Proxy Methods for Domain Adaptation

Katherine Tsai
University of Illinois Urbana-Champaign

Stephen R. Pfohl
Google Research

Olawale Salaudeen
University of Illinois Urbana-Champaign

Nicole Chiou
Stanford University

Matt J. Kusner
University College London

Alexander D'Amour
Google DeepMind

Sanmi Koyejo
Google DeepMind
Stanford University

Arthur Gretton
Google DeepMind
Gatsby Computational Neuroscience Unit

Abstract

We study the problem of domain adaptation under distribution shift, where the shift is due to a change in the distribution of an unobserved, latent variable that confounds both the covariates and the labels. In this setting, neither the covariate shift nor the label shift assumptions apply. Our approach to adaptation employs proximal causal learning, a technique for estimating causal effects in settings where proxies of unobserved confounders are available. We demonstrate that proxy variables allow for adaptation to distribution shift without explicitly recovering or modeling latent variables. We consider two settings, (i) **Concept Bottleneck**: an additional "concept" variable is observed that mediates the relationship between the covariates and labels; (ii) **Multi-domain**: training data from multiple source domains is available, where each source domain exhibits a different distribution over the latent confounder. We develop a two-stage kernel estimation approach to adapt to complex distribution shifts in both settings. In our experiments, we show that our approach outperforms other methods, notably those which explicitly recover the latent confounder.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1 Introduction

The goal of domain adaptation is to transfer an accurate model from a labeled *source* domain to an unlabeled *target* domain, which has a different but related distribution (Pan et al., 2010; Koh et al., 2021; Malinin et al., 2021). It is motivated by the fact that labeling data is often labor intensive, and sometimes requires domain expertise. For example, the distribution of patients diagnosed with a condition from hospital A and hospital B may differ due to patients' socioeconomic status, demographics, and other factors. However, labeled data might be only be available at hospital A and not at hospital B (e.g., due to less funding). As a result, an accurate model for patients from hospital A may perform poorly for patients from hospital B.

In order to provide guarantees on the accuracy of a transferred model, one of two classical assumptions have been made: *label shift* or *covariate shift*. Label shift (Buck et al., 1966; Lipton et al., 2018) assumes that the distribution of a label $P(Y)$ shifts between source and target domains, but the conditional distribution $P(X \mid Y)$ does not. Conversely, covariate shift (Shimodaira, 2000) assumes that the covariate distribution $P(X)$ shifts between domains, but the distribution $P(Y \mid X)$ stays the same. Each assumption provides theoretical guarantees on the generalization of a transferred classifier. In fact, without any assumptions, the source and target domains could differ arbitrarily, making guarantees impossible. However, these assumptions are often too restrictive to apply in real-world settings (Zhang et al., 2015; Schrouff et al., 2022). For instance, if covariates X and labels Y are confounded by a third variable U, it is possible for neither $P(X \mid Y)$ or $P(Y \mid X)$ to be equal across domains. For example, demographic information U could

confound the relationship between a diagnosis Y and a radiological image X. In this example, if two hospitals have different distributions over demographics, both label shift and covariate shift adaptation methods will fail to transfer a classifier across hospitals.

To address this, recent work has introduced a *latent shift* assumption: the distribution of U, an unobserved latent confounder of X and Y, shifts between the source and target domain (Alabdulmohsin et al., 2023). In this setting, all distributions of X and Y (without conditioning on U) may differ across the domains, violating label and covariate shift assumptions.

Contributions. We propose techniques for domain adaptation under the latent shift assumption that are guaranteed to identify the optimal predictor $\mathbb{E}[Y \mid x]$ in the target domain. We make use of proxy methods (Miao et al., 2018), which are a recently developed framework for causal effect estimation in the presence of a hidden confounder U, given indirect proxy information on U. Compared to prior work (Alabdulmohsin et al., 2023), our techniques do not require: identifying the distribution of the latent variable U, that U be discrete, or further linear independence assumptions. We consider two settings: (1) **Concept Bottleneck**: we observe in both domains a *proxy* W of the unobserved confounder U and a *concept* C that mediates the direct relationship between X and Y (Alabdulmohsin et al., 2023), or (2) **Multi-Domain**: we do not observe C in either domain, but have access to observations from multiple source domains. For both settings, we provide guarantees for identifying $\mathbb{E}[Y \mid x]$ without observing Y in the target domain. When $\mathbb{E}[Y \mid x]$ is identifiable, we develop practical two-stage kernel estimators to perform adaptation. The code is available at `https://github.com/koyejo-lab/ProxyDA`.

2 Related Work

The development of techniques for learning robust models and adapting to distribution shift has a long history in machine learning, but recently has received increased attention (Shen et al., 2021; Zhou et al., 2022; Wang et al., 2022).

Causality for domain adaptation. Our work is inspired by techniques that formulate the covariate/label shift settings as assumptions on the causal structure for domain adaptation and distributional robustness (*e.g*, Schölkopf et al. (2012); Peters et al. (2015); Zhang et al. (2015); Subbaswamy et al. (2019); Rothenhäusler et al. (2021); Veitch et al. (2021); Magliacane et al. (2018); Arjovsky et al. (2019); Ganin et al. (2016); Ben-David et al. (2010); Oberst et al. (2021)).

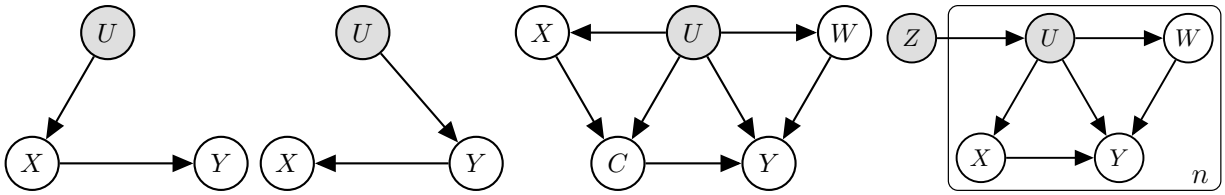
Proximal causal inference. Our identification technique is inspired by approaches used to identify causal effects with unobserved confounding with observed proxies (Kuroki and Pearl, 2014; Miao et al., 2018; Deaner, 2018; Tchetgen et al., 2020; Mastouri et al., 2021; Cui et al., 2023; Xu and Gretton, 2023). These approaches design 'bridge functions' to connect quantities involving a proxy W with those of the label Y. The beauty of this approach is that these bridge functions are implicitly a marginalization over U. This allows these approaches to identify causal quantities without identifying distributions involving U.

Latent shift. Our work is most closely related to Alabdulmohsin et al. (2023), who introduced the setting of latent shift with proxies W and concepts C. They showed that the optimal predictor $\mathbb{E}[Y \mid x]$ is identifiable in the target domain if W and C are observed in the source domain and X is observed in the target domain. To do so, they required (a) identification of distributions involving U, (b) that U is a discrete variable, (c) knowledge of the dimensionality of U, and (d) additional linear independence assumptions. In contrast, our work derives identification results for arbitrary U, and does not require any of (a)-(d). However, there is no free lunch: to achieve this, we require that proxies W are observed in the target, and either that: (i) concepts C are also observed in the target, or (ii) we observe multiple source domains. For (ii) we do not require C in either the source or the target, but for full identification we require that U is discrete.

3 Problem Framework

Let $P(\cdot)$ and $Q(\cdot)$ denote the probability distribution functions of the source domain and target domain, respectively. Let p and q indicate source and target quantities. Our goal is to study identification and estimation of the optimal target predictor $\mathbb{E}_q[Y \mid x]$ when Y is not observed in the target domain.

Concept Bottleneck. The first setting we study is described by the graph in Figure 1c. We have two additional variables: (i) proxies W, which provide auxiliary information about U, or can be seen as a noisy version of it (Kuroki and Pearl, 2014), and (ii) concepts C, which mediate or 'bottleneck' the relationship between the covariates X and labels Y (Goyal et al., 2019; Koh et al., 2020). For example, Koh et al. (2020) describe a setting where the concepts C are high-level clinical and morphological features of a knee X-ray X, which mediate the relationship with osteoporosis severity Y. In this example, U could describe demographic variations that alter symptoms X, C and outcome Y, and the proxies W could include patient background and clinical history (e.g., prior diagnoses, medications, procedures, etc). For the source domain we assume we observe $(X, C, W, Y) \sim P$ and for the target domain



(a) Covariate shift (b) Label shift (c) Concept Bottleneck shift (d) Multi-Domain shift

Figure 1: Causal diagrams. The shaded circle denotes unobserved variable and the solid circle denotes observed variable. X is the covariate, Y is the response, C is the concept, W is the proxy, Z is the domain-related variable, and U is the latent variable.

we observe $(X, C, W) \sim Q$.

We formalize the notion of latent shift, as introduced in Alabdulmohsin et al. (2023).

Assumption 1 (Concept Bottleneck, Alabdulmohsin et al. (2023))**.** *The shift between P and Q is located in unobserved U, i.e., there is a latent shift $P(U) \neq Q(U)$, but $P(V \mid U) = Q(V \mid U)$, where $V \subseteq \{W, X, C, Y\}$.*

This assumption states that every variable conditioned on U is invariant across domains. However, as $P(U) \neq Q(U)$, none of the marginal distributions are: $P(V) \neq Q(V)$ for $V \subseteq \{W, X, C, Y\}$. This assumption is a generalization of covariate shift $P(Y \mid X, U) = Q(Y \mid X, U)$ (Shimodaira, 2000) and label shift $P(X \mid Y, U) = Q(X \mid Y, U)$ (Buck et al., 1966), with associated graphs in Figure 1a–1b.

Assumption 2 (Structural assumption)**.** *Graphs in Figure 1 are faithful and Markov (Spirtes et al., 2000).*

Under Assumption 2, we have the following conditional independence properties for the graph in Figure 1c:

$$Y \perp\!\!\!\perp X \mid \{U, C\}, \quad W \perp\!\!\!\perp \{X, C\} \mid U.$$

With this conditional independence structure, $\{U, C\}$ blocks the information from X to Y and U blocks the information flow from W to $\{X, C\}$. We will see in Section 4 that these assumptions allow us to obtain $Q(Y \mid x)$ from $Q(W, C \mid x)$ in the target domain, where the latter is a function of observed quantities.

Multi-domain. In the second setting, suppose we do not observe the concepts C in any domain, but instead observe data from multiple source domains, according to the graph in Figure 1d. For instance, we may want to learn a classifier for a target hospital that has only unlabelled data, using data from several source hospitals with labelled data. Here, let Z be a random variable in $\mathcal{Z}$ denoting a prior over the source domains, and let $P(U|Z)$ be the distribution of U given Z. We make k_Z draws from Z, indexed by $r \in \{1, \ldots, k_Z\}$, and write $\{z_1, \ldots, z_{k_Z}\} =: \mathcal{Z}_p \subseteq \mathcal{Z}$. For each source domain z_r, we observe $(X, W, Y) \sim P(X, W, Y|z_r) := P_r(X, W, Y)$. For the target, we denote it with index $k_Z + 1$ and only observe $(X, W) \sim P(X, W|z_{k_Z+1}) := Q(X, W)$. In general let $P_r(V) := P(V|z_r)$ and $Q(V) := P(V|z_{k_Z+1})$ for any $V \subseteq \{W, X, Y, U\}$. For this setting we replace Assumption 1 with the following shift assumption.

Assumption 3 (Multi-Domain)**.** *For each $z, z' \in \mathcal{Z}_p$ such that $z \neq z'$, we have $P(U|z) \neq P(U|z') \neq Q(U)$.*

Note that Assumption 2 implies the following the conditional independence property in Figure 1d:

$$\{Y, X, W\} \perp\!\!\!\perp Z \mid U.$$

Note that under Assumption 3, we allow all joint distributions to be different $P(W, X, U, Y|z) \neq P(W, X, U, Y|z') \neq Q(W, X, U, Y)$ for $z \neq z' \in \mathcal{Z}_p$.

4 Identification under Latent Shifts

Our identification techniques are inspired by proximal causal inference (Tchetgen et al., 2020). The key idea is to design so-called "bridge" functions to identify distributions confounded by unobserved variables. We first show that with additional proxies and concepts, $\mathbb{E}_q[Y \mid x]$ is identifiable under any latent shift.

4.1 Identification with Concepts

To prove identifiability, we need certain assumptions to hold for the shift. The first is a regularity assumption, also known as a completeness condition, and is commonly used to identify causal estimands (D'Haultfoeuille, 2011; Miao et al., 2018).

Assumption 4 (Informative variables)**.** *Let g be any mean squared integrable function. Both the source domain and the target domain, $(f, F) \in \{(p, P), (q, Q)\}$, satisfy $\mathbb{E}_f[g(U) \mid x, c] = 0$ for all $x \in \mathcal{X}, c \in \mathcal{C}$ if and only if $g(U) = 0$ almost surely with respect to $F(U)$.*

At a high level, completeness states that the X must have sufficient variability related to the change of U. This is a common assumption made in proximal causal inference (cf. Condition (ii) in Miao et al. (2018) and

Assumption 3 in Mastouri et al. (2021)). For more details on the justification of completeness assumption, see the supplementary material of Miao et al. (2022).

Second, we need a guarantee on the support of $u \in \mathcal{U}$. Intuitively, if a $u \in \mathcal{U}$ has non-zero probability in the target domain, it should have non-zero probability in the source domain as well. Otherwise, it is impossible to adjust to certain shifts (as we never see these regimes in the source domain). This is similar to the positivity assumption commonly made in causality literature (Hernán and Robins, 2006).

Assumption 5 (Positivity)**.** *For any $u \in \mathcal{U}$, if $Q(u) > 0$ then $P(u) > 0$.*

If data are generated according to Figure 1c, and the regularity conditions 8–10 hold (see Appendix A.2), Miao et al. (2018) first showed the existence of the solutions $h_0^p(w,c), h_0^q(w,c)$ of the following equations:

$$\mathbb{E}_p[Y \mid c, x] = \int_{\mathcal{W}} h_0^p(w,c) \mathrm{d}P(w \mid c, x) \tag{4.1}$$
$$\mathbb{E}_q[Y \mid c, x] = \int_{\mathcal{W}} h_0^q(w,c) \mathrm{d}Q(w \mid c, x).$$

The terms $h_0^p(w,c), h_0^q(w,c)$ are called 'bridge' functions as they connect the proxy W to the label Y. If we are able to identify $h_0^q(w,c)$ then we can identify $\mathbb{E}_q[Y \mid x]$, by using eq. (4.1) to obtain $\mathbb{E}_q[Y \mid C, x]$ and marginalizing over $Q(C \mid x)$.

We show that it is possible to connect identification of $h_0^q(w,c)$ with that of $h_0^p(w,c)$, leading directly to identification of $\mathbb{E}_q[Y \mid x]$.

Theorem 4.1. *Assume that h_0^p and h_0^q exist (i.e., regularity Assumptions 8–10 hold). Then given Assumptions 1, 2, 4, 5 we have that, for any $c \in \mathcal{C}$,*

$$\int_{\mathcal{W}} h_0^p(w,c) \mathrm{d}P(w \mid u) = \int_{\mathcal{W}} h_0^q(w,c) \mathrm{d}Q(w \mid u),$$

almost surely with respect to $Q(U)$. This implies that

$$\mathbb{E}_q[Y \mid x] = \int_{\mathcal{W} \times \mathcal{C}} h_0^p(w,c) \mathrm{d}Q(w, c \mid x).$$

The proof is given in Appendix B.1. Hence, given h_0^p and (W, X, C) from the target Q, we are able to adapt to arbitrary distribution shifts in unobserved U. The advantage of this approach is that it will not require estimating any distributions involving U. We demonstrate this in Section 5.

