Journal of Machine Learning Research 22 (2021) 1-80 Submitted 10/20; Revised 11/21; Published 11/21

Domain adaptation under structural causal models

Yuansi Chen YUANSI.CHEN@STAT.MATH.ETHZ.CH

Peter Bühlmann BUHLMANN@STAT.MATH.ETHZ.CH

Seminar for Statistics
ETH Zürich
Zürich, Switzerland

Editor: Isabelle Guyon

Abstract

Domain adaptation (DA) arises as an important problem in statistical machine learning when the source data used to train a model is different from the target data used to test the model. Recent advances in DA have mainly been application-driven and have largely relied on the idea of a common subspace for source and target data. To understand the empirical successes and failures of DA methods, we propose a theoretical framework via structural causal models that enables analysis and comparison of the prediction performance of DA methods. This framework also allows us to itemize the assumptions needed for the DA methods to have a low target error. Additionally, with insights from our theory, we propose a new DA method called CIRM that outperforms existing DA methods when both the covariates and label distributions are perturbed in the target data. We complement the theoretical analysis with extensive simulations to show the necessity of the devised assumptions. Reproducible synthetic and real data experiments are also provided to illustrate the strengths and weaknesses of DA methods when parts of the assumptions in our theory are violated.

Keywords: anticausal, conditionally invariant components, domain generalization, domain invariant projection, label shift, structural equation models

1. Introduction

Domain adaptation (DA) is a statistical machine learning problem in which one aims at learning a model from a labeled source dataset and expecting it to perform well on an unlabeled target dataset drawn from a different but related data distribution. Domain adaptation is considered a sub-field of transfer learning and also a sub-field of semi-supervised learning (Pan and Yang, 2009). The possibility of DA is inspired by the human ability to apply knowledge acquired on previous tasks to unseen tasks with minimal or no supervision. For example, it is common to believe that humans who learned driving in sunny days would adapt their skills to drive reasonably well in a rainy day without additional training. However, the scenario where the source and target data distribution is different (e.g. sunny vs. rainy) is difficult to handle for many machine learning systems. This is mainly because the classical statistical learning theory mostly focuses on statistical learning methods and guarantees when the training and test data are generated from the same distribution.

While existing DA theory is limited, more and more application scenarios have emerged where DA is needed and useful. DA is desired especially when obtaining unlabeled data is

cheap while labeling data is difficult. The difficulties of labeling data typically arise due to the required human expertise or the large amount of human labor. For example, to annotate the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset (Russakovsky et al., 2015) with more than 1.2 million labeled images, it costs an average worker on the Amazon Mechanical Turk (www.mturk.com) about 0.5 seconds per image (Fei-Fei, 2010). Thus, annotating more than 1.2 million images requires more than 150 human hours. The actual time needed for labeling is much longer, due to the additional time spent on label verification and on labeling the images which are not used in the final challenge. Now if one is only interested in a similar but different task such as classifying objects in oil paintings, one cannot expect the ILSVRC dataset which contains pictures of natural objects to be representative. As a consequence, one needs to collect new data. It is relatively easy to collect digital images for the new task, but it is costly to label them if human experts have to be involved. Similar situations where the labeling is costly emerge in many other fields such as part-of-speech tagging (Ratnaparkhi, 1996; Ben-David et al., 2007), web-page classification (Blum and Mitchell, 1998), etc.

Despite the broad need of DA methods in practice, a priori it is impossible to provide a generic solution to DA. In fact, the DA problem is ill-posed if assumptions on the relationship between the source and target datasets are absent. One cannot learn a model with good performance if the target dataset can be arbitrary. As it is often difficult to specify the relationship between the source and target datasets, a large body of existing DA work are driven by applications. This line of work often focuses on developing new DA methods to solve very specific data problems. Despite the empirical success on these specific problems, it is not clear why DA succeeds or how applicable the proposed DA methods are to new related data problems. The growing development of domain adaptation calls forth a theoretical framework to analyze the existing methods and to guide the design of new procedures. More specifically, can one formulate the assumptions needed for a DA method to have a low target error? Can these assumptions be itemized, so that once specified the performance of different DA methods can be compared? More specifically, does the causal direction of the data generation or the type and location of data perturbation (Figure 1) affect the choice of the best DA method?

