# Domain-wise Data Acquisition to Improve Performance under Distribution Shift

Yue He [*1] Dongbai Li [*1] Pengfei Tian [2] Han Yu [1] Jiashuo Liu [1] Hao Zou [3] Peng Cui [1]

## Abstract

Despite notable progress in enhancing the capability of machine learning against distribution shifts, training data quality remains a bottleneck for cross-distribution generalization. Recently, from a data-centric perspective, there have been considerable efforts to improve model performance through refining the preparation of training data. Inspired by realistic scenarios, this paper addresses a practical requirement of acquiring training samples from various domains on a limited budget to facilitate model generalization to target test domain with distribution shift. Our empirical evidence indicates that the advance in data acquisition can significantly benefit the model performance on shifted data. Additionally, by leveraging unlabeled test domain data, we introduce a Domain-wise Active Acquisition framework. This framework iteratively optimizes the data acquisition strategy as training samples are accumulated, theoretically ensuring the effective approximation of test distribution. Extensive real-world experiments demonstrate our proposal's advantages in machine learning applications. The code is available at https://github.com/dongbaili/DAA.

## 1. Introduction

In the past years, the field of machine learning has witnessed unprecedented growth. Traditional machine learning paradigms often assume that training data and test data are drawn from the same distribution. However, this hypothesis starkly contrasts with the real-world scenarios where data exhibits significant heterogeneity, leading to changed distributions across different domains. In addressing the challenges of deploying machine learning models in real applications, there is an increasing focus (Koh et al., 2021; Arjovsky et al., 2019; Ben-David et al., 2010; Shen et al., 2021) on enhancing the model performance when confronted with the distribution shift.

In order to facilitate model generalization to target test domain, domain adaptation methods propose to align the source and target domains in terms of population distribution (Li & Zhang, 2019), data representation (Hoffman et al., 2017), and model parameters (Motiian et al., 2017), respectively. In the scenarios where only unlabeled data from the test domain is available, the solutions of unsupervised domain adaptation include mitigating distribution discrepancies by sample reweighting (Zhang et al., 2018), generating pseudo labels via self-supervised learning (French et al., 2018). However, these methods suppose the training samples are predetermined, thereby their model performance closely depends on the training data quality.

Recently, data-centric machine learning has been receiving increasing attention. By refining the data preparation process (Chen et al., 2023; Mazumder et al., 2024; Schmarje et al., 2022), they have successfully employed more effective training samples to breakthrough the bottlenecks in model performance. To address the distribution shift in test domain, Fu et al. (2021); Xie et al. (2022) propose to select and label the most representative samples from the unlabeled test data, enabling efficient access to test domain information. However, in many real-world applications, sample labels cannot be provided by a third party. For instance, manually determining if an individual has public health insurance based on profile features is difficult without self-disclosure.

In this paper, we consider a new problem setting for learning under distribution shift from a data-centric perspective. It is noteworthy that in many applications, not only the annotations but also the unlabeled data point are difficult to obtain. Therefore, it is limited that we can at most acquire $T$ pairs of both training data and corresponding label as the budget.

[*] Equal contribution [1] Department of Computer Science and Technology, Tsinghua University, Beijing, China [2] Qiuzhen College, Tsinghua University, Beijing, China [3] Zhongguancun Laboratory, Beijing, China. Emails: hy865865@gmail.com, lidongbai30@gmail.com, e9tian@gmail.com, yuh21@mails.tsinghua.edu.cn, liujiashuo77@gmail.com, zouh@zgclab.edu.cn, cuip@tsinghua.edu.cn. Correspondence to: Peng Cui <cuip@tsinghua.edu.cn>.

*Proceedings of the $41^{st}$ International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

It means that there is no predefined large data pool for acquiring label in this problem, which is in distinction with the practice of traditional active learning. Due to the high dimensional property and complex relationship in data feature, it is unfeasible to directly control the data acquisition at the sample-wise level. Fortunately, we usually can access to the meta-information of the data population. Based on the meta-information, the potential acquired data can be divided into several domains. Instead, we can control the domain in which the data is acquired. Assuming the number of data domains is $Q$, our problem becomes how to allocate the sample budget $T$ to the $Q$ domains for better generalization to the test domain. Our empirical evidence indicates that an advanced data acquisition can enhance model performance on shifted data significantly. This emphasizes the pivotal role of raising the training data quality in overcoming the distribution shift.

To tackle the proposed problem, we further put forward a Domain-wise Active Acquisition (DAA) framework, as illustrated in Figure 1. Our approach splits the budget $T$ into $K$ portions, acquiring $T/K$ samples in each round. Starting with a uniform acquisition from $Q$ domains in the initial round, DAA utilizes the previously accumulated $(T * (k - 1))/K$ samples along with unlabeled test data to calculate the domain-wise acquisition weights $W$. These weights, designed to reduce the distribution distance between training and test data, guide the data acquisition process in the $k^{th}$ round. By iteratively optimizing data acquisition strategy through $K$ rounds, our framework theoretically guarantees an effective approaching toward test distribution. Extensive real-world experiments validates the superiority of ours in improving model performance under distribution shift.

In summary, our contributions are highlighted as follows:

1. We investigate to improve model performance under distribution shift from a data-centric perspective, and introduce a practical challenge of pursuit of advanced data acquisition from diverse domains to enhance model generalization towards target test domain.

2. To address the problem, we propose a Domain-wise Active Acquisition (DAA) framework. By iteratively optimizing the data acquisition strategy as training samples are accumulated, it theoretically ensures the effective approximation of test distribution.

3. Extensive real-world experiments demonstrates the significance of training data acquisition in the context of distribution shift, and the advantages of our approach in data acquisition for better model performance.

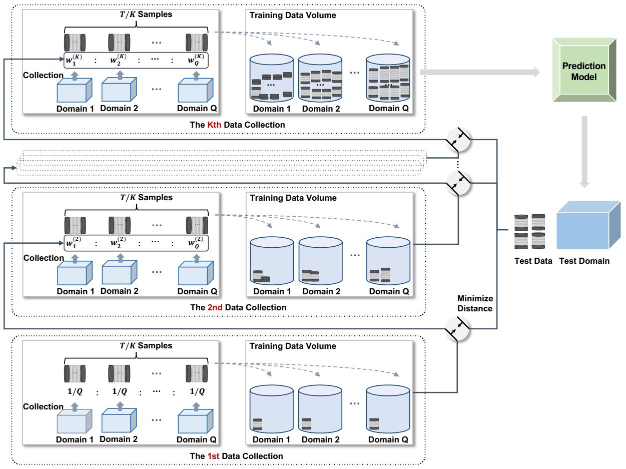


*Figure 1.* The framework of Domain-wise Active Acquisition. It iteratively optimizes the data acquisition strategy as training samples are accumulated, pursuing the approximation of test distribution.