While concepts can ensure identifiability, they may not be available in practice. In this case, a natural question is whether the optimal target predictor $\mathbb{E}_q[Y \mid x]$ is still identifiable. In the next section we show that if we instead have access to data from multiple source domains, $\mathbb{E}_q[Y \mid x]$ may again be identifiable.

4.2 The Blessings of Multiple Domains

We now turn to the multi-domain setting. The graphical structure in Figure 1d is similar to the structure in Figure 1c with C replaced by X, X replaced by Z, and the arrow between U and Z flipped. Although the bridge function proposed by Miao et al. (2018) assumes an edge from U to Z, changing the direction from Z to U does not change the conditional independence structure (Pearl, 2009). The main difference is we will only be able to guarantee full identification when U is discrete. We start by demonstrating this, and then give an example of the inherent difficulty of identification when U is continuous.

To begin, for simplicity, assume U and W are discrete (with dimensionalities k_U and k_W). We have finitely many samples from Z, denoted as $z_1, \ldots, z_{k_Z}$, corresponding to our training domains. We seek a bridge function (in this case, a matrix $M_0(w_i, x)$) satisfying

$$\mathbb{E}_r[Y \mid x] = \sum_{i=1}^{k_w} M_0(w_i, x) P_r(w_i \mid x), \tag{4.2}$$

for all $r = 1, \ldots, k_Z$, where $\mathbb{E}_r[Y \mid x]$ is the conditional expectation obtained in domain r, and $P_r(W \mid x) = P(W \mid x, z_r)$.

In order to identify $M_0(w_i, x)$ and $\mathbb{E}_q[Y \mid x]$, we need enough source domains to capture the variability of U. The following result describes how many we need.

Proposition 4.2. *Suppose that we have k_Z source domains and W, U have k_W and k_U categories respectively. Then, if $k_W, k_Z \geqslant k_U$ and subject to appropriate rank conditions (see proof in Appendix B.2), the bridge function is identifiable and does not depend on the specific z.*

This result generalizes the identification analysis developed in Miao et al. (2018). If the number of observed source domains k_Z is greater than the dimension of the latent U, then subject to appropriate identifiability requirements (detailed in Appendix B.2), we can recover the bridge $M_0(w_i, x)$.

Now, consider the case where U is discrete but all observed variables W, X, Y are continuous. In this case we have the following system

$$\mathbb{E}_r[Y \mid x] = \int_{\mathcal{W}} m_0(w, x) \mathrm{d}P_r(w \mid x), \tag{4.3}$$

for $r = 1, \ldots, k_Z$. The proof of existence of m_0 is a modification of Proposition A.2, as shown in Proposition A.3. In order to identify target $\mathbb{E}_q[Y \mid x]$, we need the following assumption.

Assumption 6. *Let g be a square integrable function on U. For each $x \in \mathcal{X}$ and for all $z \in \mathcal{Z}_p$, $\mathbb{E}[g(U) \mid x, z] = 0$ if and only if $g(U) = 0$, $P(U)$ almost surely.*

Given this assumption we can prove identifiability.

Proposition 4.3. *Given that Assumptions 1–3, 6 hold; that m_0 exists; that (W, X, Y) are observed for the sources $z \in \mathcal{Z}_p$, and (W, X) is observed from the target domain. Then $\mathbb{E}_q[Y \mid x]$ is identifiable, and for any $x \in \mathcal{X}$, we can write*

$$\mathbb{E}_q[Y \mid x] = \int_{\mathcal{W}} m_0(w, x) dQ(w \mid x). \tag{4.4}$$

The proof is given in Appendix B.3. Crucially, this result is valid only when Assumptions 6 holds, and it remains unclear when it is expected to hold. Proposition 4.2 suggests that Assumptions 6 is not vacuous when U is finite dimensional.

Now let us consider the case where U is continuous. In this case, unfortunately, Assumption 6 is unlikely to hold, preventing identification of $\mathbb{E}_q[Y \mid x]$. This is illustrated in the following example.

Example 4.4. *Recall the decomposition of both sides of (4.3). Under Assumption 2 and given the existence of m_0 (Proposition A.2),*

$$\begin{aligned}\mathbb{E}_p[Y \mid x, z] &= \int_{\mathcal{W}} m_0(w, x) \mathrm{d}P(w \mid x, z) \\ &= \int_{\mathcal{U}} \int_{\mathcal{W}} m_0(w, x) \mathrm{d}P(w \mid u) \mathrm{d}P(u \mid x, z);\end{aligned} \tag{4.5}$$

$$\mathbb{E}_p[Y \mid x, z] = \int_{\mathcal{U}} \mathbb{E}_p[Y \mid x, u] \mathrm{d}P(u \mid x, z). \tag{4.6}$$

For every x, Eqs. (4.5) and (4.6) represent projections onto $P(u \mid x, z_r)$, $r \in 1, \ldots, k_z$. Consider $\mathcal{U} := [-\pi, \pi]$ with periodic boundary conditions, and for a given x define $P(u \mid x, z_r) = (2\pi)^{-1}(1 + \cos(ru)), \forall r \in \mathbb{N}^+$ (note that cosines form an orthonormal basis). We now construct an example where (4.5) holds for some z but not for others. Define the difference

$$\begin{aligned}&\mathbb{E}_p[Y \mid x, u] - \int_{\mathcal{W}} m_0(w, x) \mathrm{d}P(w \mid u) \\ &= \cos((k_z + 1)u) =: g(u).\end{aligned} \tag{4.7}$$

In this case, $g(u) \neq 0$, and in particular, (4.5) holds for all $r \leqslant k_z$, but not for $P(u \mid x, z_{k_z+1})$.

This example illustrates a larger point: that for continuous U, no finite set of projections will suffice to completely characterize the square integrable functions on $\mathcal{U}$. That said, as more projections are employed, and subject to appropriate assumptions on the smoothness of (4.7), the error will reduce as more domains are observed. The characterization of this convergence will be the topic of future work. In experiments, we show that the adaptation can still be effective even when the latent variable $U|z_r$ is continuous valued and follows different Beta distributions for each distinct r, given just two training source domains.

5 Kernel Bridge Function Estimation

We introduce kernel methods to estimate the bridge functions and subsequently leverage the estimates to adapt to distribution shifts. Section 4 shows that bridge functions for both settings can be adapted to the target domain, so we drop the domain specific indices and use h_0 and m_0 to denote the bridge functions. We begin by introducing the notation.

Notation. Let $\otimes$ be the tensor product, $\overline{\otimes}$ be the columnwise Khatri-Rao product and $\odot$ be the Hadamard product. For any space $\mathcal{V} \in \{\mathcal{X}, \mathcal{C}, \mathcal{W}, \mathcal{Y}\}$, let $k : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ be a positive semidefinite kernel function and $\phi(v) = k(v, \cdot)$ for any $v \in \mathcal{V}$ be the feature map. We denote $\mathcal{H}_\mathcal{V}$ to be the RKHS on $\mathcal{V}$ associated with kernel function k. The RKHS has two properties: (i) $f \in \mathcal{H}_\mathcal{V}$, $f(v) = \langle f, k(v, \cdot)\rangle$ for all $v \in \mathcal{V}$ and (ii) $k(v, \cdot) \in \mathcal{H}_\mathcal{V}$. We denote $\langle \cdot, \cdot \rangle$ as the inner product and $\|\cdot\|_{\mathcal{H}_\mathcal{V}}$ as the induced norm. For notation simplicity, we denote the product space $\mathcal{H}_\mathcal{V} \times \mathcal{H}_{\mathcal{V}'}$ associated with operation $\mathcal{H}_\mathcal{V} \otimes \mathcal{H}_{\mathcal{V}'}$ as $\mathcal{H}_{\mathcal{V}\mathcal{V}'}$. We define the kernel mean embedding as $\mu_V = \mathbb{E}[\phi(V)] = \int k(v, \cdot)p(v)dv$ (Smola et al., 2007) and the conditional mean embedding as $\mu_{V|y} = \int k(v, \cdot)p(v \mid y)dv$ (Song et al., 2009; Singh et al., 2019). For $V \in \{W, X, C\}$, we denote the a-th batch of *i.i.d.* samples as $V_a = \{v_{a,i}\}_{i=1}^{n_a}$. Define the Gram matrices as $\mathcal{K}_{V_a} = \left[k(v_{a,i}, v_{a,j})\right]_{i,j} \in \mathbb{R}^{n_a \times n_a}$, $\mathcal{K}_{V_{ab}} = \left[k(v_{a,i}, v_{b,j})\right]_{i,j} \in \mathbb{R}^{n_a \times n_b}$. Let $\Phi_{V_a} = \left[\phi(v_{a,1}), \ldots, \phi(v_{a,n_a})\right]^\top \in \mathcal{H}_\mathcal{V}^{n_a}$ be the vectorized feature map such that $\Phi_{V_a}(v') = \left[k(v_{a,1}, v'), \ldots, k(v_{a,n_a}, v')\right]^\top \in \mathbb{R}^{n_a}$.

5.1 Adaptation with Concepts

Suppose that for the bridge function $h_0 \in \mathcal{H}_{\mathcal{W}\mathcal{C}}$, where $\mathcal{H}_{\mathcal{W}\mathcal{C}}$ is a RKHS. It follows from Theorem 4.1 that

$$\begin{aligned}\mathbb{E}_q[Y \mid X = x] &= \mathbb{E}_q[h_0(W, C) \mid x] \\ &= \mathbb{E}_q[\langle h_0, \phi(W) \otimes \phi(C)\rangle \mid x] \\ &= \langle h_0, \mu^q_{WC|x}\rangle.\end{aligned} \tag{5.1}$$

To adapt to the distribution shifts, we estimate the bridge function h_0 in the source domain and the conditional mean embedding $\mu^q_{WC|x} = \mathbb{E}_q[\phi(W) \otimes \phi(C) \mid x]$ in the target domain. The empirical estimate of the conditional mean embedding along with the consistency proof have been provided in (Song et al., 2009; Grünewälder et al., 2012) thus we focus on the estimation procedure of the bridge function h_0.

To estimate the bridge function h_0, we employ the regression method developed in Mastouri et al. (2021).

Recall $\mathbb{E}[Y \mid c, x] = \mathbb{E}[h_0(W, c) \mid c, x]$. We define the population risk function in the source domain as:

$$R(h_0) = \mathbb{E}_p[(Y - G_{h_0}(C, X))^2]; \quad (5.2)$$
$$G_{h_0}(x, c) = \langle h_0, \mu^p_{W|c,x} \otimes \phi(c) \rangle.$$

The procedure to optimize (5.2) involves two stages. In the first stage, we estimate the conditional mean embedding $\mu^p_{W|c,x} = \mathbb{E}_p[\phi(W) \mid c, x]$, which we will use as a plug-in estimator to estimate h_0 in the second step. Given n_1 *i.i.d.* samples $(X_1, W_1, C_1) = \{(x_{1,i}, w_{1,i}, c_{1,i})\}_{i=1}^{n_1}$ from the source distribution p and a regularizing parameter $\lambda_1 > 0$, we denote $\mathcal{K}_{X_1} \in \mathbb{R}^{n_1 \times n_1}$, $\mathcal{K}_{C_1} \in \mathbb{R}^{n_1 \times n_1}$ as the Gram matrices and $\Phi_{X_1} \in \mathcal{H}^{n_1}_{\mathcal{X}}$, $\Phi_{C_1} \in \mathcal{H}^{n_1}_{\mathcal{C}}$ as n_1-dimensional vectorized feature maps of X_1, C_1 respectively. Following the procedure developed in Song et al. (2009), the estimate of $\mu^p_{W|x,c}$ is

$$\widehat{\mu}^p_{W|c,x} = \sum_{i=1}^{n_1} b_i(x, c)\phi(w_{1,i}); \quad (5.3)$$
$$b(x, c) = (\mathcal{K}_{X_1} \odot \mathcal{K}_{C_1} + \lambda_1 n_1 I)^{-1} (\Phi_{X_1}(x) \odot \Phi_{C_1}(c)).$$

In the second stage, we replace $\mu^p_{W|x,c}$ with $\widehat{\mu}^p_{W|x,c}$ in (5.2) and define the empirical risk. Consider n_2 *i.i.d.* samples $(X_2, Y_2, C_2) = \{(x_{2,i}, y_{2,i}, c_{2,i})\}_{i=1}^{n_2}$ from the source distribution and a regularization parameter $\lambda_2 > 0$, we want to minimize

$$\operatorname*{argmin}_{h_0 \in \mathcal{H}_{\mathcal{WC}}} \frac{1}{2n_2} \sum_{i=1}^{n_2} \left(y_{2,i} - \langle h_0, \phi(c_{2,i}) \otimes \widehat{\mu}^p_{W|c_{2,i}, x_{2,i}} \rangle \right)^2 + \lambda_2 \|h_0\|^2_{\mathcal{H}_{\mathcal{WC}}}. \quad (5.4)$$

We follow the same analysis procedure derived in Mastouri et al. (2021). The solution to (5.4) is shown in the following.

Proposition 5.1. *Let $\mathcal{K}_{W_1} \in \mathbb{R}^{n_1 \times n_1}$, $\mathcal{K}_{C_2} \in \mathbb{R}^{n_2 \times n_2}$ be the Gram matrices of W_1 and C_2, respectively. Let $\mathcal{K}_{X_{12}} \in \mathbb{R}^{n_1 \times n_2}$, $\mathcal{K}_{C_{12}} \in \mathbb{R}^{n_1 \times n_2}$ be the cross Gram matrices of (X_1, X_2) and (C_1, C_2), respectively. For any $\lambda_2 > 0$, there exists a unique optimal solution to (5.4) of the form*

$$\widehat{h}_0 = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \alpha_{ij} \phi(w_{1,i}) \otimes \phi(c_{2,j});$$
$$vec(\alpha) = (I \otimes \Gamma)(\lambda_2 n_2 I + \Sigma)^{-1} y_2,$$

where $\Sigma = (\Gamma^\top \mathcal{K}_{W_1} \Gamma) \odot \mathcal{K}_{C_2}$, $\Gamma = (\mathcal{K}_{X_1} \odot \mathcal{K}_{C_1} + \lambda_1 n_1 I)^{-1} (\mathcal{K}_{X_{12}} \odot \mathcal{K}_{C_{12}})$, and $y_2 = \left[y_{2,1}, \ldots, y_{2,n_2}\right]^\top$.

Proposition 5.1 is an application of the Representer theorem (Schölkopf et al., 2001) – the optimal estimate of the infinite dimensional operator is a finite rank operator spanned by the feature space of W_1 and C_2.

Finally, given estimate $\widehat{\mu}^q_{WC|x}$ and a new sample x_{new}, we can construct the empirical predictor of (5.1) as

$$\widehat{y}_{\text{pred}} = \langle \widehat{h}_0, \widehat{\mu}^q_{WC|x_{\text{new}}} \rangle.$$

This completes the full adaptation procedure.

