.33	33.33	65.33	36.00	40.00	14.00
KFD	78.67	38.67	46.00	74.67	47.33	49.33	34.00	19.33	76.00
L-SVM	56.00	60.00	64.00	62.00	60.67	64.67	45.33	46.00	59.33
DICA	**93.33**	84.67	76.00	**84.00**	84.67	54.00	***95.33***	71.33	**88.67**
SCA	79.33	72.00	**84.67**	57.33	76.00	59.33	84.67	61.33	81.33
CIDG	90.67	**87.33**	74.67	77.33	**86.67**	**83.33**	**92.00**	**82.00**	86.00
MDA	***96.67***	***96.00***	***97.33***	***94.00***	***94.00***	***91.33***	***95.33***	***94.00***	***94.00***

source sample sets, validated on one target sample set, and tested on the other target sample set. In real data experiments, we first selected hyper-parameters by 5-fold cross-validation using only labeled source sample sets. Then the model with optimal parameter settings was applied on the target domain. The classification accuracy on the target domain serves as the evaluation criterion for different methods. Since measures in section 3.2 are defined as the averaged distance, we naturally put them on a equal footing by setting $\beta = 0.5$, $\alpha = \gamma = 1$. Thus in practice, these parameters are set to be an interval containing values in the above balanced case. The hyper-parameters required by each method and the values validated in the experiments are given in Appendix E.

Table 4: Accuracy (%) of Office+Caltech Dataset

Target	A	C	A, C	W, D	W, C	D, C
1NN	89.80	84.16	78.63	80.60	86.29	85.28
SVM	91.96	**85.75**	77.66	**84.51**	87.31	86.72
KPCA	89.87	83.35	66.46	79.65	85.83	84.45
KFD	91.75	85.66	74.68	82.96	87.59	86.64
L-SVM	91.64	85.39	**80.55**	83.33	**88.09**	**87.10**
CCSA	90.98	83.37	77.56	80.04	85.80	84.91
DICA	**92.59**	83.17	63.67	83.85	87.59	86.25
SCA	91.96	83.35	73.04	83.85	87.31	86.25
CIDG	92.38	81.39	69.87	82.74	87.45	85.63
MDA	***93.47***	***86.89***	***82.56***	***84.89***	***88.91***	***88.23***

5.2 SYNTHETIC DATA

We investigate the influence of variation in the class prior distribution, $\mathbb{P}(Y)$, on different DG methods. Two-dimensional data is generated from three different domains and each domain consists of three classes. Each dimension of the data follows a Gaussian distribution $\mathcal{N}(\mu, \sigma)$, where μ is the mean and σ is the standard deviation. The settings of the distribution of the synthetic data are listed in Table 2. Domains 1 and 2 are source domains and domain 3 is the target domain. The setting in Table 2 is the base condition where class prior distributions are uniform in all domains, i.e., $\mathbb{P}^1(Y) = \mathbb{P}^2(Y) = \mathbb{P}^3(Y)$. Then we change $\mathbb{P}(Y)$ of one source domain to be distributions shown in Figure 2b to 2e and keep $\mathbb{P}(Y)$ of the other source domain and target domain uniform to compare different DG methods. Note that CCSA is based on convolutional neural network and thus not suitable for 2-dimensional synthetic data.

The results of different methods on different settings of class prior distributions in source domains are given in Table 3 (also visualized in Appendix F). The accuracy of 1NN is 33.33% in all cases thus omitted in Table 3. It can be seen that MDA performs best in the base setting, as well as all settings with different $\mathbb{P}(Y)$ in source domains. DICA performs equally well as MDA in (2a, 2c) setting but its accuracy is heavily influenced by the variation in $\mathbb{P}(Y)$. Compared with other methods, MDA is much more robust against the variation in $\mathbb{P}(Y)$, which is consistent with our expectation because we essentially work with the class-conditional, not the marginal, distributions.