## 2. Related Works

**Domain Adaptation** To generalize a model to the target domain with shifted data, instance-based domain adaptation methods (Jiang & Zhai, 2007; Dai et al., 2007) reweight or select representative training samples to mirror the test distribution; feature-based methods (Tzeng et al., 2014; Sun & Saenko, 2016) align or transfer features between training and test distributions; model-based methods (Motiian et al., 2017; Ganin et al., 2016) develop models either robust to domain shifts or customized for the test domain. Additionally, unsupervised domain adaptation is proposed to address the challenge of unlabeled test domain data. Typically, Zhang et al. (2018); Courty et al. (2017) transfer the distribution from source to target domains through sample reweighting. Self-supervised techniques (French et al., 2018; Deng et al., 2019) generate pseudo-labels for unlabeled data to expand the usable dataset. Another strategies (Taigman et al., 2017; Tzeng et al., 2017) involve using generative models to produce samples for the target domain, utilizing adversarial training to extract features aligned between the source and target domains, and etc. However, these approaches always suppose a predetermined training dataset, leading to their effectiveness heavily relying on the quality of training data.

**Data-centric Methodologies** Data-centric machine learning (Johnson & Khoshgoftaar, 2023; Patel et al., 2022; Mazumder et al., 2024; Schmarje et al., 2022) focuses on enhancing the quality of dataset and the way data is processed, rather than primarily optimizing the algorithm or model architecture. It improves the model performance in a sustainable and often more cost-effective manner, garnering increasing attention in recent years. Among data-centric methodologies, data acquisition (Chen et al., 2023; Agarwal et al., 2019) aims to obtain training data within limited budget to achieve more accurate and robust prediction performance. Yet, the utility of data acquisition in the context of distribution shifts remains underexplored. Addressing such shifts, active domain adaptation (Fu et al., 2021; Xie et al., 2023) selects highly representative samples from the test domain for annotation, boosting the model generalization ability. However, in real-world settings, labeling can be problematic and not always feasible through external annotators. To meet practical demands, this paper introduces a novel protocol designed to optimize training data acquisition to improve model performance in the test domain with distribution shift.

## 3. Preliminaries

**Notations** Throughout the paper, we use upper-case letters $X, Y$ to denote random variables or vectors, lower-case letters $x, y$ to denote their realizations/observations, bold capital letters $\mathbf{X}, \mathbf{Y}$ to denote the sets containing all the observations of $X, Y$.

### 3.1. Problem Statement

In this paper, we explore the impact of the collected training data on model generalization ability. Concretely, we investigate how to effectively attain the distribution of training data from multiple domains and thereby determine the data acquisition size of different domains to improve the model performance on the test data with distribution shift. The formal definition is shown in Problem 3.1.

**Problem 3.1** (Domain-wise Data Acquisition)**.** Given $Q$ available training domains with underlying data distributions $\{P_j^{tr}(X, Y)\}_{1\leq j\leq Q}$ and the total sample budget $T$, we aim to determine the data acquisition size between training domains $\{n_1, n_2, ..., n_Q\}$ with $\sum_{i=1}^{Q} n_i = T$, so that the trained prediction model from the acquired data can perform well on the test domain. The test domain is represented by unlabeled samples $\mathbf{X}_{test} = \{x_i^{test}\}_{i=1}^{M}$ drawn marginally from the distribution $P^{te}(X, Y)$.

*Remark* 3.2. To be clarified, we can decide only the data acquisition size for each domain, and the samples of each domain are sampled independently from the corresponding distribution in random. This is in distinction with the setting of traditional active learning.

Since the data heterogeneity widely presents in the real world, the data distributions differ among training domains $\{P_j^{tr}(X, Y)\}_{1\leq j\leq Q}$ and test domain $P^{te}(X, Y)$. To ensure the model trained from training domains being able to generalize to the test domain, we take studies in the scenarios that adhere the covariate shift assumption here.

**Assumption 3.3** (Covariate Shift)**.** The test domain differs from the training domains in the covariate distribution, which means $\forall 1 \leq j \leq Q, P_j^{tr}(X) \neq P^{te}(X)$. In contrast, the conditional distribution of label $P(Y|X)$ keeps consitent among all the domains, formally $\forall 1 \leq j \leq Q, P_j^{tr}(Y|X) = P^{te}(Y|X)$.

### 3.2. Preliminary Studies

#### 3.2.1. Empirical Risk Minimization

In the context under traditional IID hypothesis, Empirical Risk Minimization (ERM) aims to minimize the average empirical errors of training samples $\{(x_i, y_i)\}_{i=1}^{N}$. Given a loss function in form of $\mathcal{L}(x, y; \theta)$ (e.g. $(\theta^\top x - y)^2$ used in least square regression) where $\theta$ denotes the parameter set, ERM minimizes the following objective:

$$\min_{\theta} \mathbb{E}_{P^{tr}}[\mathcal{L}(X, Y; \theta)] \approx \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(x_i, y_i; \theta) \tag{1}$$