To answer the above questions, this work develops a theoretical framework via structural causal models that enables the comparison of various existing DA methods. Since there are no clear winning DA methods in general, the performance of DA methods has to be analyzed and compared with precise assumptions on the underlying data structure. Through analysis on simple models, we aim to give insights on when and why one DA method outperform others.

Our contributions: Our contributions are three-fold. First, we develop a theoretical framework via structural causal models (SCM) to analyze and compare the prediction performance of existing DA methods, such as domain invariant projection (DIP) (Pan et al., 2010; Baktashmotlagh et al., 2013) and conditional invariance penalty (CIP) or conditional transferable component (Gong et al., 2016; Heinze-Deml and Meinshausen, 2021), under precise assumptions relating source and target data. In particular, we show that under linear SCM the popular DA method DIP is guaranteed to have a low target error when the prediction problem is anticausal without label distribution perturbation. However, DIP

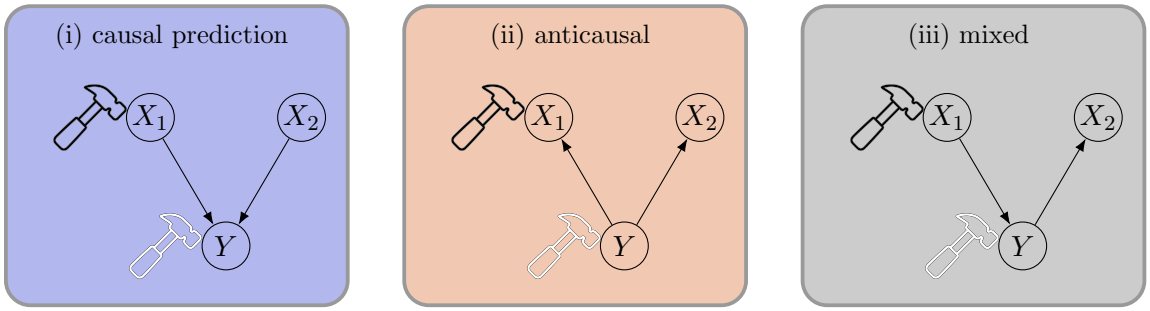


Figure 1: (i) causal prediction setting where the label Y is the descendant of the covariates X. The hammers indicate the location of data perturbation. (2) anti-causal prediction setting where all the covariates X are descendants of the label Y. (3) mixed-causal-anticausal setting where some covariates are descendants of the label Y and some covariates are ancestors of Y. Would the causal direction of the data generation affect DA performance? Would changing the location of the perturbation (black hammer vs. white hammer) cause some DA methods to fail?

fails to outperform the estimator trained solely on the source data when there are perturbations on the label distribution or when the prediction problem is either causal or mixed-causal-anticausal. Second, based on our theory, we introduce a new DA method called CIRM and develop its variants which can have better prediction performance than DIP in mixed-causal-anticausal DA scenarios and those with label distribution perturbation. Third, we illustrate via extensive simulations and real data experiments that our theoretical DA framework enables a better understanding of the success and failure of DA methods even in cases where presumably not all assumptions in our theory are satisfied. Our theory and experiments make it clear that knowing the relevant information about the data generation process such as the causal direction or the existence of label distribution perturbation is the key to the success of domain adaptation.

The rest of the paper is organized as follows. In Section 2 we review previous ways to mathematically formulate the DA problem and summarize the existing DA methods. Section 3 contains background on structural causal models, our problem setup, a formal introduction of the DA methods to study and three simple motivating examples. In Section 4 we analyze and compare the performance of three DA methods (DIP, CIP and CIRM) under our theoretical framework. Based on whether the DA problem is causal, anticausal or mixed and whether there is label distribution perturbation, we identify scenarios where these DA methods are guaranteed to have low target errors. Section 5 contains numerical experiments on synthetic and real datasets to illustrate the validity of our theoretical results and the implication of our theory for practical scenarios.