On classification tasks. For classification tasks, where the label is $Y \in \{1, \ldots, k_Y\}$, we treat the multi-task regressor as a classifier. We encode Y by a one-hot encoder and then regress on the encoded $\widetilde{Y} \in \{0, 1\}^{k_Y}$. Each label ℓ has a corresponding bridge function $h_{0,\ell}$ for $\ell \in \{1, \ldots, k_Y\}$. For $i = 1, \ldots, n_2$, let the encoded $y_{2,i}$ be $\widetilde{y}_{2,i} = \left[\widetilde{y}_{2,i,1}, \ldots, \widetilde{y}_{2,i,k_Y}\right]^\top \in \{0, 1\}^{k_Y}$. Then for each ℓ, we can estimate $h_{0,\ell}$ by replacing $y_{2,i}$ in (5.4) with $\widetilde{y}_{2,i,\ell} \in \{0, 1\}$. For each new sample x_{new}, the predicted score of label ℓ is $\widehat{y}_{\text{pred},\ell} = \langle \widehat{h}_{0,\ell}, \widehat{\mu}^q_{WC|x_{\text{new}}} \rangle$, and we select the label that has the highest prediction score: $\operatorname{argmax}_\ell \widehat{y}_{\text{pred},\ell}$.

5.2 Adaptation with Multiple Domains

In the multiple source domain setting, the estimation of m_0 follows similarly to that of h_0. Assuming that $m_0 \in \mathcal{H}_{\mathcal{WX}}$, then (4.3) can be written as

$$\mathbb{E}_r[Y \mid x] = \mathbb{E}_p[\langle m_0, \mu_{W|x,r} \otimes \phi(x) \rangle \mid x],$$

for $r = 1, \ldots, k_Z$. The task is to estimate m_0 from the source domain and then apply it to the target domain. We can define the population risk function as

$$R(m_0) = \sum_{r=1}^{k_Z} \mathbb{E}_r[(Y - G_{m_0}(r, X))^2]; \quad (5.5)$$
$$G_{m_0}(r, x) = \langle m_0, \mu_{W|r,x} \otimes \phi(x) \rangle.$$

We employ the two-stage estimation procedure as we did for estimating h_0: (i) we first estimate $\mu_{W|r,x}$ and then (ii) plug the estimate $\widehat{\mu}_{W|r,x}$ to estimate m_0.

At the r-th domain, we observe the samples: $\{(w_{r,i}, x_{r,i}, r)\}_{i=1}^{n_r}$. As with (5.3), we learn a conditional mean embedding $\widehat{\mu}_{W|r,x} = \sum_{i=1}^{n_r} d_{r,i}(x)\phi(w_{r,i})$, where $d_r(x) = (\mathcal{K}_{X_r} + \lambda_3 I)^{-1} (\Phi_{X_r}(x)) \in \mathbb{R}^{n_r}$ and $\lambda_3 > 0$ for $r = 1, \ldots, k_Z$. In the second stage, given another batch of independent samples: $\{(y_{r,i}, x_{r,i}, r)\}_{i=1}^{n_r}$ for $r = 1, \ldots, k_Z$, we minimize:

$$\frac{1}{2\sum_{r=1} n_r} \sum_{r=1}^{k_Z} \sum_{i=1}^{n_r} \left(y_{r,i} - \langle m_0, \phi(x_{r,i}) \otimes \widehat{\mu}_{W|r,x_{r,i}} \rangle \right)^2 + \lambda_4 \|m_0\|^2_{\mathcal{H}_{\mathcal{WX}}}. \quad (5.6)$$

Then, $\widehat{m}_0$ yields an analytical solution in similar form to $\widehat{h}_0$ shown in Proposition 5.1 (see Appendix C.2 for details). Finally, with the estimated conditional mean embedding $\widehat{\mu}^q_{W|x}$ and a new sample x_{new} from the target test set, we have

$$\widehat{y}_{\text{pred}} = \langle \widehat{m}_0, \widehat{\mu}^q_{W|x_{\text{new}}} \otimes \phi(x_{\text{new}}) \rangle.$$

We convert the regression task with m_0 to the classification task by learning k_Y bridge functions, where each bridge function $m_{0,\ell}$ corresponds to label ℓ.

6 Experiments

We verify our theory with both simulated and real data, demonstrating robustness to latent shifts and transferablility of the bridge functions.

For the setting with concept variables present, we compare our method with baselines: Empricial Risk Minimization (ERM), Covariate shift weighting (COVAR) (Shimodaira, 2000), Label shift weighting (LABEL) (Buck et al., 1966), and the spectral (LSA-S) and Wasserstein Autoencoder (LSA-WAE) latent shift adaptation approaches (Alabdulmohsin et al., 2023). For the multi-domain setting, we compare our method with baselines: Simple Adaptation (SA) (Mansour et al., 2008), Weighted Combination of Source Classifiers (WCSC) (Zhang et al., 2015), and Marginal Kernel (MK) (Blanchard et al., 2011). We also compare with multi-domain generalization baselines (Muandet et al., 2013): Domain Adversarial Neural Networks (DANN) (Ganin et al., 2016), Maximum Mean Discrepancy (MMD) (Gretton et al., 2012). Additionally, we modify the ERM method to the multi-domain setting by concatenating the source samples to learn one ERM model (Cat-ERM) or taking the average result of each source domain ERM model (Avg-ERM). The ORACLE model is a model that is trained on target distribution samples. and evaluated on held-out target distribution samples. The tuning parameters for all models including the proposed model are selected using five-fold cross-validation. Details regarding the setups are in Appendix D.

Classification task. The task designed in Alabdulmohsin et al. (2023) is a binary classification problem with $Y \in \{0,1\}$ and the latent variable $U \in \{0,1\}$ is a Bernoulli random variable. Additionally, $X \in \mathbb{R}^2, W \in \mathbb{R}$ are continuous random variables and $C \in \mathbb{R}^3$ is a discrete variable. We have one source domain with $P(U=1) = 0.1$. We evaluate the models on the target distribution with $Q(U)$ shifting from $Q(U=1) \in \{0.1, \ldots, 0.9\}$. The goal of this task is to investigate whether the adaptation method is robust to any arbitrary shift of U.

The ORACLE and ERM model are implemented as MultiLayer Perceptrons (MLP). The kernel function used in the proposed method is the Gaussian kernel.

We compare the proposed method with the LSA-S and Wasserstein Autoencoder adaptation LSA-WAE approaches developed in Alabdulmohsin et al. (2023). While all three methods are designed to adjust shift for the same graph in Figure 1c, our method takes additional W, C, X as training samples in the target domain while LSA-S and LSA-WAE only take X. For all three methods, only X is observed in the test data.

While the identification theory developed in (Alabdulmohsin et al., 2023) does not require W, C in the target domain, we are aware that in practice, having more information in the target domain may improve estimation. To make the methods more directly comparable, we design an additional step to incorporate W from the target in the LSA-S algorithm. We describe this procedure in more detail in Appendix D.1.

Results are shown in Figure 2a. The proposed method is more robust to the shift compared to baselines and is close to the ORACLE model. It is shown that with observed W in the target domain, LSA-S does not improve the performance compared to LSA-S without W. We also compare results under different noise levels and observe similar trends as discussed in Appendix D.

dSprites dataset regression task. We test the proposed procedure on the dSprites dataset (Matthey et al., 2017), an image dataset described by five latent parameters (shape, scale, rotation, posX, and posY). Motivated by Matthey et al. (2017)'s experiments, we design a regression task where the dSprites images ($64 \times 64 = 4096$-dimensional) are $X \in \mathbb{R}^{64\times 64}$ and subject to a nonlinear confounder $U \in [0, 2\pi]$ which is a rotation of the image. $W \in \mathbb{R}$ and $C \in \mathbb{R}$ are continuous random variables. For this experiment, we have 7000 training samples and 3000 test samples. Further details about the procedure are in Appendix D.

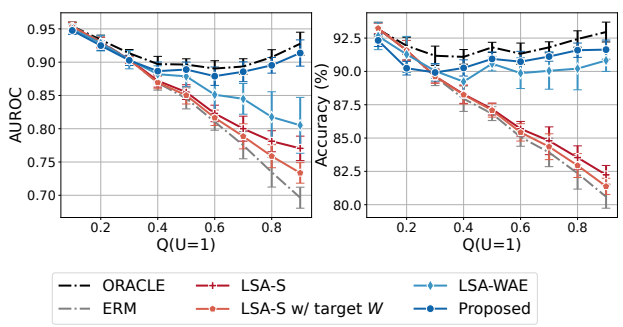
In the results in Figure 2b, we vary a, which controls which region of the source distribution that the target distribution concentrates. We design the experiment such that increasing a shifts the target distribution to increasingly low mass regions of the source distribution. We compute the mean squared error of each method on test examples from the target distribution.

We find that, while the baseline methods degrade as the target distributions shift increases, the proposed method adapts and maintains low error, nearly matching the error achieved by the oracle, which is trained on target distribution samples.

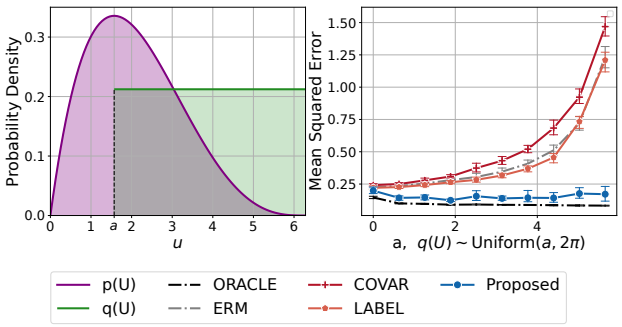
6.1 Multi-Domain Adaptation

In the multi-domain setting, we use the same classification dataset provided in Alabdulmohsin et al. (2023) as Section D.6. We assume that C is not observed in any domain and generate multiple datasets drawn with different distributions on U.

Classification task. We construct three different tasks with different settings of $P(U)$ over the source and target domains. For each task, we construct three



(a) Classification task on simulated data.



(b) Regression on the dSprites dataset.

Figure 2: **Adaptation results with concept and proxy**. Shown is the average evaluation metric on held-out target distribution samples across 10 independent replicates of the data. The proposed method is robust to the latent shift compared to the baselines in both cases. (a) We set $P(U = 1) = 0.1$. Both the AUROC and accuracy remains nearly constant in various degree of shifts, while the performance of other baselines drops as $Q(U = 1)$ moves to 0.9. (b) The left figure denotes the density function of U, the overlapping area of two distribution shrinks as a moves rightward. The result on the right shows that our method is robust even when the overlapping area between two distributions is small.

Table 1: **Multi-domain adaptation result**. The values are the average AUROC of 10 independent replicates of the data. Each task has three source domains with different $P_r(U)$ and one target domain. The proposed method has outperformed other baselines and is close to the Oracle in task 2.

Task	ORACLE	Cat-ERM	Avg-ERM	SA	MK	WCSC	DANN	MMD	Proposed
Task 1	0.9425 ±0.0039	0.8030 ±0.0155	0.7916 ±0.0148	0.7918 ±0.0148	0.5848 ±0.0593	0.5221 ±0.0299	0.8039 ±0.0229	0.8055 ±0.0248	**0.8848** ±0.0120
Task 2	0.9431 ±0.0061	0.8942 ±0.0084	0.8953 ±0.0079	0.8953 ±0.0079	0.8054 ±0.0204	0.8144 ±0.0474	0.9158 ±0.0125	0.9149 ±0.0135	**0.9318** ±0.0063
Task 3	0.8876 ±0.0085	0.8483 ±0.0134	0.8427 ±0.0130	0.8408 ±0.0132	0.8002 ±0.0311	0.7428 ±0.0311	0.8480 ±0.0166	0.8470 ±0.0181	**0.8569** ±0.0095

source domains and one target domain, drawing 3200 random training samples for the each source domain and 9600 random training samples for the target domain. The set of source domains of of Task 1–3 have different combinations of distribution on U documented in Appendix D.3.

The backbone models for ORACLE, Cat-ERM, Avg-ERM, and SA (Mansour et al., 2008) are simple MLPs; MK (Blanchard et al., 2011) is a weighted kernel support vector machine; WCSC (Zhang et al., 2015) is a re-weighted kernel density estimator. SA (Mansour et al., 2008) assumes that $Q(X)$ is the convex combinations of $P_r(X)$ for $r = 1, \ldots, k_Z$; WCSC (Zhang et al., 2015) assumes that $Q(X \mid Y)$ is a linear mixture of $P_r(X|Y)$ for $r = 1, \ldots, k_Z$ domain is an i.i.d. realization from the general distribution.

The results are shown in Table 1. Overall, we find our approach performs better than ERM and baseline multi-domain adaptation methods. All methods perform better in the setting of Task 2 than for Task 1, informally demonstrating the effect of the closeness of the source domains to the target domain. For Task 3, while our proposed approach performs best, ERM also performs well, and substantially better than the domain adaptation baselines.

Regression task. We consider two regression tasks, where U is either a Bernoulli or a Beta random variable. We present the results in Appendix D.

6.2 Concept and multi-domain adaptation with MIMIC-CXR

We conduct a small-scale experiment using a sample of chest X-ray data extracted from the MIMIC-CXR dataset (Johnson et al., 2019). We briefly describe the experimental design and results here, and include a complete description in Appendix D.7. We consider classification of the absence of a radiological finding from low-dimensional embeddings of the X-rays (Sellergren et al., 2022), using the absence of a radiological finding in the radiology report as the target of pre-

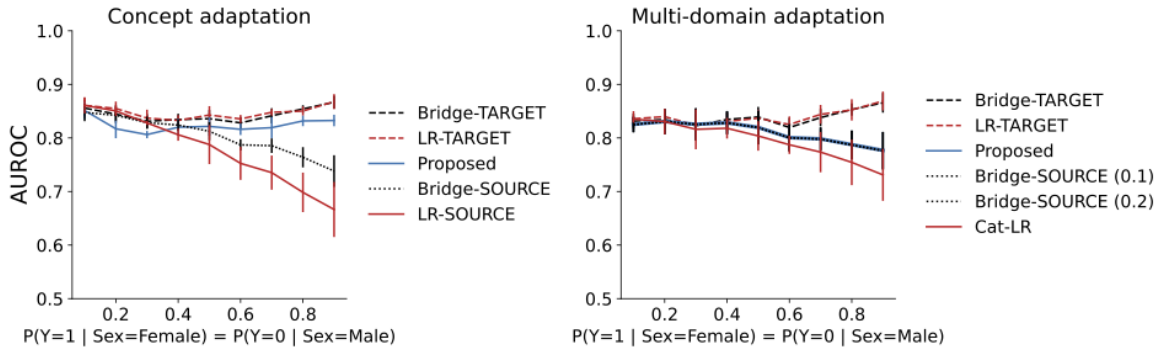


Figure 3: **Concept and multi-domain adaptation with MIMIC-CXR**. Shown are the mean ± SD AUROC of concept (left) and multi-domain adaptation (right) for classification of "No finding" from embeddings of chest X-rays over five replicates of a sampling procedure that introduces a shift in the prevalence of "No finding" with patient sex subgroups, where radiology report embeddings serve as concept variables C and patient age serves as the proxy W. In the concept adaptation experiment, the source domain corresponds to $P(U = 1) = P(Y = 1 \mid \text{Sex} = \text{Female}) = P(Y = 0 \mid \text{Sex} = \text{Male}) = 0.1$. In the multi-domain adaptation experiment, we consider two source domains $P(U = 1) = \{0.1, 0.2\}$.

diction. This corresponds to the "No Finding" label defined by Irvin et al. (2019).