#### 3.2.2. Sample Reweighting

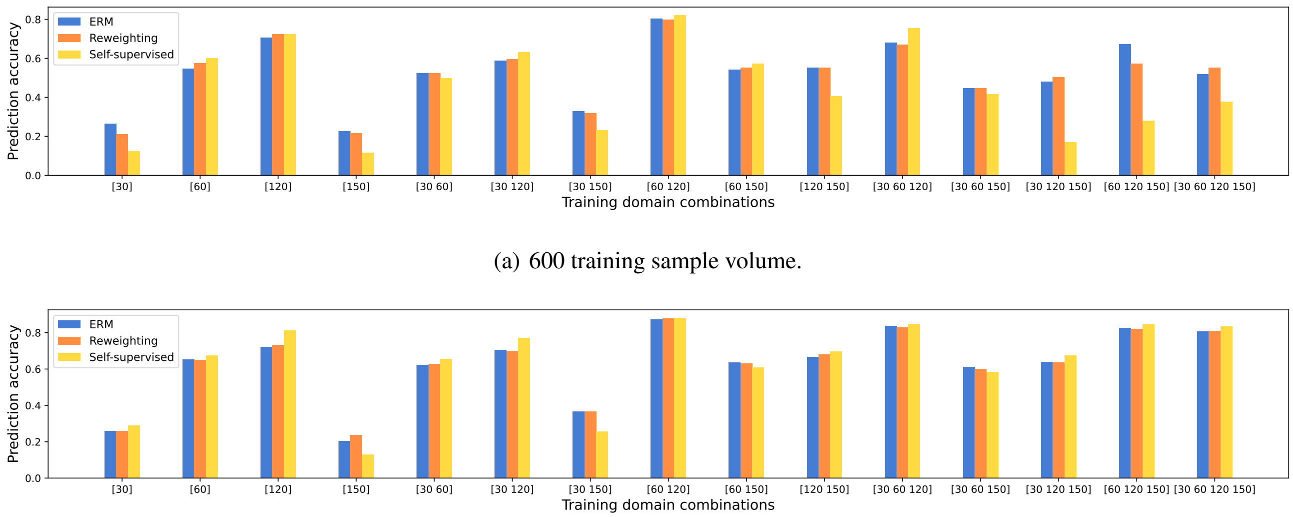
Under the distribution shift, sample reweighting based methods are proposed to align the distributions of training and test domain. The importance sampling weights (Farahani et al., 2020) of training samples are defined as $w(x, y) = \frac{P^{te}(x,y)}{P^{tr}(x,y)}$. A typical approach called density ratio estimation determines the sample weights by learning the probability of a sample belonging to either the test or training distributions in a framework of binary classification. After obtaining the sample weights, it can approximate the expectation of test risk using the weighted empirical errors of training data, then minimizes the following objective:

$$\begin{aligned}\min_{\theta} \mathbb{E}_{P^{te}}[\mathcal{L}(X, Y; \theta)] &= \int \frac{P^{te}(x, y)}{P^{tr}(x, y)} \mathcal{L}(x, y; \theta) P^{tr}(x, y) dx dy \\ &\approx \sum_{i=1}^{N} w(x_i, y_i) \mathcal{L}(x_i, y_i; \theta)\end{aligned} \tag{2}$$

#### 3.2.3. Self-supervised Learning

To utilize the unlabeled data $\mathbf{X}_{te}$ from test domain for training, self-supervised learning methods aim to generate pseudo-labels with the model automatically. After learning a prediction model from the training data $(\mathbf{X}_{tr}, \mathbf{Y}_{tr})$, self-supervised learning employs this model to impute pseudo-labels for unlabeled data. The labeled samples are then incorporated into the training process to fine-tune the model in turn. Through iterative optimization of imputed pseudo-label and prediction model, self-supervised learning enhances the model's adaptability to the test domain. Suppose $\hat{\mathbf{Y}}_{te}$ is the pseudo label of $\mathbf{X}_{te}$ in one iteration, it minimizes the following objective to finetune the prediction model, where $\psi$ is a hyper-parameter to balance the prediction loss of training data and imputed test data.

$$\min_{\theta} \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(x_i, y_i; \theta) + \psi \frac{1}{M}\sum_{i=1}^{M} \mathcal{L}(x_i^{te}, \hat{y}_i^{te}; \theta) \tag{3}$$



(a) 600 training sample volume.

(b) 1800 training sample volume.

*Figure 2.* The performance of model in test domain when it learns from different training data. We take 30, 60, 120, 150 degree rotations as the training domain candidates, 90 degree rotation as the test domain. The training data has a great influence on the model performance.

### 3.2.4. Motivation

To investigate the impact of training data acquisition on model performance under distribution shift, we conduct experimental analyses on the Rotated MNIST dataset (Ghifary et al., 2015). This dataset builds upon the original MNIST dataset (LeCun et al., 1998), which contains the handwritten digits ranging from 0 to 9, by introducing a modification: the images of digits are rotated by various angles. It simulates a real-world phenomenon where the orientation of objects varies from training to test phases.

In our study, we establish rotation angles of 30, 60, 120, and 150 degrees as the training domains, reserving the 90 degree rotation as the test domain. In different experiments, we uniformly acquire 600 or 1800 training samples from distinct combinations of domains among [30, 60, 120, 150] degrees, and get 100 unlabeled samples from 90 degree domain additionally. Subsequently, we employ the ERM, sample reweighting, and self-supervised learning methods based on MNIST_CNN[1] in Gulrajani & Lopez-Paz (2021) independently. In the test phase, we evaluate the prediction models using 600 samples from the 90-degree domain.

From the results in Figure 2, we find that:

1. The same model exhibits notably performance variations when trained in different domain combinations, indicating that acquiring distinct training samples substantially impacts the model performance under distribution shift.
2. The advanced model can enhance generalization ability, but cannot fully mitigate the difference in performance caused by training data acquisition. For instance, ERM model trained with the best domain combination [60, 120] exhibits better performance than self-supervised learning model trained with the second-best combination [30, 60, 120].
3. A lower data acquisition budget leads to a greater influence of training data on model performance, highlighting the necessity for advanced data acquisition.

Empirical evidence suggests that training data acquisition is closely concerned to model generalization performance, underscoring its importance in the context of distribution shift.

## 4. Algorithm

To achieve optimal data acquisition for test data with distribution shift, this paper proposes an algorithm, called Domain-wise Active Acquisition (DAA), that iteratively optimizes the data acquisition strategy from heterogeneous training domains under limited sample budget.

### 4.1. Domain-wise Active Acquisition

Since we do not have prior knowledge about the data distribution of each training domain at the initial stage, we split the whole data acquisition procedure into $K$ rounds and approximate the distribution empirically with the collected data in the previous rounds.