2. Related work

Domain adaptation is a sub-field of the broader research area of transfer learning. More specifically, domain adaptation is also named transductive transfer learning (Pan and Yang, 2009; Redko et al., 2020). In this paper, we concentrate on the DA problem without diving

into the broader field of transfer learning. Consequently, many previous work on transfer learning are omitted for the sake of space. We direct the interested readers to the survey paper by Pan and Yang (2009) and references therein for a literature review on transfer learning. Focusing on the DA problem, we first review existing theoretical frameworks of DA and then provide an overview of DA methods and algorithms.

Theoretical DA frameworks: The DA problem is ill-posed if one does not specify any assumptions on the relationship between the source and target data distribution. For this reason, depending on how this relationship is specified, many ways to formulate the DA problem have been introduced.

Ben-David et al. (2007) were the first to provide a DA prediction performance bound via Vapnik-Chervonenkis (VC) theory for classifiers from a generic hypothesis class. Since this bound is obtained without making explicit assumptions on the relationship between the source and target data distribution, it involves a divergence term that characterizes the closeness of the source and target distribution. A follow-up from Ben-David et al. (2010) further formally proves the necessity of assuming similarity between source and target distribution to ensure learnability. Ben-David et al.'s work laid the foundation of many further studies that attempt to bound the target and source prediction performance difference via divergence measures (Mansour et al., 2009; Cortes and Mohri, 2011, 2014; Cortes et al., 2015; Hoffman et al., 2018). See also the survey paper by Redko et al. (2020) for a complete review of VC-theory-type DA prediction performance bounds.

One natural way to explicitly relate the source and target data distribution is to assume that both of them are generated via the same well-specified generative model and to treat the missing target label problem as a missing data problem. Given a well-specified probabilistic generative model, the problem of imputing the missing data is well-studied and it is commonly solved via the expectation maximization (EM) algorithm (see McLachlan and Krishnan, 2007). This idea of casting a DA problem to a missing data problem has been introduced in Amini and Gallinari (2003) and Nigam et al. (2006).

It is also possible to loosely relate the source and target data distribution by assuming that they differ only by a small amount. How to specify this "small amount" depends on applications. If one assumes that the source data distribution is contaminated so that it is ϵ total variation distance away from the target data distribution, then the problem goes back to the classical robust statistics literature (Huber, 1964; Yuan et al., 2019). More recently, Wasserstein distance or f-divergence have been considered to describe the difference between source and target data distribution. These work and contributions are referred to as distributional robust learning (Sinha et al., 2018; Duchi and Namkoong, 2018; Gao et al., 2017). A closely related line of work directly assumes that target data points can be interpreted as source data points contaminated with arbitrary small additive noise quantified via norm constraints (ℓ_1 or ℓ_∞). This direction is called adversarial machine learning (Goodfellow et al., 2018; Raghunathan et al., 2018).

Another way to make the DA problem tractable is to assume that the conditional distribution $Y \mid X$ is invariant across source and target data, where Y and X denote the response (label) and covariates, respectively. The only difference between source and target distributions comes from the change in the distribution of the covariates X. This type of assumption is called the *covariate shift* assumption (Quionero-Candela et al., 2009; Sugiyama

and Kawanabe, 2012; Storkey, 2009). Alternatively, assuming the other conditional distribution $X \mid Y$ distribution to be invariant is also plausible in certain applications. This approach has a similar name called the *label shift* assumption (Lipton et al., 2018; Azizzadenesheli et al., 2019; Garg et al., 2020).

Finally on the causality side, it has been pointed out by Pearl and Bareinboim (2014) that full specification of a structural causal model (SCM) allows to study transportability of learning methods on the relationship between variables in the structural causal model. We refer to this approach as full-SCM transfer learning. The full-SCM transfer learning approach is very powerful in describing many data generation models. However, the main drawback of this framework is that the full specification of the structural causal model might be difficult to learn in many applications with limited data. On the other hand, the pioneering work by Schölkopf et al. (2012) reveals that distinguishing between the causal and anticausal prediction may already be useful to facilitate the selection of DA and semi-supervised learning methods. Figuring out the right amount of causal information needed to carry out DA is one of our main motivations in this paper.