5.3 OFFICE+CALTECH DATASET

We evaluate the performance of different DG methods on Office+Caltech dataset [Gong et al., 2012], which is a widely used benchmark for DG tasks. Office+Caltech consists of photos from four different datasets: Amazon (A), Webcam (W), DSLR (D), and Caltech-256 (C) [Griffin et al., 2007]. Since there are 10 shared classes in these datasets, photos of these classes are selected and those from the same original dataset form one domain in Office+Caltech. Thus, the domains within Office+Caltech corresponds to the biases of different data collection procedures [Torralba and Efros, 2011]. The 4096-dimensional DeCAF_6 features [Donahue et al., 2014] are adopted in the experiments to ensure that the feature spaces, $\mathcal{X}$, are consistent across all domains.

Table 5: Accuracy (%) of VLCS Dataset

Target	V	L	C	S	V, L	V, C	V, S	L, C	L, S	C, S
1NN	60.19	53.57	89.94	55.74	57.26	58.54	50.59	66.06	58.13	66.25
SVM	***68.57***	59.26	***93.99***	***65.27***	***61.80***	***64.39***	**55.89**	70.08	***64.10***	**71.09**
KPCA	60.69	54.86	83.89	55.61	57.54	57.50	49.46	67.48	56.05	66.15
KFD	61.64	**60.54**	86.78	58.75	57.33	46.84	53.20	70.03	61.64	67.87
L-SVM	58.14	39.87	75.56	52.92	52.25	56.64	48.27	61.24	56.65	66.27
CCSA	60.39	58.80	86.88	59.87	59.27	55.02	51.56	69.94	61.41	68.49
DICA	62.71	59.38	86.15	57.28	58.11	55.08	55.17	70.01	61.44	70.30
SCA	62.13	58.24	88.48	**60.66**	**60.66**	57.59	54.66	**71.90**	61.57	70.71
CIDG	64.16	57.91	90.11	59.48	60.54	54.56	55.77	70.74	62.48	69.83
MDA	**66.86**	***61.78***	**92.64**	59.58	59.60	**63.72**	***55.98***	***72.88***	**62.83**	***72.00***

The accuracies on different choices of target domains are shown in Table 4. MDA again performs best, yet by a smaller margin of improvement compared to that of the synthetic experiment. In particular, MDA is the only kernel-based method that outperforms 1NN in (A, C) case which is probably because of the newly proposed average class discrepancy (3). L-SVM outperforms other kernel-based methods and ranks the second. Note that other 4 cases, such as A, D, C $\rightarrow$ W, are not reported since 1NN baseline could already achieve accuracies higher than 90%.

5.4 VLCS DATASET

The second real data experiment uses the VLCS dataset. It consists of photos of five common classes extracted from four datasets: Pascal VOC2007 (V) [Everingham et al., 2010], LabelMe (L) [Russell et al., 2008], Caltech-101 (C) [Griffin et al., 2007], and SUN09 (S) [Choi et al., 2010]. Photos from the same dataset form one domain in VLCS. DeCAF$_6$ features of 4096 dimensions are again adopted in the experiments to ensure the consistency of feature spaces over different domains. The training and test procedures are the same as in experiments on the Office+Caltech dataset. The parameters of L-SVM were trained (validated) on 70% (30%) source instances due to its high complexity.

The accuracies are given in Table 5. It is interesting to see that SVM baseline outperforms all DG methods in 6 cases. This is probably because many instances of different classes are overlapped in VLCS, so using 1NN in the transformed space is more likely to misclassify them compared with SVM. Apart from SVM baseline, MDA performs best in 8 out of 10 cases compared with other DG methods. CCSA outperforms MDA in the case of S being the target domain, which may indicate that neural networks extracted better features in this case. Inspired by the results of SVM, kernel-based methods together with SVM classifier may be a promising direction for further VLCS accuracy improvement.