[1] https://github.com/facebookresearch/DomainBed

**Algorithm 1** Domain-wise Active Acquisition (DAA)

```
**Input:** $Q$ available training domains, unlabeled test samples $\mathbf{X}_{test}$, the total acquisition volume $T$, and the acquisition count $K$.
**First Iteration:** Uniformly collect $T/K$ samples from the training data domains, obtaining $T/(K \times Q)$ samples from $j^{th}$ domain, denoted as $\mathbf{X}_j^1$.
**for** $C = 2$ to $K$ **do**
    Calculate the average value $\{\bar{\phi}_j\}_{1 \le j \le Q}$ using the collected samples of the corresponding training domains $\{\bigcup_{i=1}^{C-1} \mathbf{X}^i\}_{1 \le j \le Q}$ , and the average value $\overline{\phi^{te}}$ of test domain using $\mathbf{X}_{test}$.
    Learn the domain weight $W = (w_j)_{1 \le j \le Q}$ based on $\{\bar{\phi}_j\}_{1 \le j \le Q}$ and $\overline{\phi^{te}}$ according to Equation (7).
    Collect $T/K$ samples from the training data domains following $W$, obtaining $(T \times w_j)/K$ samples from $j^{th}$ domain, denoted as $\mathbf{X}_j^C$.
**end for**
**return:** $T$ training samples $\{\bigcup_{i=1}^{K} \mathbf{X}_j^i\}_{1 \le j \le Q}$.
```

In the first round of data collection, we uniformly collect $n^1 = T/K$ samples from the training data domains, obtaining initial training samples $\{\mathbf{X}_j^1\}_{1 \le j \le Q}$ because of the absence of knowledge about data distribution. Specifically, it means that the domain-specific acquisition size satisfies $n_j^1 = T/(K \times Q)$ for each domain $j \in \{1, 2, 3, ..., Q\}$.

For the following rounds, we also equally assign total acquisition size $n^i = T/K, \forall i \in \{2, 3, 4, ...K\}$, which is same to the first round. However, since we have obtained the distribution information manifested by the collected samples before, we can adaptively determine the acquisition size of each domain for non-trivial performance improvement. Inspired by the motivation of sample reweighting technologies, we decide to reduce the distribution discrepancy between the acquired samples and test samples as our criteria for determining the domain-specific acquisition size.

To achieve this, we try to learn the domain weights $W = (w_i)_{1 \le i \le Q}$ so that the weighted combination of training domain distribution $P_w^{tr} = \sum_{j=1}^{Q} w_j \cdot P_j^{tr}$ is close to the test domain distribution $P^{te}$. By collecting samples with the domain-specific acquisition sizes $n_j^i$ proportional to the learned weight $w_j$, we can obtain a collection of samples resembling the test domain, which can lead the resulting trained model to perform better on test domain. Formally, our target for learning domain weights can be described as following:

$$W = \underset{W}{\operatorname{argmin}}\, Dist(P_w^{tr}, P^{te}), \tag{4}$$

$$s.t. \sum_{j=1}^{Q} w_j = 1, W \ge 0 \tag{5}$$

To instantiate this strategy, we choose Maximum Mean Discrepancy (MMD) distance (Iyer et al., 2014) between the average covariate vector $\bar{X}$ in this paper. It minimizes the following objective:

$$W = \underset{W}{\operatorname{argmin}} \left|\left|\mathbb{E}_{x \sim P_w^{tr}}[\phi(x)] - \mathbb{E}_{x \sim P^{te}}[\phi(x)]\right|\right|_2^2, \tag{6}$$

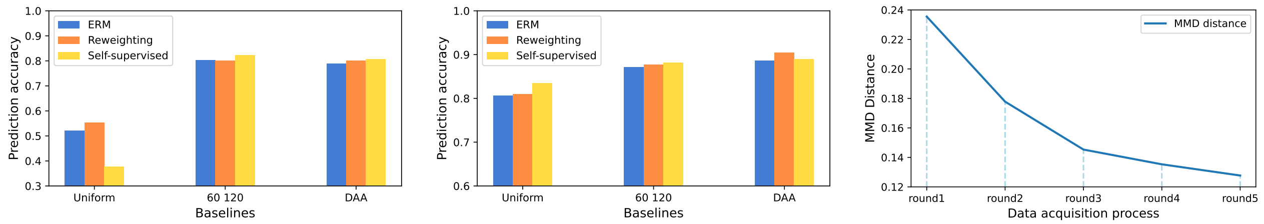
The feature map $\phi : \mathcal{X} \to \mathcal{H}$ is defined with the embedding space $\mathcal{H}$ of a reproducing kernel Hilbert space (RKHS) function.

In practice, we can empirically estimate the distribution distance in Equation 6 based on the previously acquired samples and learn the optimal domain weights for the current round. Specifically, with the collected samples of each training domain $\{\bigcup_{i=1}^{C} \mathbf{X}_j^i\}_{1 \le j \le Q}$ ($\mathbf{X}_j^i$ is the set of samples acquired in the $j^{th}$ training domain at the $i^{th}$ round) after $C$ rounds, we learn $W$ that minimizes the following function:

$$W = \underset{W}{\operatorname{argmin}} \left|\left| \sum_{j=1}^{Q} w_j \cdot \bar{\phi}_j - \overline{\phi^{te}} \right|\right|_2^2 + \lambda \cdot ||w||_1 \tag{7}$$

where $\bar{\phi}_j$ and $\overline{\phi^{te}}$ are respectively the average embedding of the collected samples in the $j^{th}$ training domain and test domain.

After calculating domain weights $W$, we acquire $n^{C+1} = \frac{T}{K}$ samples for the $(C+1)^{th}$ round from training domains with the domain-specific acquisition size $n_j^{C+1}$ proportional to the domain weight $w_j$. With the samples drawn randomly from the distribution of each training domain, we obtain the new samples of the current round $\{\mathbf{X}_j^{(C+1)}\}_{j=1}^{Q}$, where the sample sizes satisfy the condition $|\mathbf{X}_1^{(C+1)}| : \cdots : |\mathbf{X}_Q^{(C+1)}| = w_1 : \cdots : w_Q$. However, sampling bias in finite samples can easily lead to estimation errors for $w$. As a result, we iteratively optimize sampling strategies and acquire data until the total budget $T$ is exhausted. The framework of our method can be found in Figure 1. And we also list the pseudo code of algorithm in Algorithm 1.



(a) The model performance using different training data upon 600 sample volume.

(b) The model performance using different training data upon 1800 sample volume.

(c) MMD distance between training and test data with model convergence.

*Figure 3.* The experimental results in Rotated MNIST dataset. DAA can approach optimal data acquisition until it achieves convergence.

After the whole training data $\{\bigcup_{i=1}^{K}\mathbf{X}_j^i\}_{1\leq j\leq Q}$ are collected, we can learn the prediction models that can generalize to the test domain.

### 4.2. Theoretical Analysis

Due to the sampling error induced by finite samples, the estimated domain weights $W$ differ from the ideal counterpart $W^*$, which may degrade the performance of the trained model on the test domain. In this section, we theoretically analyze that as the number of acquired samples increases, this error of the estimated domain weights $W$ diminish.