Previous DA methods: While it is in general helpful to have theoretical DA frameworks to relate the source and target data, theory is not essential for the development of new DA methods. A large number of DA methods and algorithms were introduced with the focus of addressing DA for specific datasets. Here we highlight several popular ones.

Self-training (Amini and Gallinari, 2003) is one of the earliest DA methods which is originated in the semi-supervised learning literature (see the book by Chapelle et al., 2009). The self-training algorithm begins with an estimator trained on the source data, and gradually labels a part of unlabeled target data and then updates the estimator with appropriate regularization after combining newly labeled target data. It has been shown to have good empirical performance on several computer vision domain adaptation tasks with small labeled source datasets (Xie et al., 2020; Carmon et al., 2019). A theoretical analysis of the performance of self-training under a gradual shift assumption on Gaussian mixture data was recently provided by Kumar et al. (2020).

An important line of DA methods relate the source and target data by assuming the existence of a common subspace. Pan et al. (2010) first came up with the idea of projecting the source and target data onto a reproducing kernel Hilbert space to preserve common properties and applied the idea to text classification datasets. The existence of an intermediate subspace that relates source and target data was made precise in Gopalan et al. (2011) for visual objection recognition datasets. This method was further developed and analyzed for sentiment analysis and web-page classification (Blitzer et al., 2011; Gong et al., 2012; Muandet et al., 2013). Baktashmotlagh et al. (2013) simplified the idea of enforcing a common subspace to adding a regularization term based on the maximum mean discrepancy (MMD) (Gretton et al., 2012). Their method is named domain invariant projection (DIP) because the regularization term enforces a projection of the source and target data on the subspace to be invariant in distribution. Recently, with the development of deep neural networks and the introduction of generative adversarial nets based distributional distance measures, the common subspace approach was further extended to allow for neural network implementations (Ganin et al., 2016; Peng et al., 2019).

The empirical success of DIP like DA methods and their neural network variants such as in Ganin et al. (2016) has sparked a wide discussion on the general validity of these methods. Zhao et al. (2019) constructed a simple counterexample showing that domain invariant projection is not sufficient to guarantee successful domain adaptation. Furthermore, Zhao et al. (2019) provided a lower bound of the target error when the label distribution is perturbed. Johansson et al. (2019) illustrated via a simple example the danger of the blind use of distributional distance such as MMD for domain invariant penalization. Li et al. (2019) and Tachet des Combes et al. (2020) discussed the failure of DIP in the presence of target label perturbation and proposed label shift correction using conditional invariant features. However, it is not clear whether their proposed algorithms have any guarantees for estimating the conditional invariant features or achieving low target error. The mixed messages about the success and failure of DIP like DA methods motivate us to set up rigorous target error comparisons of DIP and other DA methods.

Another line of DA methods that is worth mentioning is the one that only makes use of source data. Gong et al. (2016) introduced conditional transferable components which consist of features that are have invariant distributions given the label. The search of conditional transferable components is achieved via a penalty that matches the conditional distribution for any label across source environments. A related idea was proposed by Heinze-Deml and Meinshausen (2021). They extract the conditionally invariant (or core) components (CICs) across data points that share the same identifier but have different style features. The invariance is enforced via adding a conditional variance penalty to the training loss. Enforcing the conditional invariance allows them to learn models that are robust across perturbed computer vision datasets. Later, other concepts of invariance beyond conditional invariance across source environments were developed, such as invariant risk minimization (Arjovsky et al., 2019). It should be noted that, though with a different focus, the idea of using the heterogeneity across multiple source datasets to learn invariant or robust models has also appeared in the causal inference literature (Peters et al., 2016; Meinshausen, 2018; Rothenhäusler et al., 2021). Based on these ideas, Rojas-Carulla et al. (2018) and Magliacane et al. (2018) provided guarantees for domain adaptation under the assumption that the conditional distribution of the label given some subset of covariates is invariant.