We consider distribution shifts similar to settings in Makar et al. (2022), where patient sex is considered as a possible "shortcut" in the classification of the absence of a radiological finding. We impose distribution shift through structured resampling of the data where $P(U = 1) = P(Y = 1 \mid \text{Sex} = \text{Female}) = P(Y = 0 \mid \text{Sex} = \text{Male})$ and $P(\text{Sex} = \text{Female}) = P(\text{Sex} = \text{Male}) = 0.5$ is held constant. We perform both concept adaptation and multi-domain adaptation experiments with the MIMIC-CXR data. For the concept adaptation experiment, we consider the concept variable C to be the embedding of a radiology report associated with the chest X-ray. We experiment with the use of patient age as a potential proxy W for U due to a hypothesized correlation between the presence of radiological findings and patient age.

The results are summarized in Figure 3. For both experiments, we find that the performance of baseline models fit using only information from the source domain(s) degrades under distribution shift. In the concept adaptation experiment, adaptation is relatively successful, as much of the performance of comparator models fit using target domain data is recovered by the adaptation procedure.

However, we find that the multi-domain adaptation procedure is not successful. In this case, we find that while the multi-domain adaptation procedure marginally outperforms a model fit using the concatenated source domain data under distribution shift, it recovers substantially less of the performance of the target domain model than the concept adaptation procedure does. Furthermore, the adapted model does not outperform the kernel estimators that only leverage information from the source domains. The lack of success in this setting could potentially be explained by insufficient number or diversity of domains relative to the level of noise induced by sampling variability and limited sample size.

7 Discussion

We propose a strategy for adaptation under distribution shift in a latent variable using a bridge function approach (Miao et al., 2018; Tchetgen et al., 2020). This approach allows for identification of the optimal predictor in the target domain without identifying the distribution of the latent variable and without distributional assumptions on the form of the latent. We require that proxies of the latent variable are present and that (i) mediating concepts are available or (ii) data from multiple source domains are present.

We argue our approach is useful for two reasons. First, the latent distribution in general is only identifiable under strict distributional assumptions (Locatello et al., 2019). Second, recovery of the latent variable may be challenging in practice even if it is identifiable (Rissanen and Marttinen, 2021). For example, because most latent variable estimation methods are designed to model the data generating process (Kingma and Welling, 2013), one might allocate substantial modeling capacity to variability in the data and the latent variable that are irrelevant to modeling the shift in the conditional distribution of $Y \mid X$. By contrast, we model only the components of the observable variables relevant to the adaptation.

Acknowledgments: We thank Zhu Li and Dimitri Meunier for helpful discussions. AG was partly supported by the Gatsby Charitable Foundation. OS was partly supported by the UIUC Beckman Institute Graduate Research Fellowship, NSF-NRT 1735252. KT was partly supported by NSF Graduate Research Fellowship Program. SK was partly supported by the NSF III 2046795, IIS 1909577, CCF 1934986, NIH 1R01MH116226-01A, NIFA award 2020-67021-32799, the Alfred P. Sloan Foundation, and Google Inc. This study was funded by Google LLC and/or a subsidiary thereof ('Google').

References

I. Alabdulmohsin, N. Chiou, A. D'Amour, A. Gretton, S. Koyejo, M. J. Kusner, S. R. Pfohl, O. Salaudeen, J. Schrouff, and K. Tsai. Adapting to latent subgroup shifts via concepts and proxies. In *International Conference on Artificial Intelligence and Statistics*, pages 9637–9661. PMLR, 2023.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

G. Blanchard, G. Lee, and C. Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24, 2011.

A. Buck, J. Gart, et al. Comparison of a screening test and a reference test in epidemiologic studies. ii. a probabilistic model for the comparison of diagnostic tests. *American Journal of Epidemiology*, 83(3):593–602, 1966.

Y. Cui, H. Pu, X. Shi, W. Miao, and E. Tchetgen Tchetgen. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, pages 1–12, 2023.

B. Deaner. Proxy controls and panel data. *arXiv preprint arXiv:1810.00283*, 2018.

X. D'Haultfoeuille. On the completeness condition in nonparametric instrumental problems. *Econometric Theory*, 27(3):460–471, 2011.

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Y. Goyal, A. Feder, U. Shalit, and B. Kim. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*, 2019.

A. Gretton, K. Borgwardt, M. Rasch, B. Schoelkopf, and A. Smola. A kernel two-sample test. *JMLR*, 13: 723–773, 2012.

S. Grünewälder, G. Lever, L. Baldassarre, S. Patterson, A. Gretton, and M. Pontil. Conditional mean embeddings as regressors-supplementary. *arXiv preprint arXiv:1205.4656*, 2012.

M. A. Hernán and J. M. Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7):578–586, 2006.

J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

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

M. Kuroki and J. Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2): 423–437, 2014.

Z. Lipton, Y.-X. Wang, and A. Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018.

F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

S. Magliacane, T. Van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. *Advances in neural information processing systems*, 31, 2018.

M. Makar, B. Packer, D. Moldovan, D. Blalock, Y. Halpern, and A. D'Amour. Causally motivated shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics*, pages 739–766. PMLR, 2022.

A. Malinin, N. Band, G. Chesnokov, Y. Gal, M. J. Gales, A. Noskov, A. Ploskonosov, L. Prokhorenkova, I. Provilkov, V. Raina, et al. Shifts: A dataset of real distributional shift across multiple large-scale tasks. *arXiv preprint arXiv:2107.07455*, 2021.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation with multiple sources. *Advances in neural information processing systems*, 21, 2008.

A. Mastouri, Y. Zhu, L. Gultchin, A. Korba, R. Silva, M. Kusner, A. Gretton, and K. Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *International conference on machine learning*, pages 7512–7523. PMLR, 2021.

L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

W. Miao, Z. Geng, and E. J. Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

W. Miao, W. Hu, E. L. Ogburn, and X.-H. Zhou. Identifying effects of multiple treatments in the presence of unmeasured confounding. *Journal of the American Statistical Association*, pages 1–15, 2022.

K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *International conference on machine learning*, pages 10–18. PMLR, 2013.

M. Oberst, N. Thams, J. Peters, and D. Sontag. Regularizing towards causal invariance: Linear models with proxies. In *International Conference on Machine Learning*, pages 8260–8270. PMLR, 2021.

S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE transactions on neural networks*, 22(2): 199–210, 2010.

J. Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference using invariant prediction: identification and confidence intervals, 2015.

S. Rissanen and P. Marttinen. A critical look at the consistency of causal estimation with deep latent variable models. *Advances in Neural Information Processing Systems*, 34:4207–4217, 2021.

D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):215–246, 2021.

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

J. Schrouff, N. Harris, S. Koyejo, I. M. Alabdulmohsin, E. Schnider, K. Opsahl-Ong, A. Brown, S. Roy, D. Mincu, C. Chen, et al. Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. *Advances in Neural Information Processing Systems*, 35:19304–19318, 2022.

A. B. Sellergren, C. Chen, Z. Nabulsi, Y. Li, A. Maschinot, A. Sarna, J. Huang, C. Lau, S. R. Kalidindi, M. Etemadi, et al. Simplified transfer learning for chest radiography models using less data. *Radiology*, 305(2):454–465, 2022.

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

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

R. Singh, M. Sahani, and A. Gretton. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32, 2019.

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A hilbert space embedding for distributions. In *International conference on algorithmic learning theory*, pages 13–31. Springer, 2007.

L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968, 2009.

P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT press, 2000.

A. Subbaswamy, P. Schulam, and S. Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127. PMLR, 2019.

E. J. T. Tchetgen, A. Ying, Y. Cui, X. Shi, and W. Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

V. Veitch, A. D'Amour, S. Yadlowsky, and J. Eisenstein. Counterfactual invariance to spurious correlations in text classification. *Advances in Neural Information Processing Systems*, 34:16196–16208, 2021.

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

L. Xu and A. Gretton. Kernel single proxy control for deterministic confounding. *arXiv preprint arXiv:2308.04585*, 2023.

K. Zhang, M. Gong, and B. Schölkopf. Multi-source domain adaptation: A causal view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Supplementary Materials

A Identification of the Distribution

In this section, we demonstrate the existence of the bridge functions h_0 and m_0 under certain regularity conditions. We first discuss the discrete case and then generalize to the continuous case.

A.1 The Discrete Case of the Bridge Function h_0

The idea of bridge function h_0 may seem abstract in the continuous setting. When every variable is discrete, however, the construction of the bridge function is demonstrated by solving series of matrix problems. This idea originates from Miao et al. (2018) and we apply the technique to show the construction of bridge function when every variable (W, U, C, X, Y) is discrete.

Let $\mathbf{P}(W \mid u) = \begin{bmatrix} P(w_1 \mid u) & \dots & P(w_{k_W} \mid u) \end{bmatrix}^\top \in \mathbb{R}^{k_W}$, $\mathbf{P}(W \mid U) = \begin{bmatrix} \mathbf{P}(W \mid u_1) & \dots & \mathbf{P}(W \mid u_{k_U}) \end{bmatrix} \in \mathbb{R}^{k_W \times k_U}$ be a column vector, and a matrix, respectively. We define similarly $\mathbf{P}(U \mid x, c) = \begin{bmatrix} P(u_1 \mid c, x) & \dots & P(u_{k_U} \mid c, x) \end{bmatrix}^\top \in \mathbb{R}^{k_U}$, $\mathbf{P}(U \mid X, c) = \begin{bmatrix} \mathbf{P}(U \mid x_1, c) & \dots & \mathbf{P}(U \mid x_{k_X}, c) \end{bmatrix} \in \mathbb{R}^{k_U \times k_X}$ for $c \in \mathcal{C}$. We define $\mathbf{P}(Y \mid X, c) = \begin{bmatrix} \mathbf{P}(Y \mid x_1, c) & \dots & \mathbf{P}(Y \mid x_{k_X}, c) \end{bmatrix} \in \mathbb{R}^{k_Y \times k_X}$, $\mathbf{P}(Y \mid U, c) = \begin{bmatrix} \mathbf{P}(Y \mid u_1, c) & \dots & \mathbf{P}(Y \mid u_{k_X}, c) \end{bmatrix} \in \mathbb{R}^{k_Y \times k_X}$, $\mathbf{P}(W \mid X, c) = \begin{bmatrix} \mathbf{P}(W \mid x_1, c) & \dots & \mathbf{P}(W \mid x_{k_X}, c) \end{bmatrix} \in \mathbb{R}^{k_W \times k_X}$ analogously. As an alternative to finding a $h_0(w, c)$ such that

$$\mathbb{E}[Y \mid c, x] = \sum_{i=1}^{k_W} h_0(w_i, c) p(w_i \mid c, x),$$

the proxy problem is converted to finding a $\widetilde{H}_0(Y, W, c)$ such that

$$\mathbf{P}(Y \mid X, c) = \widetilde{H}_0(Y, W, c)\mathbf{P}(W \mid X, c), \quad c \in \mathcal{C}.$$

First, under the condition that $W \perp\!\!\!\perp \{X, C\} \mid U$, we can write

$$\mathbf{P}(W \mid X, c) = \mathbf{P}(W \mid U)\mathbf{P}(U \mid X, c). \tag{A.1}$$

Similarly, under the condition that $Y \perp\!\!\!\perp X \mid \{U, C\}$, we have

$$\mathbf{P}(Y \mid X, c) = \mathbf{P}(Y \mid U, c)\mathbf{P}(U \mid X, c) \tag{A.2}$$

We introduce the following assumption:

Assumption 7. *Columns of $\mathbf{P}(W \mid U)$ are linearly independent. For every $c \in \mathcal{C}$, the columns of $\mathbf{P}(W \mid X, c)$ satisfy $\mathbf{P}(W \mid x, c) \in \mathcal{N}(\mathbf{P}(W \mid U)^*)^\perp$ for all $x \in \mathcal{X}$.*

Assumption 7 is the requirement for the least-squares problem to have an unique solution. Hence, by Assumption 7, we have

$$\mathbf{P}(U \mid X, c) = \mathbf{P}(W \mid U)^\dagger \mathbf{P}(W \mid X, c),$$

where $\mathbf{P}(W \mid U)^\dagger$ is the generalized inverse of $\mathbf{P}(W \mid U)$. Plug the above equation into (A.2), we see that

$$\mathbf{P}(Y \mid X, c) = \underbrace{\mathbf{P}(Y \mid U, c)\mathbf{P}(W \mid U)^\dagger}_{\widetilde{H}(Y, W, c)} \mathbf{P}(W \mid X, c).$$

A.2 Existence of the Bridge Function h_0

The sufficient conditions of existence of h_0 are originally discussed in Miao et al. (2018), we adapt them to our setting and provide a brief review in this section. We assume the following completeness assumption and regularity conditions. This assumption is equivalent to Condition (iii) in Miao et al. (2018).

Assumption 8. *For any mean squared integrable function g and for $c \in \mathcal{C}$, $\mathbb{E}[g(X) \mid W, c] = 0$ almost surely if and only if $g(X) = 0$ almost surely.*

Let f be either the distribution from p or q, we consider $K_c : L_2(W \mid c) \to L_2(X \mid c)$ as the conditional expectation operator associated with the kernel function

$$k(w, x, c) = \frac{f(w, x \mid c)}{f(w \mid c) f(x \mid c)}.$$

Then it follows that $\mathbb{E}[Y \mid c, x] = K_c h_0$:

$$\begin{aligned}\mathbb{E}[Y \mid c, x] &= \int_{\mathcal{W}} h_0(w, c) f(w \mid x, c) \mathrm{d}w \\ &= \int k(w, x, c) h_0(w, c) f(w \mid c) \mathrm{d}w = K_c h_0.\end{aligned}$$

To find the solution h_0, we assume the followings.

Assumption 9. *For any $c \in \mathcal{C}$, $\int_{\mathcal{W}} \int_{\mathcal{X}} f(w \mid c, x) f(x \mid c, w) \mathrm{d}w \mathrm{d}x < \infty$.*

This is a sufficient condition to ensure that K_c is a compact operator (Carrasco et al., 2007, Example 2.3). Hence, by the definition of a compact operator, there exists a singular system $\{\lambda_{c,i}, \phi_{c,i}, \psi_{c,i}\}_{i \in \mathbb{N}}$ of K_c for every $c \in \mathcal{C}$.

Assumption 10. *For fixed $c \in \mathcal{C}$:*

1. $\mathbb{E}[Y \mid X, c] \in L_2(X \mid c)$;
2. $\sum_{i \in \mathbb{N}} \lambda_{c,i}^{-2} \left| \langle \mathbb{E}[Y \mid X, c], \psi_{c,i} \rangle \right|^2 < \infty$.

The above two assumptions are restatements of Conditions (v)–(vii) in Miao et al. (2018). We adapt the results from Proposition 1 in Miao et al. (2018) to the graph in Figure 1c which replaces the node X by C and node Z by X.