For the ease of elaboration, we denote the expected value of covariate embedding in the $q^{th}$ training domain and test domain as $\eta_q$ and $\eta$ respectively. Formally, it can be formulated by

$$\eta_q = \mathbb{E}_{X\sim P_q^{tr}(X)}[\phi(X)], \quad \eta = \mathbb{E}_{X\sim P^{te}(X)}[\phi(X)].$$

Therefore, the ideal domain weights $W^*$ satisfies:

$$\sum_{q=1}^{Q} w_q^*\eta_q = \eta, \quad \sum_{q=1}^{Q} w_q^* = 1.$$

However, due to sampling limitation, we can only use $\bar{\phi}_q = (n_q)^{-1}\sum_{i\in\mathcal{I}_q}\phi(X_i)$ to empirically approximate $\eta_q$ ($\mathcal{I}_q$ is the index set of the collected samples in $q^{th}$ training domain). Correspondingly, the actual learned domain weight $W$ satisfies the following conditions:

$$\sum_{q=1}^{Q} w_q\bar{\phi}_q = \eta, \quad \sum_{q=1}^{Q} w_q = 1.$$

Intuitively, more samples in one domain indicate that the empirically estimated $\bar{\phi}_q$ is a more accurate estimation of the embedding estimate of $\eta_q$. Consequently, our estimated domain weights $W$ will be more closer to the ideal weights $W^*$. This is theoretically proved in the following theorem.

**Theorem 4.1.** *Denote $W = (w_1, \ldots, w_Q)^{\mathrm{T}}$ and $W^* = (w_1^*, \ldots, w_Q^*)^{\mathrm{T}}$, and $||\cdot||$ is any norm. We assume $A$ and $\Delta A$ as the coefficient matrix and the perturbation of the matrix respectively as follows:*

$$A = \begin{bmatrix} \eta_1 & \eta_2 & \ldots & \eta_Q \\ 1 & 1 & \ldots & 1 \end{bmatrix},$$

$$\Delta A = \begin{bmatrix} \bar{\phi}_1 - \eta_1 & \bar{\phi}_2 - \eta_2 & \ldots & \bar{\phi}_Q - \eta_Q \\ 0 & 0 & \ldots & 0 \end{bmatrix}.$$

*When $||A^{-1}||\cdot||\Delta A|| < 1$, $V_q = \mathrm{var}_{X\sim P_q^{tr}(X)}[\phi(X)] < \infty$, we have*

$$||W - W^*|| \leq \frac{||A^{-1}||}{1 - ||A^{-1}||\cdot||\Delta A||}\cdot\left|\left|\sum_{q=1}^{Q} w_q^*(\bar{\phi}_q - \eta_q)\right|\right|.$$

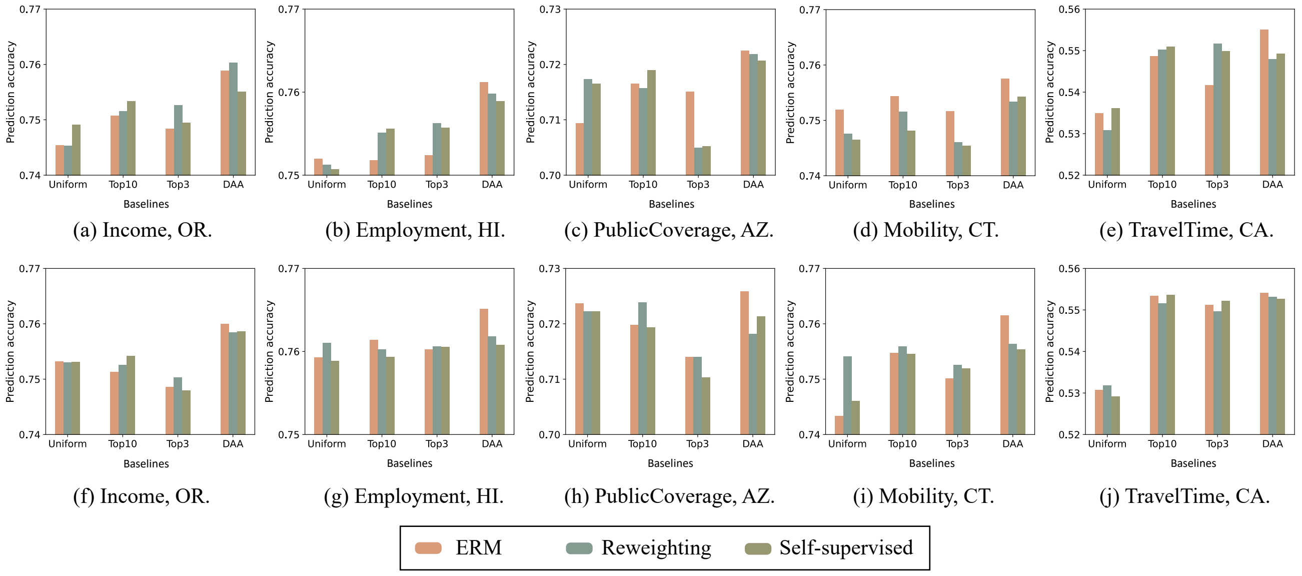
*Furthermore, when $\min_q n_q \to \infty$,*

$$\sum_{q=1}^{Q} w_q^*(\bar{\phi}_q - \eta_q) \xrightarrow{d} \mathcal{N}\left(0, \sum_{q=1}^{Q}\frac{(w_q^*)^2}{n_q}V_q\right).$$

*Remark* 4.2. When the sampling noise is homogeneous, i.e. $V_q = V$, then the asymptotic variance $\sum_{q=1}^{Q}(w_q^*)^2/n_q V \geq V/n$, when $n_q/n = w_q^*$ obtain the optimal asymptotic variance. When the sampling noise is heterogeneous, for any unit vector $e$, we have that $e^{\mathrm{T}}(\sum_{q=1}^{Q}(w_q^*)^2/n_q V_q)e \geq (\sum_{q=1}^{Q} w_q^*\sqrt{e^{\mathrm{T}}V_q e})^2/n$, when $n_q \propto w_q^*\sqrt{e^{\mathrm{T}}V_q e}$ obtain the optimal asymptotic variance. The optimal direction depends on the direction $e$, but they have the same coefficient $w_q$.

According to Theorem 4.1, we find that:

- As training samples accumulated per domain increases, the sampling error in each domain diminishes, resulting in a more accurate estimation of the optimal $W^*$.
- The closer the proportion of samples collected from each environment already is to the optimal $W^*$, the more accurate the following estimation of the optimal $W^*$ will become.