6 CONCLUSION

In this paper, we proposed a method called Multidomain Discriminant Analysis (MDA) to solve the DG problem of classification tasks. Unlike existing works, which typically assume stability of certain (conditional) distributions, MDA is able to solve DG problems in a more general setting where both $\mathbb{P}(Y)$ and $\mathbb{P}(X|Y)$ change across domains. The newly proposed measures, average domain discrepancy and average class discrepancy, together with two measures based on kernel Fisher discriminant analysis, are theoretically analyzed and incorporated into the objective for learning the domain-invariant feature transformation. We also prove bounds on the excess risk and generalization error for kernel-based DG methods. The effectiveness of MDA is verified by experiments on synthetic and two real benchmark datasets.

Acknowledgements

SH thanks Lequan Yu for comments on a previous draft of this paper. KZ acknowledges the support by National Institutes of Health (NIH) under Contract No. NIH-1R01EB022858-01, FAINR01EB022858, NIH-1R01LM012087, NIH-5U54HG008540-02, and FAIN- U54HG008540, by the United States Air Force under Contract No. FA8650-17-C-7715, and by National Science Foundation (NSF) EAGER Grant No. IIS-1829681. The NIH, the U.S. Air Force, and the NSF are not responsible for the views reported in this article. This work was partially funded by the Hong Kong Research Grants Council.

References

Eustasio Del Barrio, J A Cuestaalbertos, Carlos Matran, and Jes U S M Rodriguezrodriguez. Tests of goodness of fit based on the l_2-wasserstein distance. *Annals of Statistics*, 27(4):1230–1239, 1999.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2178–2186, 2011.

Myung Jin Choi, Joseph J Lim, Antonio Torralba, and Alan S Willsky. Exploiting hierarchical context on a large database of object categories. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 129–136, 2010.

Povilas Daniušis, Dominik Janzing, Joris Mooij, Jakob Zscheischler, Bastian Steudel, Kun Zhang, and Bernhard Schölkopf. Inferring deterministic causal relations. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI 2010)*, pages 143–150, 2010.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, pages 647–655, 2014.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2551–2559, 2015.

Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. 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.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014.

Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2066–2073, 2012.

Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *Proceedings of The 33rd International Conference on Machine Learning (ICML 2016)*, pages 2839–2848, 2016.

Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems (NIPS)*, pages 513–520, 2007.

Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007. URL `http://authors.library.caltech.edu/7694`.

Dominik Janzing and Bernhard Scholkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *The European Conference on Computer Vision (ECCV)*, pages 158–171, 2012.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

Da Li, Yongxin Yang, Yizhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5543–5551, 2017.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5400–5409, 2018a.

Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*, pages 3579–3587, 2018b.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *The European Conference on Computer Vision (ECCV)*, pages 647–663, 2018c.

Sebastian Mika, Gunnar Ratsch, Jason Weston, Bernhard Scholkopf, and Klaus-Robert Mullers. Fisher discriminant analysis with kernels. In *Neural networks for signal processing IX, 1999. Proceedings of the 1999 IEEE signal processing society workshop.*, pages 41–48, 1999.

Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5716–5726, 2017.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, pages 10–18, 2013.

Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015.

Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173, 2008.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001. ISBN 0262194759.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10 (5):1299–1319, 1998.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, pages 1255–1262, 2012.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. URL `http://arxiv.org/abs/1409.1556`.

Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *Proceedings of the 18th International Conference on Algorithmic Learning Theory*, pages 13–31, 2007.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert R G Lanckriet, Bernhard Scholkopf, and R A Servedio T Zhang. Injective hilbert space embeddings of probability measures. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT 2008)*, pages 111–122, 2008.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11 (Apr):1517–1561, 2010.

Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern Recognition, Fourth Edition*. Academic Press, Inc., Orlando, FL, USA, 4th edition, 2008. ISBN 1597492728, 9781597492720.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1521–1528, 2011.

Zheng Xu, Wen Li, Li Niu, and Dong Xu. Exploiting low-rank structure from latent domains for domain generalization. In *The European Conference on Computer Vision (ECCV)*, pages 628–643, 2014.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, pages 819–827, 2013.

Kun Zhang, Mingming Gong, and Bernhard Schölkopf. Multi-source domain adaptation: A causal view. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2015)*, pages 3150–3157, 2015.