Proposition A.1 (Existence of h_0, adapted from Proposition 1 in Miao et al. (2018))**.** *Under Assumption 2, 8–10, the solution to (4.1) exists.*

Proof. The proof follows directly from the result of Picard's theorem. Assumption 9 implies that K_c is a compact operator. Assumption 8 implies that $\mathcal{N}(K_c^*)^\perp = L_2(X \mid c)$. Therefore, under the first statement in Assumption 10, we have $\mathbb{E}[Y \mid X, c] \in \mathcal{N}(K_c^*)^\perp$. Along with the second statement in Assumption 10, we can apply Lemma A.3. □

A.3 Existence of Bridge Function m_0

The proof of the existence of m_0^p is similar to the analysis of h_0. Let $K_x : L_2(W \mid x) \to L_2(Z \mid x)$ be the integral operator associated with the kernel function $k(w, x, z) = p(w, z \mid x)/(p(w \mid x)p(z \mid x))$. Then, we can write

$$\mathbb{E}_p[Y \mid x, z] = \int k(w, x, z) p(w \mid x) m_0(w, x) \mathrm{d}w = K_x m_0.$$

Proposition A.2 (Existence of m_0, Proposition 1 in Miao et al. (2018))**.** *Assume that*

1. *for any mean squared integrable function g and for $x \in \mathcal{X}$, $\mathbb{E}[g(Z) \mid W, x] = 0$ almost surely if and only if $g(Z) = 0$ almost surely;*

2. *For any* $x \in \mathcal{X}$, $\int_{\mathcal{W}} \int_{\mathcal{Z}} f(w \mid x, z) f(z \mid x, w) \mathrm{d}w \mathrm{d}z < \infty$;

3. *For any* $x \in \mathcal{X}$, $\mathbb{E}[Y \mid Z, x] \in L_2(Z \mid x)$;

4. *For any* $x \in \mathcal{X}$, $\sum_{i \in \mathbb{N}} \lambda_{x,i}^{-2} \left| \langle \mathbb{E}[Y \mid Z, x], \psi_{x,i} \rangle \right|^2 < \infty$, *where* $(\lambda_{x,i}, \phi_{x,i}, \psi_{x,i})$ *is the singular system of* K_x.

Then the solution of m_0^p *exists.*

The proof of Proposition A.2 is similar to the proof of Proposition 1 in (Miao et al., 2018), where we replace $P(y|z,x)$ in Proposition 1 of Miao et al. (2018) with $\mathbb{E}[Y \mid Z, x]$. The proof for existence of m_0^q also follows similarly as Proposition A.2.

A.4 Auxiliary Lemma

We introduce the Picard's theorem as follows.

Lemma A.3 (Picard's Theorem)**.** *Let* $K : \mathcal{H}_1 \to \mathcal{H}_2$ *be a compact operator with singular system* $\{\lambda_j, \varphi_j, \psi_j\}_{j=1}^{\infty}$ *and* ϕ *be a given function in* $\mathcal{H}_2$. *Then the equation of first kind* $Kh = \phi$ *have solutions if and only if*

1. $\phi \in \mathcal{N}(K^*)^{\perp}$, *where* $\mathcal{N}(K^*) = \{h : K^* h = 0\}$ *is the null space of the adjoint operator* K^*.

2. $\sum_{j=1}^{+\infty} \lambda_j^{-2} \left| \langle \phi, \psi_j \rangle \right|^2 < \infty$.

B Transferring Bridge Functions

In this section, we discuss the identifiability results.

B.1 Proof of Theorem 4.1

For $f \in \{p, q\}$, recall that

$$\begin{aligned}
\mathbb{E}_f[Y \mid c, x] &= \int_{\mathcal{W}} h_0^f(w, c) f(w \mid c, x) \mathrm{d}w \\
&= \int_{\mathcal{W}} \int_{\mathcal{U}} h_0^f(w, c) f(w \mid c, u) f(u \mid c, x) \mathrm{d}u \mathrm{d}w \\
&= \int_{\mathcal{W}} \int_{\mathcal{U}} h_0^f(w, c) f(w \mid u) f(u \mid c, x) \mathrm{d}u \mathrm{d}w \qquad (W \perp\!\!\!\perp C \mid U).
\end{aligned}$$

Similarly, we can write

$$\mathbb{E}_f[Y \mid c, x] = \int_{\mathcal{U}} \mathbb{E}_f[Y \mid c, u] f(u \mid c, x) \mathrm{d}u \qquad (Y \perp\!\!\!\perp X \mid \{U, C\}).$$

Under Assumption 4, we have

$$\mathbb{E}_f[Y \mid c, U] = \int_{\mathcal{W}} h_0^f(w, c) f(w \mid U) \mathrm{d}w \tag{B.1}$$

almost surely with respect to $F(U)$, $F \in \{P, Q\}$.

Suppose that $u \in \mathcal{U}$ such that $Q(u) > 0$. Then, by Assumption 5 , we must have $P(u) > 0$. Hence, conditioned on the selected u and c and under Assumption 1, we have

$$\begin{aligned}
&\mathbb{E}_p[Y \mid c, u] = \int_{\mathcal{W}} h_0^p(w, c) p(w \mid u) \mathrm{d}w; \\
&\mathbb{E}_q[Y \mid c, u] = \int_{\mathcal{W}} h_0^q(w, c) p(w \mid u) \mathrm{d}w \qquad (p(w \mid u) = q(w \mid u),\ \forall c \in \mathcal{C}, w \in \mathcal{W}, u \in \mathcal{U}).
\end{aligned}$$

We then can write

$$\mathbb{E}_p[Y \mid c, u] - \mathbb{E}_q[Y \mid c, u] = \int_{\mathcal{W}} h_0^p(w, c)p(w \mid u)\mathrm{d}w - \int_{\mathcal{W}} h_0^q(w, c)q(w \mid u)\mathrm{d}w.$$

Note that, by Assumption 1, we have $\mathbb{E}_p[Y \mid c, u] = \mathbb{E}_q[Y \mid c, u]$ and hence the left hand side of the above equation is 0 and we can conclude that:

$$\int_{\mathcal{W}} h_0^p(w, c)p(w \mid U)\mathrm{d}w = \int_{\mathcal{W}} h_0^q(w, c)q(w \mid U)\mathrm{d}w$$

$Q(U)$ almost surely. We complete the first part of proof.

To show the second part of the theorem, note that we can write

$$\begin{aligned}\mathbb{E}_q[Y \mid x] &= \mathbb{E}_q[\mathbb{E}_q[Y \mid C, x] \mid x] \\ &= \mathbb{E}_q[\mathbb{E}_q[h_0^q(W, c) \mid C, x] \mid x].\end{aligned}$$

Since $p(w \mid u) = q(w \mid u)$ by Assumption 1, we can factorize the above equation as

$$\mathbb{E}_q[Y \mid x] = \int_{\mathcal{C}} \left[\int_{\mathcal{U}} \left\{\int_{\mathcal{W}} h_0^q(w, c)p(w \mid u)\mathrm{d}w\right\} q(u \mid c, x)\mathrm{d}u\right] q(c \mid x)\mathrm{d}c.$$

Let the support of U conditioned on c, x be $\mathcal{U}_{c,x}^1 = \{u : Q(u \mid c, x) > 0\}$ and $\mathcal{U}_{c,x}^0 = \{u : Q(u \mid c, x) = 0\}$. Hence, we have $\mathcal{U} = \mathcal{U}_{c,x}^0 \cup \mathcal{U}_{c,x}^1$, and $\mathcal{U}_{c,x}^0 \cap \mathcal{U}_{c,x}^1 = \varnothing$ such that $\int_{\mathcal{U}_{c,x}^0} q(u \mid c, x)\mathrm{d}u = 0$ and $\int_{\mathcal{U}_{c,x}^1} q(u \mid c, x)\mathrm{d}u = 1$. Then, we can further decompose the above as

$$\begin{aligned}\mathbb{E}_q[Y \mid x] &= \int_{\mathcal{C}} \left[\int_{\mathcal{U}_{c,x}^0} \left\{\int_{\mathcal{W}} h_0^q(w, c)p(w \mid u)\mathrm{d}w\right\} q(u \mid c, x)\mathrm{d}u\right] q(c \mid x)\mathrm{d}c \\ &\quad + \int_{\mathcal{C}} \left[\int_{\mathcal{U}_{c,x}^1} \left\{\int_{\mathcal{W}} h_0^q(w, c)p(w \mid u)\mathrm{d}w\right\} q(u \mid c, x)\mathrm{d}u\right] q(c \mid x)\mathrm{d}c \\ &= \int_{\mathcal{C}} \left[\int_{\mathcal{U}_{c,x}^1} \left\{\int_{\mathcal{W}} h_0^q(w, c)p(w \mid u)\mathrm{d}w\right\} q(u \mid c, x)\mathrm{d}u\right] q(c \mid x)\mathrm{d}c.\end{aligned}$$

Given c, x, since the support of $Q(U \mid c, x)$ is included in the support of $Q(U)$, so if $u \in \mathcal{U}_{c,x}^1$, we must have $Q(u) > 0$ and hence $P(u) > 0$ by Assumption 5, and we can swap h_0^q with h_0^p.

$$= \int_{\mathcal{C}} \left[\int_{\mathcal{U}_{c,x}^1} \left\{\int_{\mathcal{W}} h_0^p(w, c)p(w \mid u)\mathrm{d}w\right\} q(u \mid c, x)\mathrm{d}u\right] q(c \mid x)\mathrm{d}c.$$

Since $\int_{\mathcal{U}_{c,x}^0} \left\{\int_{\mathcal{W}} h_0^p(w, c)p(w \mid u)\mathrm{d}w\right\} q(u \mid c, x)\mathrm{d}u = 0$, we can add it to the above term and arrive at

$$\begin{aligned}&= \int_{\mathcal{C}} \left[\int_{\mathcal{U}} \left\{\int_{\mathcal{W}} h_0^p(w, c)p(w \mid u)\mathrm{d}w\right\} q(u \mid c, x)\mathrm{d}u\right] q(c \mid x)\mathrm{d}c \\ &= \int_{\mathcal{C}} \int_{\mathcal{W}} h_0^p(w, c)q(w, c \mid x)\mathrm{d}w\mathrm{d}c. \qquad \text{(B.2)}\end{aligned}$$

Since we can identify h_0^p from the observable (W, X, Y, C) of the source domain by solving the linear system (4.1), given observable (W, C, X) from the target domain, we can identify $\mathbb{E}_q[Y \mid x]$.

B.2 Proof of Proposition 4.2

The following proof is a generalization of the proof of Miao et al. (2018), suited to the multidomain case. All variables besides Z are assumed to be discrete-valued and multivariate: V can take k_v values for $V \in \{U, X, Y, W\}$.

Let $\mathbf{P}(W \mid U) = \begin{bmatrix}\mathbf{P}(W \mid u_1) & \dots & \mathbf{P}(W \mid u_{k_U})\end{bmatrix} \in \mathbb{R}^{k_W \times k_U}$. Similarly, define $\mathbf{P}(Y \mid U, x) = \begin{bmatrix}\mathbf{P}(Y \mid u_1, x) & \dots & \mathbf{P}(Y \mid u_{k_U}, x)\end{bmatrix} \in \mathbb{R}^{k_Y \times k_U}$. This notation carries through to the remaining variables.

The approach we will take differs from the concept case (and standard proxy case) in the following way: we do not observe Z in the training or test domains, nor do we know its true dimension (indeed Z may be continuous valued). Rather, we assume that we have at least k_Z distinct draws z_r from Z in training, where $r \in \{1, \dots, k_Z\}$ is the domain index, and that $k_Z \geqslant k_U$. We also suppose that in test, we observe a distinct draw z_{k_Z+1} which was not seen in training.

Our goal is to obtain a bridge function, which in the categorical case will be a bridge *matrix* of dimension $M_{w,x} \in \mathbb{R}^{k_Y \times k_W}$. Define $P_r(V \mid x) := P(V \mid x, z_r)$ for $V \in \{U, Y, W\}$. We assume that for each x,

$$\text{rank}\left(P_{1:k_Z}(U \mid x)\right) = k_U, \qquad P_{1:k_Z}(U \mid x) := \begin{bmatrix} P_1(U \mid x) & \dots & P_{k_Z}(U \mid x) \end{bmatrix}$$

which implies that $P(U \mid x, z_r)$ varies with z_r, and that we see a sufficient diversity of domains to span the space of vectors on U.

The graphical model supports the conditional independence relation

$$\{Y, X, W\} \perp\!\!\!\perp Z \mid U,$$

however we will only require the standard proxy assumptions

$$W \perp\!\!\!\perp X, Z \mid U,$$
$$Y \perp\!\!\!\perp Z \mid X, U.$$

Next, as in the concept case, we require

$$P(Y|U, x) = M_{w,x} P(W|U),$$

where we assume $\text{rank}(P(W|U)) = k_u$ (as in the first condition of Assumption 7). The matrix $M_{w,x}$ is invariant to the distribution $P(U)$ by construction. If we can solve for $M_{w,x}$, then given a novel domain corresponding to the draw z_{k_z+1}, we have

$$P(Y|U, x) P_{k_z+1}(U|x) = M_{w,x} P(W|U) P_{k_z+1}(U|x)$$
$$P_{k_z+1}(Y|x) = M_{w,x} P_{k_z+1}(W|x).$$

This allows us to compute conditional expectations under $P(Y \mid x)$ in the novel domain, based on observations of (W, X) in this domain.

To solve for $M_{w,x}$, we project both sides on a basis over U arising from the training domains,

$$P(Y|U, x) P_{1:k_Z}(U \mid x) = M_{w,x} P(W|U) P_{1:k_Z}(U \mid x),$$

where we define $P_{1:k_Z}(Y|x) = \begin{bmatrix} P_1(Y \mid x) & \dots & P_{k_Z}(Y \mid x) \end{bmatrix}$, and likewise $P_{1:k_Z}(W \mid x)$. Then the above becomes

$$P_{1:k_Z}(Y|x) = M_{w,x} P_{1:k_Z}(W \mid x)$$
$$M_{w,x} = P_{1:k_Z}(Y|x) P^{\dagger}_{1:k_Z}(W \mid x). \tag{B.3}$$

This demonstrates that we can recover the domain-invariant $M_{w,x}$ purely from observed data.

One domain is not enough: We illustrate with an example, where we again consider the case where all variables are categorical:

$$P(Y|x) = M_{w,x} P(W|x), \tag{B.4}$$

where $P(Y \mid x)$ is a $k_Y \times 1$ vector of probabilities, $P(W \mid x)$ is a $k_W \times 1$ vector of probabilities, and M is a $k_Y \times k_W$ matrix for which we wish to solve. We have too few equations for the number of unknowns.

One solution to (B.4) is the matrix of conditional probabilities $M_{w,x} = P(Y|W, x)$. This matrix is *not* invariant to changes to $P(U)$, however:

$$p(Y|W, x) = p(Y|U, x) P(U|W, x).$$

The posterior $P(U|W, x)$ changes when the prior $P(U)$ changes. In contrast, the solution in (B.3) is guaranteed to be domain invariant.