(a) Income, OR. (b) Employment, HI. (c) PublicCoverage, AZ. (d) Mobility, CT. (e) TravelTime, CA.

(f) Income, OR. (g) Employment, HI. (h) PublicCoverage, AZ. (i) Mobility, CT. (j) TravelTime, CA.

*Figure 4.* The experimental results in ACS PUMS dataset. The first row (a-e) and the second row (f-j) presents the results upon 1000 sample volume, and 2000 sample volume, respectively. DAA achieves the best results in almost all cases.

Therefore, after acquiring $T \times (C-1)/K$ samples through $(C-1)$ rounds, our approach can effectively obtain the data acquisition strategy manifested by the domain weights $W^C$ at the $C^{th}$ round. Consequently, the proportion of accumulated samples after $C$ rounds approaches $W_*$ closer, thereby enabling a better data acquisition strategy $W^{(C+1)}$ at the $(C+1)^{th}$ round, which correspondingly more closely approximates the optimal $W_*$. As a result, our proposed DAA approach can effectively approach the optimal data collection through iteratively optimizing the data acquisition strategy with domain weights $W$ and acquiring training samples.

## 5. Experiments

In this section, we validate the effectiveness of our proposal towards better data acquisition under distribution shift.

**Baselines** Given a total budget $T$ and $Q$ training domains, the traditional data acquisition approach involves uniformly collecting T samples across the Q domains (i.e., $\frac{T}{Q}$ samples per domain). We refer to this as uniform acquisition. Further, an optimized approach initially collects partial data uniformly from $Q$ domains, and then utilizing unlabeled test domain data to calculate the distribution distance between training and test domains. Thereby, the closest Top-$\mathcal{K}$ training domains are selected to uniformly collect the remaining samples. On the other hand, our approach DAA achieves optimal data acquisition through iterative optimization of acquiring strategy towards the test domain. Once the data acquisition process finishes, we obtain the prediction model by employing ERM, sample reweighting, and self-supervised learning methods, respectively. We report the prediction accuracy of the model in the test domain.

### 5.1. Rotated MNIST

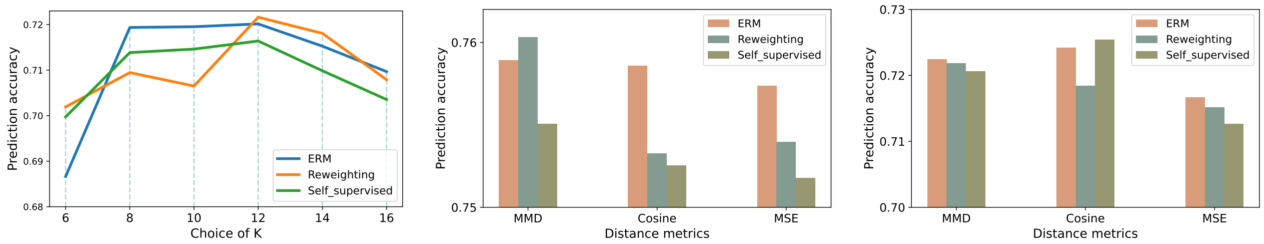
#### 5.1.1. Experimental Result

Following the experimental setting in Section 3.2.4, we evaluate our method with in 600 and 1800 training data volume, respectively. From the results in Figure 3(a) and 3(b), we can see that:

- Without exhaustive searching, our method can effectively select appropriate domains for sample acquisition, significantly enhancing the model's performance on test domain compared to uniform acquisition.
- Compared to the best domain combination [60, 120] in Figure 2, our method can even achieve further improvements, benefiting from our strategic data acquisition across different domains.

#### 5.1.2. convergence process

Figure 3(c) shows the variation in MMD distance between accumulated training data and test data throughout the data acquisition process. As shown, their distance decreases gradually as the number of data acquisition rounds increases until convergence, indicating that our method can effectively mirror the test domain.



(a) The results of DAA based on different numbers of acquisition rounds.

(b) The results of DAA based on various distance measures on Income task.

(c) The results of DAA based on various distance measures on PublicCoverage task.

*Figure 5.* The model analysis upon ACS PUMS dataset, in terms of different distance measures and acquisition rounds.

## 5.2. US-Wide ACS PUMS Data

### 5.2.1. Experimental Setting

Ding et al. (2021) [2] creates a new tabular dataset from US Census sources to enhance research on algorithmic fairness, as well as distribution shift problems. It introduces five prediction tasks based on the American Community Survey (ACS) Public Use Microdata Sample (PUMS) Data. More details of five tasks are provided in Appendices.

ACS PUMS Data comes from US 50 states. In each task, we randomly choose one state as the test domain, and take other states as the training domains. We conduct experiments with total training data volume of 1000 and 2000, respectively, and get 100 unlabeled samples from test domain. In the test phase, we use 6000 samples from test domain to verify the model performance.

### 5.2.2. Experimental Result

From the results in Figure 4, we can observe that:

- The impact of dataset quality on model performance varies across various tasks and test domains, yet it consistently holds the critical importance. The crude uniform acquisition method often results in the worst outcomes, especially when the acquisition budget is small.
- The Top-$\mathcal{K}$ approach is sensitive to the number of domain selections. It can pinpoint the beneficial training domains, thus generally surpassing the uniform acquisition method. However, it may sometimes perform much worse due to its lack of consideration for the synergistic effects of multi-domains, potentially leading to a biased selection.
- In contrast, our method, by effectively approximating the test domains through active acquiring, can significantly improve the model's test performance in almost all cases.
- Optimizations at the model level can improve the model's generalization capability, but they fall short of compensating for the impact of discrepancies in the training data. Moreover, the ERM model can achieve exceptionally high performance when the training data is of adequate quality.

### 5.2.3. Impact of acquiring rounds

Figure 5(a) exhibits the model performances when training samples are acquired over different numbers of rounds given fixed total acquisition volume $T$. If the rounds are few, the uniform acquisition in the first round comprises large proportion and may hinder the model performance; Conversely, excessive rounds result in few samples at each round, and potentially introduce the uncertainty deteriorating the domain weight learning.

### 5.2.4. Comparison of distance metrics

In addition to MMD distance, we explore the effect of various distances on data acquisition, including mean square error (MSE) distance and cosine similarity between average covariate vector. As shown in Figure 5(b) and 5(c), the performance of MMD and cosine are competitive, with a slight improvement over MSE. A solid distance metric might better assist our method in approaching the test domain.