B.3 Proof of Proposition 4.3

For all $r = 1, \ldots, k_Z$, we can write

$$\begin{aligned}\mathbb{E}_r[Y \mid x] = \mathbb{E}[Y \mid x, z_r] &= \int_{\mathcal{W}} m_0(w, x) \mathrm{d}P(w \mid x, z_r) \\ &= \int_{\mathcal{U}} \int_{\mathcal{W}} m_0(w, x) \mathrm{d}P(w \mid u) \mathrm{d}P(u \mid x, z_r); \qquad \text{(B.5)} \\ \mathbb{E}[Y \mid x, z_r] &= \int_{\mathcal{U}} \mathbb{E}[Y \mid x, u] \mathrm{d}P(u \mid x, z_r). \qquad \text{(B.6)}\end{aligned}$$

By Assumption 6, the integrands of (B.5)–(B.6) have the following property

$$\mathbb{E}[Y \mid x, u] = \int_{\mathcal{W}} m_0(w, x) \mathrm{d}P(w \mid u), \qquad \text{(B.7)}$$

almost surely with respect to $P(U)$. We will show that m_0 can be transferred to identify the distribution in the target domain.

We define the support set $\mathcal{S}_q(x) = \{u : Q(u \mid x) > 0\}$. Therefore, we can write

$$\begin{aligned}\mathbb{E}_q[Y \mid x] &= \int_{\mathcal{U}} \mathbb{E}[Y \mid u, x] \mathrm{d}Q(u \mid x) \\ &= \int_{\mathcal{S}_q(x)} \mathbb{E}[Y \mid u, x] \mathrm{d}Q(u \mid x).\end{aligned}$$

Furthermore, since we have $\mathcal{S}_q(x) \subseteq \{u : P(u) > 0\}$, we can apply (B.7) to obtain

$$\begin{aligned}\mathbb{E}_q[Y \mid x] &= \int_{\mathcal{W}} \int_{\mathcal{U}} m_0(w, x) \mathrm{d}P(w \mid u) \mathrm{d}Q(u \mid x) \\ &= \mathbb{E}_q[m_0(W, x) \mid x].\end{aligned}$$

We complete the proof.

C Estimation Procedure

The estimation procedure of $\widehat{h}_0$ is discussed in Section C.1 and the estimation procedure of $\widehat{m}_0$ is discussed in Section C.2. In Section C.3, we discuss the case when either Z or C is a discrete variable.

C.1 Proof of Proposition 5.1

The proof of Proposition 5.1 simply follows the result in (Mastouri et al., 2021) which extends from the representer theorem (Schölkopf et al., 2001). There exists a $\gamma \in \mathbb{R}^{n_2}$ such that

$$\widehat{h}_0 = \sum_{j=1}^{n_2} \gamma_j \widehat{\mu}_{W|c_{2,j}, x_{2,j}} \otimes \phi(c_{2,j}). \qquad \text{(C.1)}$$

From Song et al. (2009), we have $\widehat{\mu}_{W|c_{2,j}, x_{2,j}} = \sum_{i=1}^{n_1} b_i(c_{2,j}, x_{2,j}) \phi(w_{1,i})$ and b_i is the i-th element of b, a function on $\mathcal{C} \times \mathcal{X}$: $b(c, x) = (\mathcal{K}_{X_1} \odot \mathcal{K}_{C_1} + \lambda_1 n_1 I)^{-1} (\Phi_{X_1}(x) \odot \Phi_{C_1}(c))$. If we expand (C.1) with the previous expression, we have

$$\widehat{h}_0 = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \alpha_{ij} \phi(w_{1,i}) \otimes \phi(c_{2,j}),$$

where $\alpha_{ij} = b_i(c_{2,j}, x_{2,j}) \gamma_j$. Hence, the rest of the proof will focus on finding the expression of α_{ij}. Following the proof technique developed in (Mastouri et al., 2021), we introduce two following lemmas that assist the analysis.

Lemma C.1. *The square of the operator norm of $\widehat{h}_0$, denoted as $\|\widehat{h}_0\|^2_{\mathcal{H}_{\mathcal{WC}}}$, can be represented as*

$$\|\widehat{h}_0\|^2_{\mathcal{H}_{\mathcal{WC}}} = \mathrm{vec}(\alpha)^\top (\mathcal{K}_{C_2} \otimes \mathcal{K}_{W_1}) \mathrm{vec}(\alpha).$$

Proof of Lemma C.1. Write

$$\begin{aligned}\langle \widehat{h}_0, \widehat{h}_0 \rangle &= \left\langle \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \alpha_{ij}\phi(w_{1,i}) \otimes \phi(c_{2,j}), \sum_{m=1}^{n_1}\sum_{r=1}^{n_2} \alpha_{mr}\phi(w_{1,m}) \otimes \phi(c_{2,r}) \right\rangle \\ &= \sum_{i,m=1}^{n_1}\sum_{j,r=1}^{n_2} \alpha_{ij}\alpha_{mr} k(w_{1,i}, w_{1,m}) k(c_{2,j}, c_{2,r}) \\ &= \operatorname{tr}\left(\alpha^\top \mathcal{K}_{W_1} \alpha \mathcal{K}_{C_2}\right) \\ &= \operatorname{vec}(\alpha)^\top \operatorname{vec}(\mathcal{K}_{W_1}\alpha\mathcal{K}_{C_2}).\end{aligned}$$

Using the fact that $\operatorname{vec}(ABC) = (C^\top \otimes A)\operatorname{vec}(B)$, the above display can be written as

$$= \operatorname{vec}(\alpha)^\top (\mathcal{K}_{C_2} \otimes \mathcal{K}_{W_1}) \operatorname{vec}(\alpha).$$

□

Lemma C.2. *For any $c \in \mathcal{C}$, $x \in \mathcal{X}$,*

$$\langle \widehat{h}_0, \phi(c) \otimes \widehat{\mu}_{W|c,x} \rangle = \Phi_{C_2}(c)^\top \alpha^\top \mathcal{K}_{W_1} (\mathcal{K}_{X_1} \odot \mathcal{K}_{C_1} + \lambda_1 n_1 I)^{-1} (\Phi_{C_1}(c) \odot \Phi_{X_1}(x)).$$

Proof of Lemma C.2. Write

$$\begin{aligned}\langle \widehat{h}_0, \phi(c) \otimes \widehat{\mu}_{W|c,x} \rangle &= \left\langle \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \alpha_{ij}\phi(w_{1,i}) \otimes \phi(c_{2,j}), \phi(c) \otimes \sum_{r=1}^{n_1} b_r(c,x)\phi(w_{1,r}) \right\rangle \\ &= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{r=1}^{n_1} \alpha_{ij} k(w_{1,i}, w_{1,r}) k(c_{2,j}, c) b_r(c,x).\end{aligned}$$

Summing over i, j, the above equation is equivalent as

$$\begin{aligned}&= \sum_{r=1}^{n_1} \Phi_{C_2}(c)^\top \alpha^\top \Phi_{W_1}(w_{1,r}) b_r(c,x) \\ &= \Phi_{C_2}(c)^\top \alpha^\top \mathcal{K}_{W_1} b(c,x) \\ &= \Phi_{C_2}(c)^\top \alpha^\top \mathcal{K}_{W_1} (\mathcal{K}_{X_1} \odot \mathcal{K}_{C_1} + \lambda_1 n_1 I)^{-1} (\Phi_{X_1}(x) \odot \Phi_{C_1}(c)) \\ &= (\Phi_{X_1}(x) \odot \Phi_{C_1}(c))^\top (\mathcal{K}_{X_1} \odot \mathcal{K}_{C_1} + \lambda_1 n_1 I)^{-1} \mathcal{K}_{W_1} \alpha \Phi_{C_2}(c).\end{aligned}$$

□

With Lemma C.1–C.2, we can write (5.4) as

$$\frac{1}{2n_2}\|y_2 - D^\top \operatorname{vec}(\alpha)\|_2^2 + \lambda_2 \operatorname{vec}(\alpha)^\top E \operatorname{vec}(\alpha), \tag{C.2}$$

where

$$D = \mathcal{K}_{C_2} \overline{\otimes} \left\{ \mathcal{K}_{W_1} (\mathcal{K}_{X_1} \odot \mathcal{K}_{C_1} + \lambda_1 n_1 I)^{-1} (\mathcal{K}_{X_{12}} \odot \mathcal{K}_{C_{12}}) \right\}, \quad E = \mathcal{K}_{C_2} \otimes \mathcal{K}_{W_1}.$$

Then by setting the gradient of (C.2) with respect to $\operatorname{vec}(\alpha)$ to zero, we will obtain

$$\operatorname{vec}(\alpha) = \left(DD^\top + \lambda_2 n_2 E\right)^{-1} D y_2.$$

Apply Woodbury matrix identity, the above display is equivalent as

$$= E^{-1} D (\lambda_2 n_2 I + D^\top E^{-1} D)^{-1} y_2. \tag{C.3}$$

Using the fact that for matrices A, B, C, F, $(A \otimes B)(C \overline{\otimes} F) = AC \overline{\otimes} BF$, we can simplify $E^{-1}D$ as

$$\begin{aligned} E^{-1}D &= \left(\mathcal{K}_{C_2}^{-1} \otimes \mathcal{K}_{W_1}^{-1}\right)\left[\mathcal{K}_{C_2} \overline{\otimes} \left\{\mathcal{K}_{W_1}(\mathcal{K}_{X_1} \odot \mathcal{K}_{C_1} + \lambda_1 n_1 I)^{-1}(\mathcal{K}_{X_{12}} \odot \mathcal{K}_{C_{12}})\right\}\right] \\ &= I \overline{\otimes} (\mathcal{K}_{X_1} \odot \mathcal{K}_{C_1} + \lambda_1 n_1 I)^{-1}(\mathcal{K}_{X_{12}} \odot \mathcal{K}_{C_{12}}) \\ &= I \overline{\otimes} \Gamma. \end{aligned}$$

Hence, using the fact that $(A \overline{\otimes} B)^\top (C \overline{\otimes} F) = (A^\top C) \odot B^\top F$, we have

$$D^\top E^{-1} D = (\mathcal{K}_{C_2} \overline{\otimes} \mathcal{K}_{W_1}\Gamma)^\top (I \overline{\otimes} \Gamma) = \mathcal{K}_{C_2} \odot (\Gamma^\top \mathcal{K}_{W_1} \Gamma)$$

Hence, we can write (C.3) as

$$\mathrm{vec}(\alpha) = (I \overline{\otimes} \Gamma)\left\{\lambda_2 n_2 I + \mathcal{K}_{C_2} \odot (\Gamma^\top \mathcal{K}_{W_1}\Gamma)\right\}^{-1} y_2.$$

C.2 Proof of Kernel Bridge Function m_0

We begin with the results.

Proposition C.3. *Let $\mathcal{K}_{W_3} \in \mathbb{R}^{n_3 \times n_3}$, $\mathcal{K}_{X_4} \in \mathbb{R}^{n_4 \times n_4}$ be the Gram matrices of W_3 and X_4, respectively. Let $\mathcal{K}_{X_{34}} \in \mathbb{R}^{n_3 \times n_4}$, $\mathcal{K}_{Z_{34}} \in \mathbb{R}^{n_3 \times n_4}$ be the cross Gram matrices of (X_3, X_4) and (Z_3, Z_4), respectively. For any $\lambda_4 > 0$, there exists a unique optimal solution to (5.6) of the form*

$$\begin{aligned} \widehat{m}_0 &= \sum_{i=1}^{n_3}\sum_{j=1}^{n_4} \alpha_{ij}\phi(w_{3,i}) \otimes \phi(x_{4,j}); \\ vec(\alpha) &= (I \overline{\otimes} \Gamma)(\lambda_4 n_4 I + \Sigma)^{-1} y_4, \end{aligned}$$

where $\Sigma = (\Gamma^\top \mathcal{K}_{W_3}\Gamma) \odot \mathcal{K}_{X_4}$, $\Gamma = (\mathcal{K}_{X_3} \odot \mathcal{K}_{Z_3} + \lambda_3 n_3 I)^{-1}(\mathcal{K}_{X_{34}} \odot \mathcal{K}_{Z_{34}})$, and $y_4 = \left[y_{4,1}, \ldots, y_{4,n_4}\right]^\top$.

The proof of Proposition C.3 follows exactly as the proof of Proposition 5.1, with X replaced by Z and C replaced by X.

C.3 Estimation with discrete Z or C

In the case when C or Z happen to be discrete variables, a more efficient alternative to the estimator introduced in Section 5.1 which requires kernelized features of C (or Z), is to solve a separate regression of W on X for each $c \in \mathcal{C}$ (or $z \in \mathcal{Z}$). Define the index set $\Xi_1(c) = \{i : c_{1,i} = c, i = 1, \ldots, n_1\}$, we modify (5.3) as

$$\begin{aligned} \widehat{\mu}^p_{W|c,x} &= \sum_{i=1}^{n_1} b_i(x)\phi(w_{1,i})\mathbb{1}(c_{1,i} = c); \\ b(x) &= (\mathcal{K}_{X_{1,c}} + \lambda_1 I)^{-1}\Phi_{X_{1,c}}(x), \end{aligned}$$

where $\mathcal{K}_{X_{1,c}} = [k(x_{1,i}, x_{1,j})]_{i,j}$ and $\Phi_{X_{1,c}} = \left[\phi(x_{1,i})\right]_i^\top$ with $i, j \in \Xi_1(c)$. Alternatively, one can apply the form in (5.3) but use binary kernel on C (or Z).

D Experiments

In this section we discuss the experimental settings and implementation details. We start with introducing the implementation details of all the baselines and proposed method. Then, we discuss the experimental settings.

D.1 Baselines of Adaptation with Concepts and Proxies

We introduce the baseline methods for the adaptation task with C and W. This includes the baselines methods COVARS, LABELS, ORACLE, LSA-W, LSA-S, LSA-S w/ target W and the proposed method. To select the parameters for the regression task on dSprite, we apply five-fold cross-validation with mean squared error as the metric to select the kernel length scale and the ridge regularization penalty.

COVARS. We fit a domain classifier using logistic regression, compute instance weights following Shimodaira (2000), and learn a weighted kernel ridge regressor with a Gaussian kernel function on the source training samples.

LABELS. The label shift baseline assumes oracle access to labels in the target domain. For the classification task, we compute instance weights $q(Y)/p(Y)$ using the observed frequencies in the validation set for the source domain and the training set for the target domain. For the regression task, we compute the weights by fitting a Gaussian kernel density estimator using the source validation set and the target training set separately. We then use the fitted densities to estimate $q(Y)/p(Y)$ for each sample in the source training set. Finally, we learn a sample-weighted kernel ridge regressor with a Gaussian kernel on the source training samples.

ORACLE. For regression tasks, we learn a kernel ridge regressor with a Gaussian kernel on target training samples. For the classification task, we use a standard MLP trained with sample in the target domain. Details of the model structure are documented in Section D.2.

LSA-W. The estimation procedure follows Section 6 in Alabdulmohsin et al. (2023). In this case, we discretize the values of W by applying additional transform $\text{sign}(w)$ for each sample w.

LSA-S. The estimation procedure follows Algorithm 2–5 in Alabdulmohsin et al. (2023).