## 5.3. Functional Map of the World (FMOW) Dataset

### 5.3.1. Experimental Setting

Christie et al. (2018) compiles satellite images to promote research on addressing the dynamic nature of satellite data shaped by environmental changes. Each image is categorized by building or land use type, with its domain indicating the capture year and geographical region. In experiments, we consider the setting where the shift appears in regions. We use 19 countries[3] covering 4 continents as the training

[2] https://github.com/zykls/folktables

[3] countries include: USA RUS CHN ITA BRA GBR JPN AUS TUR DEU ESP UKR CAN ARG CHL IND MEX NLD PHL.

*Table 1.* The experimental results in FMOW dataset. E, R, S denote ERM, Reweighting and Self-supervised learning model, respectively.

| Method | 2000 sample volume | | | 4000 sample volume | | | 6000 sample volume | | |
|---|---|---|---|---|---|---|---|---|---|
| | E | R | S | E | R | S | E | R | S |
| Uniform | 83.08% | 84.49% | 83.96% | 85.23% | 84.56% | 86.28% | 86.51% | 87.66% | 87.47% |
| Top-10 | 84.31% | 84.82% | 84.64% | 86.01% | 85.28% | 86.74% | 86.84% | 87.98% | 88.09% |
| Top-3 | 84.33% | 82.90% | 84.46% | 86.39% | 85.67% | 86.82% | 86.69% | 87.00% | 87.35% |
| DAA | **84.85**% | **85.10**% | **85.28**% | **86.84**% | **86.25**% | **87.35**% | **87.51**% | **88.25**% | **88.11**% |

*Table 2.* Important domains selected by approaches upon 4000 sample volume. Each country's data size locates in ().

| | |
|---|---|
| Uniform | Each of 19 countries (210) |
| Top-10 | USA (362) RUS (362) ITA (362) AUS (362) TUR (362) UKR (362) CAN (362) CHL (362) NLD (362) PHL (362) Other country (42) |
| Top-3 | ITA (1110) CHL (1110) NLD (1110) Other country (42) |
| DAA | USA (116) RUS (106) CHN (60) ITA (1539) BRA (95) GBR (87) JPN (45) AUS (43) TUR (149) DEU (74) ESP (76) UKR (136) CAN (43) ARG (613) CHL (44) IND (169) MEX (44) NLD (241) PHL (320) |

domains, and take France as the test domain. In the training phase, we acquire 2000, 4000 or 6000 samples from training domains, respectively, and get 100 unlabeled samples from test domain. In the test phase, we use 12000 test samples to examine the models under region shift. We utilize ResNet-50 as the backbone of prediction models.

#### 5.3.2. Experimental Result

Table 1 reports the model performance upon different data acquisition approaches. Strategic data acquisition, as opposed to uniform acquisition, brings about a noticeable enhancement in performance. Our method intuitively identifies key data acquisition domains and their combinations, optimally aiding in improving the model generalization performance. leading to its consistent superior results.

#### 5.3.3. Interpretability

Furthermore, we show the important domains selected by different approaches in Table 2. Compared to others, our approach not only concerns countries across Europe but also focuses more on Italy, which is most closely related to France. Moreover, it is capable of uncovering relevant countries in other continents, such as Argentina, which features many buildings of French architectural style.

## 6. Conclusion

In this paper, we highlight the crucial challenge of employing advanced data acquisition from diverse domains to improve model performance under distribution shift. To tackle this issue, we present the Domain-wise Active Acquisition (DAA) framework, a novel approach that iteratively refines the data acquisition strategy to theoretically enable a model to more accurately mirror the test distribution. Our extensive real-world experiments demonstrate the critical role of developing data acquisition methodology to copy with the distribution shift and the superiority of our proposal in enhancing model generalization to target test domain.

## Acknowledgements

This work was supported in part by National Natural Science Foundation of China (No. 62141607), China National Postdoctoral Program for Innovative Talents (BX20230195). All opinions in this paper are those of the authors and do not necessarily reflect the views of the funding agencies. We would like to all the anonymous reviewers for the helpful feedback.

## Impact Statement

This paper presents work aimed at improving the performance of machine learning models by adopting a data-centric approach to address real-world data distribution shifts encountered in open-environment deployments. We identify no inherent ethical issues within our research. Additionally, there are several potential societal consequences of our work, including enhancing the generalization capabilities of machine learning models when facing test domains, ensuring the reliability of model applications; optimizing the data acquisition process, achieving optimal resource allocation, thereby increasing the economic efficiency of training machine learning models.

## References

Agarwal, A., Dahleh, M. A., and Sarkar, T. A marketplace for data: An algorithmic solution. In Karlin, A., Immorlica, N., and Johari, R. (eds.), *Proceedings of the 2019 ACM Conference on Economics and Computation, EC 2019, Phoenix, AZ, USA, June 24-28, 2019*, pp. 701–726. ACM, 2019. doi: 10.1145/3328526.3329589.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *CoRR*, abs/1907.02893, 2019.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175, 2010. doi: 10.1007/S10994-009-5152-4.

Chen, L., Acun, B., Ardalani, N., Sun, Y., Kang, F., Lyu, H., Kwon, Y., Jia, R., Wu, C., Zaharia, M., and Zou, J. Data acquisition: A new frontier in data-centric AI. *CoRR*, abs/2311.13712, 2023. doi: 10.48550/ARXIV.2311.13712.

Christie, G. A., Fendley, N., Wilson, J., and Mukherjee, R. Functional map of the world. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 6172–6180. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00646.

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. Joint distribution optimal transportation for domain adaptation. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 3730–3739, 2017.

Dai, W., Yang, Q., Xue, G., and Yu, Y. Boosting for transfer learning. In Ghahramani, Z. (ed.), *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, volume 227 of *ACM International Conference Proceeding Series*, pp. 193–200. ACM, 2007. doi: 10.1145/1273496.1273521.

Deng, Z., Luo, Y., and Zhu, J. Cluster alignment with a teacher for unsupervised domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 9943–9952. IEEE, 2019. doi: 10.1109/ICCV.2019.01004.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring adult: New datasets for fair machine learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 6478–6490, 2021.

Farahani, A., Voghoei, S., Rasheed, K., and Arabnia, H. R. A brief review of domain adaptation. *CoRR*, abs/2010.03978, 2020.