LSA-S w/ target W. We briefly describe the procedure to incorporate target W to LSA-S. Alabdulmohsin et al. (2023) showed that $Q(Y|x)$ can be decomposed as

$$Q(Y \mid x) = \sum_{\widetilde{u}} \underbrace{P(Y \mid \widetilde{u}, x)}_{(a)} \underbrace{Q(\widetilde{u} \mid x)}_{(b)} \tag{D.1}$$

$$\propto \sum_{\widetilde{u}} \underbrace{P(Y \mid \widetilde{u}, x)}_{(a)} \underbrace{P(\widetilde{u} \mid x)}_{(c)} \underbrace{\frac{Q(\widetilde{u})}{P(\widetilde{u})}}_{(d)} \frac{P(x)}{Q(x)}, \tag{D.2}$$

where $\widetilde{u}$ is a permutation of original u. Both LSA-WAE and LSA-S are multi-stage procedures to compute (a), (c), (d) individually and combine the results using formula (D.2) to obtain the predicted target distribution. Step (a) corresponds to Algorithm 5, (c) corresponds to Equation (17), and (d) corresponds to Algorithm 4 in (Alabdulmohsin et al., 2023).

With the additional W from target, we can obtain (b) by slightly modifying the one estimation step in LSA-S. We test on this procedure, namely LSA-S w/ target W, with (c), (d) replaced by (b). Suppose that U takes values in $1, \ldots, k_U$ and $\widetilde{U}$ be a permutation of U. Define the matrix $\mathbf{G}$ as:

$$\mathbf{G} = \begin{bmatrix} \langle \widehat{P}(W \mid \widetilde{U} = 1), \widehat{P}(W \mid \widetilde{U} = 1) \rangle & \cdots & \langle \widehat{P}(W \mid \widetilde{U} = 1), \widehat{P}(W \mid \widetilde{U} = k_U) \rangle \\ \vdots & \ddots & \vdots \\ \langle \widehat{P}(W \mid \widetilde{U} = k_U), \widehat{P}(W \mid \widetilde{U} = 1) \rangle & \cdots & \langle \widehat{P}(W \mid \widetilde{U} = k_U), \widehat{P}(W \mid \widetilde{U} = k_U) \rangle \end{bmatrix},$$

where $\widehat{P}(W \mid \widetilde{U} = i)$ is the estimated conditional kernel density function obtained by Algorithm 3 in Alabdulmohsin et al. (2023). The step (b) is computed by solving the following least-squares:

$$\widehat{Q}(\widetilde{\mathbf{U}} \mid x) = \arg\min \left\| \begin{bmatrix} \langle \widehat{Q}(W \mid x), \widehat{P}(W \mid \widetilde{U} = 1) \rangle \\ \vdots \\ \langle \widehat{Q}(W \mid x), \widehat{P}(W \mid \widetilde{U} = k_U) \rangle \end{bmatrix} - \mathbf{G} \begin{bmatrix} Q(\widetilde{U} = 1 \mid x) \\ \vdots \\ Q(\widetilde{U} = k_U \mid x) \end{bmatrix} \right\|_F^2,$$

$$\text{subject to} \quad 0 \leqslant Q(\widetilde{U} = i \mid x) \leqslant 1, \quad i = 1, \ldots, k_U;$$

$$\sum_{i=1}^{k_U} Q(\widetilde{U} = i \mid x) = 1.$$

Then, we compute the predicted conditional probability based on (D.1).

Proposed Method. For the regression task using the dSprite dataset, we employ the Gaussian kernel function as the feature map for both X and W. In the classification task, we also utilize the Gaussian kernel function for X and W. Additionally, we make use of a columnwise binary kernel for C, which performs a binary kernel

operation on each entry and computes the product of all function outputs. To compute $\widehat{h}_0$, we apply one-hot encoder on Y and apply the results in Proposition 5.1 For choosing the kernel length scale for the classification task, we use the validation set with AUROC metric.

D.2 Baselines of Multi-Source Adaptation

For the first three baselines: Cat-ERM, Avg-ERM, and SA, we use a *standard MLP* model as the backbone structure. It is a single hidden layer MLP with size 100 and ReLU activation functions. The network is trained using Adam optimizer (Kingma and Ba, 2014) with learning rate 10^{-3}. The batch size is set to be 200 and the maximum number of iteration is set to be 300.

Cat-ERM. We concatenate all the samples across environments into one dataset. Then, we train the model with a standard MLP model as specified above.

Avg-ERM. For each environment, we train a standard MLP model. During testing, we take the average of predictions from all models.

Simple Adaptation (SA) (Mansour et al., 2008). To implement the method, we build kernel density estimators with Gaussian kernel function to estimate the density $p_r(x)$ for $r = 1, \ldots, k_Z$. We then reweigh the output of the classifier, a standard MLP, of each domain with the normalized weight $P_r(x_{\text{new}}) / \{\sum_r P_{r'}(x_{\text{new}})\}$. The kernel length scale is chosen using five-fold cross-validation with AUROC metric.

Marginal Kernel (MK) (Blanchard et al., 2011). This method involves a kernel SVM with a product kernel on $(\mathcal{X}, P(X))$. For any $x, x' \in \mathcal{X}$ and a distribution on X, P, P', the kernel function is defined as $k((x, P), (x', P')) = k_1(x, x')k_2(P, P')$. Let n be the number of samples. Here k_1 is a Gaussian kernel function, and k_2 is the mean of the Gram matrix $[k(x_i, x'_j)]_{ij} \in \mathbb{R}^{n \times n}$, where x_i for $i = 1, \ldots, n$ is a *i.i.d.* sample from P and x'_j for $j = 1, \ldots, n$ is a *i.i.d.* sample from P'. To accommodate the large dataset, we precompute the Gram matrix and apply it to a linear classifier trained using Stochastic Gradient Descent (SGD) implemented in the package *scikit-learn* (Pedregosa et al., 2011). The kernel length scale is chosen using five-fold cross-validation with AUROC metric.

Weighted Combination of Source Classifiers (WCSC) (Zhang et al., 2015). For each source environment, we estimate the conditional probability $X \mid y$ using kernel density estimator with the Gaussian kernel function. The rest of the estimation procedure follows Section 2 in Zhang et al. (2015). The kernel length scale is chosen using five-fold cross-validation with AUROC metric.

Proposed Method. We use columnwise Gaussian kernel function as the feature map of X, a Gaussian kernel function as the feature map of W. The conditional mean embedding $\widehat{\mu}^p_{W|x,z}$ is estimated using the approach introduced in Section C.3. The analytical solution of $\widehat{m}_0$ is discussed in Proposition C.3. All the kernel length scale and the regularization parameters λ_3, λ_4 are selected using five-fold cross-validation with AUROC metric.

ORACLE. The model is $\langle \widehat{m}_0, \widehat{\mu}^q_{W|x} \rangle$, where both the bridge function $\widehat{m}_0$ and $\widehat{\mu}^q_{W|x}$ are estimated using the target dataset, with the number of training samples equal to the training samples of the source domain. All the kernel length scale and the regularization parameters λ_3, λ_4 are selected using five-fold cross-validation with AUROC metric.

D.3 Classification Task

The classification task discussed in Section D.6 is first introduced Alabdulmohsin et al. (2023). Let $\mathbf{o}(\cdot)$ be the one-hot encoder, we follow their data generation procedure and generate samples using the following data generation process:

$$
\begin{aligned}
U &\sim \text{Categorical}(\boldsymbol{\pi}); \\
W \mid U = u &\sim \mathcal{N}(\mathbf{o}(u)\mathbf{M}_{W|U}, 1); \\
X \mid U = u &\sim \mathcal{N}(\mathbf{o}(u)\mathbf{M}_{X|U}, \mathbf{I}_{k_X}); \\
C_i \mid X = x, U = u &\sim \text{Bernoulli}\Big(\text{logit}^{-1}\big([x\mathbf{M}_{C|X,U=u} + \mathbf{o}(u)\mathbf{M}_{C|U}]_i\big)\Big); \\
Y \mid C = c, U = u &\sim \text{Bernoulli}\Big(\text{logit}^{-1}\big(c\mathbf{M}_{Y|C,U=u} + \mathbf{o}(u)\mathbf{M}_{Y|U}\big)\Big),
\end{aligned}
$$

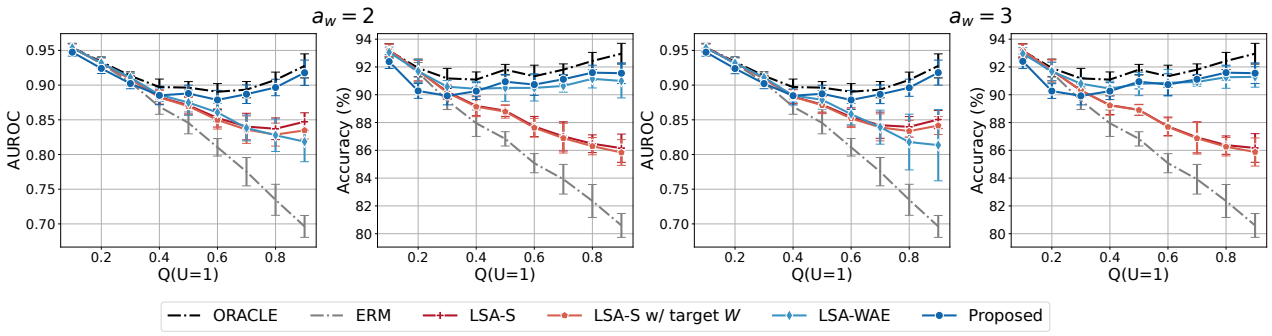


Figure 4: Classification results with $a_w = 2, 3$. The figures indicate that LSA-S and LSA-S w/ target W have comparable performance, aggregating the target W does not seem to improve the performance.

where the matrices are defined as

$$\mathbf{M}_{W|U} := \begin{bmatrix} -1 & 1 \end{bmatrix}^\top, \quad \mathbf{M}_{X|U} := a_w \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}, \quad \mathbf{M}_{C|U} := \begin{bmatrix} -2 & 2 & 2 \\ -1 & 1 & 2 \end{bmatrix};$$
$$\mathbf{M}_{C|X,U=u_0} := 3 \begin{bmatrix} -2 & 2 & -1 \\ 1 & -2 & -3 \end{bmatrix}, \quad \mathbf{M}_{C|X,U=u_1} := 3 \begin{bmatrix} 2 & -2 & 1 \\ -1 & 2 & 3 \end{bmatrix};$$
$$\mathbf{M}_{Y|U} := \begin{bmatrix} 2 & 2 \end{bmatrix}^\top, \quad \mathbf{M}_{Y|C,U=u_0} := \begin{bmatrix} 3 & -2 & -1 \end{bmatrix}^\top, \quad \mathbf{M}_{Y|C,U=u_1} := \begin{bmatrix} 3 & -1 & -2 \end{bmatrix}^\top.$$

The coefficient $a_w = 1$ in Figure 2a. Figure 4 displays additional results where $a_w = 2, 3$. We generate 7000 training samples, 1000 validation samples, and 2000 testing samples for the classification task with concepts and proxies.

In the multi-domain case, we construct 3 different tasks: Task 1 is composed of z_1, z_2, z_3 such that $P(U = 0 \mid z_1) = 0.1$, $P(U = 0 \mid z_2) = 0.2$, $P(U = 0 \mid z_3) = 0.3$ and a target domain with $Q(U = 0) = 0.9$. For task 2, we select z_4, z_5, z_6 such that $P(U = 0 \mid z_4) = 0.4$, $P(U = 0 \mid z_5) = 0.5$, $P(U = 0 \mid z_6) = 0.6$ and $Q(U = 0) = 0.9$. For task 3, we select z_7, z_8, z_9 such that $P(U = 0 \mid z_7) = 0.7$, $P(U = 0 \mid z_8) = 0.8$, $P(U = 0 \mid z_9) = 0.9$ and $Q(U = 0) = 0.4$. The results are shown in Table 1– 2.

D.4 Comparison to Domain Generalization Baselines

Table 2: **Multi-domain generalization vs. (proposed) adaptation result**. The values are the average AUROC of 10 independent runs drawn from the data generating process. Each task has three source domains with different $P_r(U)$ and one target domain. The proposed method has outperformed all domain generalization benchmarks across all tasks.

	ORACLE	ARM	CDANN	CORAL	DANN	GroupDRO	IRM	MMD	VREx	Proposed
Task 1	0.9425	0.8065	0.8061	0.8030	0.8039	0.7954	0.7989	0.8055	0.8010	**0.8848**
	±0.0039	±0.0247	±0.0252	±0.0236	±0.0229	±0.0323	±0.0283	±0.0248	±0.0279	±0.0120
Task 2	0.9431	0.9143	0.9159	0.9158	0.9158	0.9160	0.9131	0.9149	0.9136	**0.9318**
	±0.0061	±0.0150	±0.0125	±0.0132	±0.0125	±0.0125	±0.0135	±0.0135	±0.0124	±0.0063
Task 3	0.8876	0.8470	0.8456	0.8473	0.8480	0.8487	0.8469	0.8470	0.8470	**0.8569**
	±0.0085	±0.0171	±0.0164	±0.0163	±0.0166	±0.0185	±0.0186	±0.0181	±0.0132	±0.0095

Given that we observe multiple domains at test time, a natural question is: Does adaptation give us an advantage over generalization? In generalization, we cannot assume to have any observations in the target domain. We compare our adaptation method with multi-domain generalization baselines (Muandet et al., 2013): Adaptive Risk Minimization (ARM) (Zhang et al., 2021), Conditional Domain Adversarial Neural Networks (CDANN) (Long et al., 2018), Correlation Alignment (CORAL) (Sun and Saenko, 2016), Domain Adversarial Neural Networks (DANN) (Ganin et al., 2016), Distributionally Robust Optimization for Group Shifts (GroupDRO) (Sagawa et al., 2019), Invariant Risk Minimization (IRM) (Arjovsky et al., 2019), Maximum Mean Discrepancy (MMD) (Borgwardt et al., 2006), and Risk Extrapolation (REx) (Krueger et al., 2021).

In Table 2, we show that our proposed method for domain adaptation in the multi-domain setting outperforms the state-of-the-art multi-domain generalization methods.

D.5 Regression Tasks

We consider three tasks. We will first introduce the simulated task and then discuss about the task on dSprite data (Matthey et al., 2017).

D.5.1 Simulated Dataset

We consider the following data generation process.

Simulated regression task 1.

$$\begin{aligned} U &= Ber(a); \\ X &= \mathcal{N}(0,1); \\ Y &= -X\mathbf{1}_{(U=0)} + X\mathbf{1}_{(U=1)}; \\ W &= \mathcal{N}(-1,0.01)\mathbf{1}_{(U=0)} + \mathcal{N}(1,0.01)\mathbf{1}_{(U=1)}. \end{aligned} \tag{D.3}$$

There are two source domains. We set $a = 0.1$ for source domain z_1 and $a = 0.9$ for source domain z_2. According to the data generation process (D.3), Y is mostly positively correlated with X in domain z_1 and negatively correlated with X in domain z_2. For each domain, we synthesized 2000 training samples and 1000 testing samples. We sweep across $a = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ in the target domain. We run 10 replications and the results shown in Figure 5. In the next task, we set U to be a continuous random variable following a Beta distribution.

In this task, we expect the Cat-ERM method to fail drastically as we anticipate that the predicted Y versus X is a flat line – the predicted result would be an average of the downward sloping line ($U = 0$) and upward sloping line ($U = 1$). The result in Figure 5 supports our hypothesis, as the mean squared error remains nearly flat as we vary the target distribution $Q(U)$.