French, G., Mackiewicz, M., and Fisher, M. Self-ensembling for visual domain adaptation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Fu, B., Cao, Z., Wang, J., and Long, M. Transferable query selection for active domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 7272–7281. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00719.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. S. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17:59:1–59:35, 2016.

Ghifary, M., Kleijn, W. B., Zhang, M., and Balduzzi, D. Domain generalization for object recognition with multi-task autoencoders. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 2551–2559. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.293.

Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Hoffman, J., Tzeng, E., Darrell, T., and Saenko, K. Simultaneous deep transfer across domains and tasks. In Csurka, G. (ed.), *Domain Adaptation in Computer Vision Applications*, Advances in Computer Vision and Pattern Recognition, pp. 173–187. Springer, 2017. doi: 10.1007/978-3-319-58347-1_9.

Iyer, A. S., Nath, J. S., and Sarawagi, S. Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pp. 530–538. JMLR.org, 2014.

Jiang, J. and Zhai, C. Instance weighting for domain adaptation in NLP. In Carroll, J., van den Bosch, A., and Zaenen, A. (eds.), *ACL 2007, Proceedings of the 45th*

*Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics, 2007.

Johnson, J. M. and Khoshgoftaar, T. M. Data-centric AI for healthcare fraud detection. *SN Comput. Sci.*, 4(4):389, 2023. doi: 10.1007/S42979-023-01809-X.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Li, L. and Zhang, Z. Semi-supervised domain adaptation by covariance matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(11):2724–2739, 2019. doi: 10.1109/TPAMI.2018.2866846.

Mazumder, M., Banbury, C., Yao, X., Karlaš, B., Gaviria Rojas, W., Diamos, S., Diamos, G., He, L., Parrish, A., Kirk, H. R., et al. Dataperf: Benchmarks for data-centric ai development. *Advances in Neural Information Processing Systems*, 36, 2024.

Motiian, S., Piccirilli, M., Adjeroh, D. A., and Doretto, G. Unified deep supervised domain adaptation and generalization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 5716–5726. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.609.

Patel, H., Guttula, S. C., Mittal, R. S., Manwani, N., Berti-Équille, L., and Manatkar, A. Advances in exploratory data analysis, visualisation and quality for data centric AI systems. In Zhang, A. and Rangwala, H. (eds.), *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pp. 4814–4815. ACM, 2022. doi: 10.1145/3534678.3542604.

Schmarje, L., Grossmann, V., Zelenka, C., Dippel, S., Kiko, R., Oszust, M., Pastell, M., Stracke, J., Valros, A., Volkmann, N., and Koch, R. Is one annotation enough? - A data-centric image classification benchmark for noisy and ambiguous label estimation. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., and Cui, P. Towards out-of-distribution generalization: A survey. *CoRR*, abs/2108.13624, 2021.

Sun, B. and Saenko, K. Deep CORAL: correlation alignment for deep domain adaptation. In Hua, G. and Jégou, H. (eds.), *Computer Vision - ECCV 2016 Workshops - Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*, volume 9915 of *Lecture Notes in Computer Science*, pp. 443–450, 2016. doi: 10.1007/978-3-319-49409-8_35.

Taigman, Y., Polyak, A., and Wolf, L. Unsupervised cross-domain image generation. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 2962–2971. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.316.

Xie, B., Yuan, L., Li, S., Liu, C. H., Cheng, X., and Wang, G. Active learning for domain adaptation: An energy-based approach. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 8708–8716. AAAI Press, 2022. doi: 10.1609/AAAI.V36I8.20850.

Xie, M., Li, S., Zhang, R., and Liu, C. H. Dirichlet-based uncertainty calibration for active domain adaptation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Zhang, J., Ding, Z., Li, W., and Ogunbona, P. Importance weighted adversarial nets for partial domain adaptation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 8156–8164. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00851.

# Appendices

## A. Proof of Theorem 4.1

*Proof.* Denote $W = (w_1, \dots, w_Q)^{\mathrm{T}}$ and $W^* = (w_1^*, \dots, w_Q^*)^{\mathrm{T}}$, and $||\cdot||$ is any norm. When $||\Delta A||$ is small, satisfying $||A^{-1}|| \cdot ||\Delta A|| < 1$, then $||A^{-1}\Delta A|| \leq ||A^{-1}|| \cdot ||\Delta A|| < 1$. Therefore, $I + A^{-1}\Delta A$ is full rank, then $A + \Delta A = A(I + A^{-1}\Delta A)$ has full rank. Therefore,

$$\begin{aligned}||W - W^*|| &= ||(I + A^{-1}\Delta A)^{-1} A^{-1}(\Delta A W^*)|| \\ &\leq \frac{||A^{-1}||}{1 - ||A^{-1}|| \cdot ||\Delta A||} \cdot ||\Delta A \cdot W^*|| \\ &\leq \frac{||A^{-1}||}{1 - ||A^{-1}|| \cdot ||\Delta A||} \cdot ||\sum_{q=1}^{Q} w_q^*(\bar{\phi}_q - \eta_q)||.\end{aligned}$$

When $\min n_q \to \infty$, since $V_q < \infty$, by central limit theorem,

$$\sqrt{n_q}(\bar{\phi}_q - \eta_q) \xrightarrow{d} \mathcal{N}(0, V_q),$$

then we have that

$$\sum_{q=1}^{Q} w_q^*(\bar{\phi}_q - \eta_q) \xrightarrow{d} \mathcal{N}\left(0, \sum_{q=1}^{Q} \frac{(w_q^*)^2}{n_q} V_q\right).$$

□

## B. Description of ACS PUMS Data

The descriptions of 5 tasks predefined in ACS PUMS Data (Ding et al., 2021) [4] are follows.

- ACSIncome Task: The outcome variable is whether an individual's income is above $50,000$, and each sample contains features in form of 10 dimensional vector.
- ACSPublicCoverage Task: The outcome variable is whether an individual is covered by public health insurance, and each sample contains features in form of 20 dimensional vector.
- ACSMobility Task: The outcome variable is whether an individual had the same residential address one year ago, and each sample contains features in form of 21 dimensional vector.
- ACSEmployment Task: The outcome variable is whether an individual is employed, and each sample contains features in form of 16 dimensional vector.
- ACSTravelTime Task: The outcome variable is whether an individual has a commute to work that is longer than 20 minutes, and each sample contains features in form of 16 dimensional vector.

[4] https://github.com/zykls/folktables