Simulated regression task 2.

$$\begin{aligned} U &= Beta(a,b) \\ X &= \mathcal{N}(0,1) \\ Y &= (2U-1)X \\ W &= \mathcal{N}(-1,0.01)(1-U) + \mathcal{N}(1,0.01)U. \end{aligned}$$

There are two source domains, corresponding to two draws from $P(Z)$ which we write $z_r = (a, b)$. We set $a = 2, b = 4$ for the first source domain $r = 1$, and $a = 4, b = 2$ for the second source domain $r = 2$. The corresponding distributions over U are shown in Figure 6. Under this setting, we test the target domain with $a, b = 1, \ldots, 5$, with distributions shown in Figure 6. For each domain, we synthesized 2000 training samples and 1000 testing samples. We run 10 replications and the results shown in Figure 5.

D.6 Adaptation with Concepts and Proxies

D.6.1 dSprites Dataset

We test the proposed procedure on the dSprites dataset (Matthey et al., 2017), an image dataset described by five latent parameters (shape, scale, rotation, posX, and posY). Motivated by Matthey et al. (2017)'s experiments, we design a regression task where the dSprites images ($64 \times 64 = 4096$-dimensional) are $X \in \mathbb{R}^{64\times 64}$ and subject to a nonlinear confounder $U \in [0, 2\pi]$ which is a rotation of the image (Figure 7). We fix all other latent parameters – shape is heart, scale is maximized, and all others are set to their 0'th position. $W \in \mathbb{R}$ and $C \in \mathbb{R}$ are continuous random variables. The data generation process is defined as follows

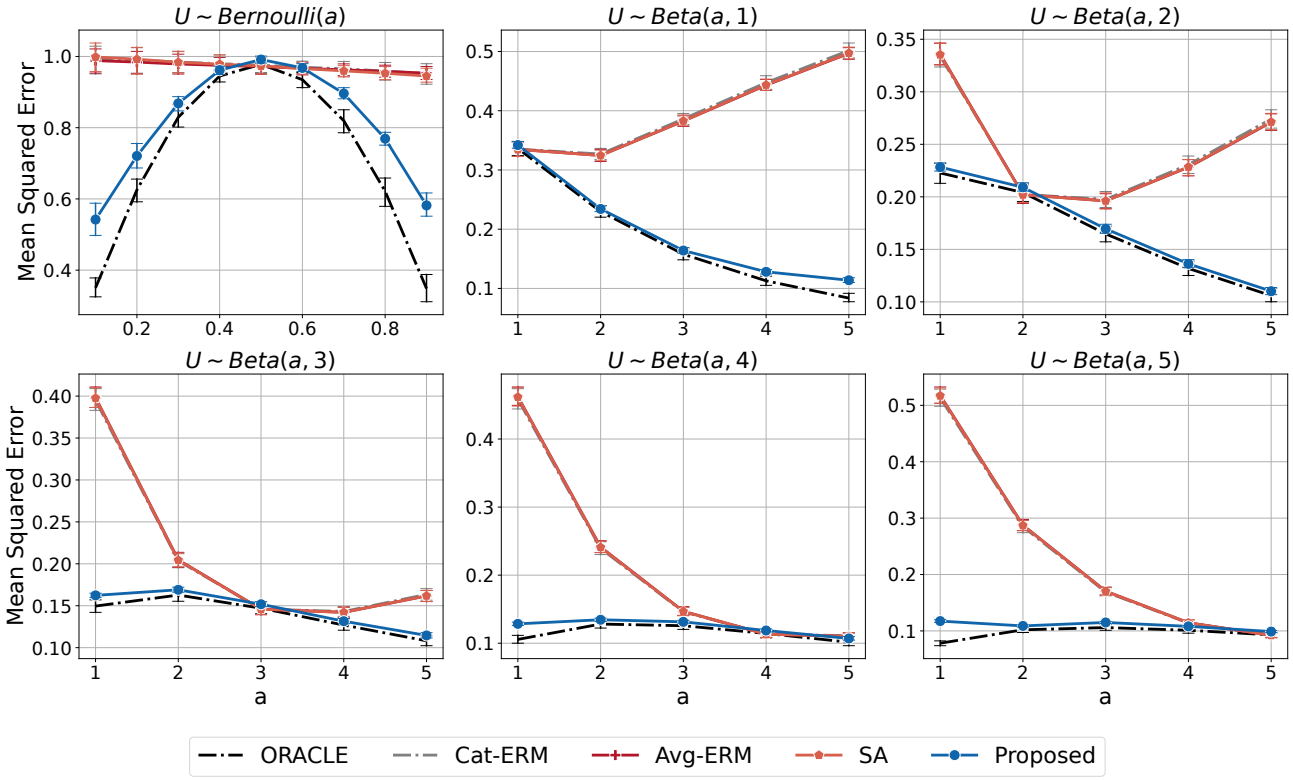


Figure 5: **Top left: results of regression task 1.** The proposed method is close to the ORACLE method as compared all other competing methods that is vulnerable to the distribution shifts. **Other figures: results of regression task 2.** In each plot, we fix b and vary a. For all plots, it appears that when $a = b$, the mean squared error of all methods converge to a point. This is the case when the target density function of U has a peak centered around 0.5, as shown in Figure 6, and hence $Y = (2U - 1)X$ is close to zero for most samples.

$$
\begin{aligned}
U^p &\sim 2\pi\text{Beta}(2,4), \quad U^q \sim \text{Uniform}(a, 2\pi);\\
X &= \text{Rotate}(\text{image}, U \text{ rads}) + \eta, \quad \eta \sim \mathcal{N}(0, 0.01 I_{64});\\
C &= \left(\frac{0.1\|X^T A\|_2^2 - 5000}{2000}\right)^2 + U + \gamma;\\
A &\sim \text{Uniform}(0,1),\ A \in \mathbb{R}^{4096\times 10}, \quad \gamma \sim \mathcal{N}(0, 0.5);\\
Y &= \frac{1}{4}C + \frac{1}{20}\sin(U) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 0.1);\\
W &= \cos(U) + \nu, \quad \nu \sim \mathcal{N}(0, 0.25).
\end{aligned}
$$

When fitting all model, both baselines and the proposed method, we project the images $\mathbb{R}^{4096}$ to $\mathbb{R}^{16}$ via Gaussian Random Projection using the *scikit-learn* implementation (Bingham and Mannila, 2001; Pedregosa et al., 2011). Additionally, for the proposed method, we use a Gaussian kernel as the feature map for X, C.

We generate 7000 training samples and 3000 test samples in our experiments. Then, we use five-fold cross-validation to select hyperparameters for baselines and proposed method for each a ($U^p \sim \text{Uniform}(a, 2\pi)$) – hyperparameters are (i) ridge regression penalty and (ii) Gaussian kernel scaling factor. Once we select a set of hyperparameters for a value of a, we perform 10 new random data regenerations to get transfer errors with 95% confidence intervals (Figure 2b).

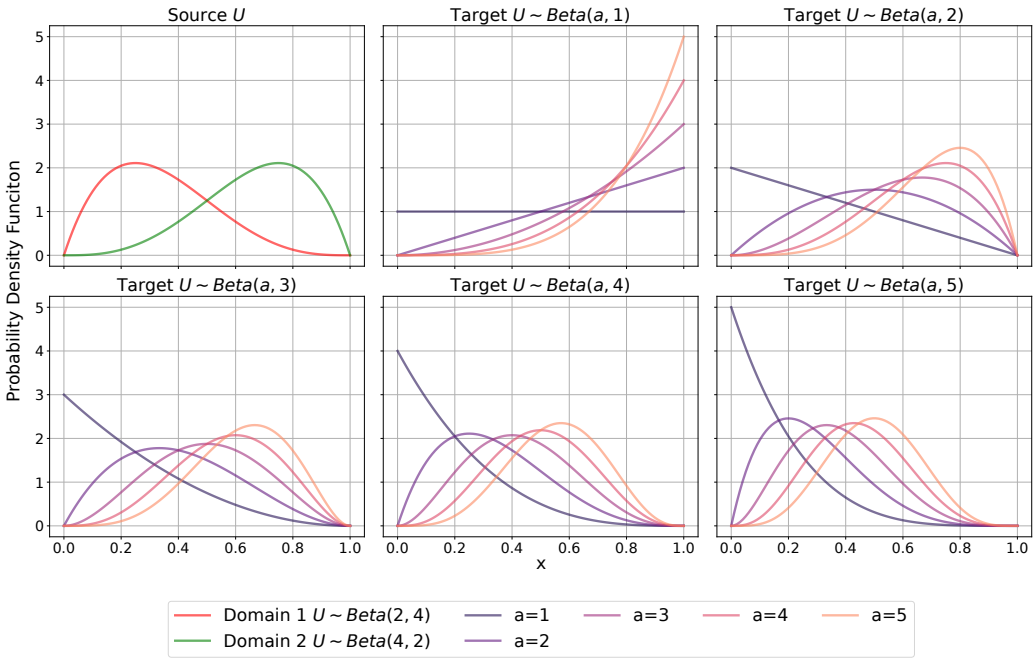


Figure 6: The probablity density function of Beta distributions with different $a, b = 1, \ldots, 5$.

Figure 7: dSprites image with confound (rotation) applied.

D.7 Classification of radiological findings with MIMIC-CXR

We conduct a small-scale experiment with chest X-ray data extracted from the MIMIC-CXR dataset (Johnson et al., 2019). We consider classification of the absence of a radiological finding in a chest X-ray. For this, we use the set of labels extracted by Irvin et al. (2019). These labels correspond to 14 categories of radiological findings extracted based on mentions in the associated radiology reports. We specifically consider classification of the "No Finding" ($Y = 1$) label, corresponding to cases where no pathology was identified as positive or uncertain in the radiology report.

To define the dataset, we consider the set of 217,536 chest X-rays with defined Chexpert labels (Irvin et al., 2019), MIMIC-IV entries, and pretrained embeddings (Sellergren et al., 2022). We then filter this dataset to the 212,567 examples considered as a part of the "train" partition provided by the MIMIC-CXR database (Johnson et al., 2019). We then partition the data into training, validation, and testing splits such that 80%, 10%, and 10% of the examples belong to each partition, respectively. For adaptation, we consider BioBERT (Lee et al., 2020) 768-dimensional embeddings of the radiology reports as concepts C and the patient's age as a proxy variable W. For simplicity, we use the patient `anchor_age` defined through linkage to the MIMIC-IV database, regardless of the patient's age at the time of the chest X-ray. Similar to the dSprites experiment, we further reduce the dimensionality of X and C to $\mathbb{R}^{64}$ using Gaussian Random Projection fit on the full training partition (170,053 examples).

To define distribution shifts, we adopt a problem formulation similar to that of Makar et al. (2022), where patient sex is considered as a possible "shortcut" in the classification of the absence of a radiological finding. As in Makar et al. (2022), we impose distribution shift through structured resampling of the data where $P(U = 1) = P(Y = 1 \mid \text{Sex} = \text{Female}) = P(Y = 0 \mid \text{Sex} = \text{Male})$. For example, when $P(U = 1) = 0.1$, the prevalence of $P(Y = 1 \mid \text{Sex} = \text{Female}) = 0.1$ and $P(Y = 1 \mid \text{Sex} = \text{Male}) = 0.9$. We implement the shift through a weighted sampling procedure that maintains the label shift invariance within patient sex subgroups, i.e., preserves $X \mid Y, A$ under the distribution shift, where A corresponds to patient sex. This procedure further fixes the total proportion of male and female patients in the population at 50%. For our experiments, we consider nine domains corresponding to cases where $P(U = 1) \in \{0.1, 0.2, \ldots, 0.9\}$.

We perform both concept adaptation and multi-domain adaptation experiments with the MIMIC-CXR data. For the concept adaptation experiment, we perform weighted sampling with replacement of 1,000 examples from each of the training, validation, and testing partitions defined previously, separately for each domain. We fix the source domain to the case where $P(U = 1) = 0.1$ and then adapt to each of the nine target domains. For the multi-domain adaptation experiment, we randomly sample 500 examples per domain and partition from the sets of 1,000 examples defined for the concept experiment. For this experiment, we consider a case where two source domains corresponding to $P(U = 1) = 0.1$ and $P(U = 1) = 0.2$ are available. To match the size of the aggregate source domain data with the size of the target domain, we sample 250 examples per partition for each source domain. We repeat the sampling procedure five times and report the mean $\pm$ standard deviation of performance metrics over the five replicates.

For both experiments, we perform two-fold cross-validation for the kernel length-scale parameters using data from the source domain(s). Here, we compare to ridge logistic regression models fit in the source and target domains, with the ridge penalty fit with five-fold cross validation. We use **LR-Target** to refer to logistic regression models fit in a target domain, **LR-SOURCE** to refer to models fit in a source domain, and **Cat-LR** to refer to logistic regression models fit with concatenated data from the multiple source domains. We use **Bridge-SOURCE** to refer to the kernel estimator that leverages the bridge function (h_0 or m_0 for the concept and multi-domain adaptation settings, respectively) and conditional mean embedding ($\mu_{WC|x}$ or $\mu_{W|z,x}$) fit on the source domain data. **Bridge-TARGET** refers to the kernel estimator where both the bridge function and conditional mean embedding are fit on the target domain data.

References

I. Alabdulmohsin, N. Chiou, A. D'Amour, A. Gretton, S. Koyejo, M. J. Kusner, S. R. Pfohl, O. Salaudeen, J. Schrouff, and K. Tsai. Adapting to latent subgroup shifts via concepts and proxies. In *International Conference on Artificial Intelligence and Statistics*, pages 9637–9661. PMLR, 2023.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Knowledge Discovery and Data Mining*, 2001.

G. Blanchard, G. Lee, and C. Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24, 2011.

K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22 14:e49–57, 2006.

M. Carrasco, J.-P. Florens, and E. Renault. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of econometrics*, 6:5633–5751, 2007.

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

A. E. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

M. Long, Z. Cao, J. Wang, and M. I. Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

M. Makar, B. Packer, D. Moldovan, D. Blalock, Y. Halpern, and A. D'Amour. Causally motivated shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics*, pages 739–766. PMLR, 2022.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation with multiple sources. *Advances in neural information processing systems*, 21, 2008.

A. Mastouri, Y. Zhu, L. Gultchin, A. Korba, R. Silva, M. Kusner, A. Gretton, and K. Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *International conference on machine learning*, pages 7512–7523. PMLR, 2021.

L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

W. Miao, Z. Geng, and E. J. Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

K. Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. In *International conference on machine learning*, pages 10–18. PMLR, 2013.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

A. B. Sellergren, C. Chen, Z. Nabulsi, Y. Li, A. Maschinot, A. Sarna, J. Huang, C. Lau, S. R. Kalidindi, M. Etemadi, et al. Simplified transfer learning for chest radiography models using less data. *Radiology*, 305(2): 454–465, 2022.

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

L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968, 2009.

B. Sun and K. Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 443–450. Springer, 2016.

K. Zhang, M. Gong, and B. Schölkopf. Multi-source domain adaptation: A causal view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

M. Zhang, H. Marklund, N. Dhawan, A. Gupta, S. Levine, and C. Finn. Adaptive risk minimization: Learning to adapt to domain shift. *Advances in Neural Information Processing Systems*, 34:23664–23678, 2021.