There are many other interesting DA methods that are less related to our work. For the sake of space, we direct the interested readers to the book by Chapelle et al. (2009) on semi-supervised learning and other surveys (Zhu, 2005; Wilson and Cook, 2020; Wang and Deng, 2018) for additional references.

3. Preliminaries and problem setup

In this section, we first provide a brief summary of structural causal models, which are essential components of our theoretical framework. Then we formalize our domain adaptation problem setup and introduce the DA methods we study. Finally, we provide three simple motivating examples to illustrate the need of a DA theory.

3.1 Background on structural causal models

Structural causal models (SCMs) (see Pearl, 2000) are introduced to describe causal relationships between variables in a system. A SCM integrates the structural equation models

(SEMs) used in economics and social sciences, the potential-outcome framework of Neyman (1923) and Rubin (1974), and the graphical models developed for probabilistic reasoning and causal analysis. A SCM can be seen as a set of generative equations that describe not only the data generation process of the observational data, but also that of the intervention data. We refer the readers to Chapter 7 of the book by Pearl (2000) for a detailed description of SCM for causal inference. In the context of DA, SCMs can be used to describe both the data generation process of the source and target domains (or environments). The SCMs are specified via a set of structural equations with a corresponding causal graph to describe the relationship between variables.

While SCMs are very powerful tools to describe data generation processes in interventional environments, fully specifying a SCM for a DA problem has two main drawbacks in practice: first, defining the functional forms that relate variables in a SCM can be difficult for data involving many variables; second, even if the functional forms are specified, learning all the functions from data may result in a more complicated statistical learning task than the original DA problem. Focusing on solving DA problems, the way we address the two main drawbacks differentiates our work from the full-SCM transfer learning approach by Pearl and Bareinboim (2014).

To address the first drawback and to ease our theoretical analysis, we adopt a simplification that replaces all the functional models in the SCM with linear models as in previous work (Peters et al., 2016; Rothenhäusler et al., 2021, 2019). We call this simplified SCM a linear SCM. The simplification allows us to develop rigorous DA theory and to study more complicated DA problems as extensions of the linear case. Regarding the second drawback, we focus on DA methods that can be applied without relying on the full specification of the SCM. Only when we analyze the performance of these DA methods, we bring in SCMs to set forth the assumptions needed for the DA methods to have low target errors.

3.2 Domain adaptation problem setup

In this subsection, we set up the domain adaptation problem with M ($M \geq 1$) labeled source environments and one unlabeled target environment. Although SCMs are general enough to also handle classification problems, we focus our theory on regression to simplify the presentation. Later in Section 5, we show the adaptation of our results to the classification setting through numerical experiments.

For the m-th source environment ($m \in \{1, 2, \cdots, M\}$), we observe n_m i.i.d. samples $S^{(m)} = \left((x_1^{(m)}, y_1^{(m)}), \cdots, (x_{n_m}^{(m)}, y_{n_m}^{(m)})\right)$ drawn from the source data distribution $\mathcal{P}^{(m)}$, with $(x_k^{(m)}, y_k^{(m)}) \in \mathbb{R}^{d+1}$ for each $k \in \{1, 2, \cdots, n_m\}$. The m-th dataset is also called the m-th source environment. Furthermore, there are $\tilde{n}$ i.i.d. samples $\tilde{S} = ((\tilde{x}_1, \tilde{y}_1), \cdots, (\tilde{x}_{\tilde{n}}, \tilde{y}_{\tilde{n}}))$ from the target distribution $\widetilde{\mathcal{P}}$, but we only observe the covariates $\tilde{S}_X = (\tilde{x}_1, \cdots, \tilde{x}_{\tilde{n}})$ from $\widetilde{\mathcal{P}}_X$. Here $\widetilde{\mathcal{P}}_X$ is used to denote the marginal distribution of $\widetilde{\mathcal{P}}$ on X. The goal of the DA problem is to estimate a function $f_\beta : \mathbb{R}^d \mapsto \mathbb{R}$, mapping covariates to the label, parametrized by $\beta \in \Theta$ so that the *target population risk* is "small". Here the parameter space Θ is a subset of a finite-dimensional space. The performance metric *target population risk* for an estimator f is defined as

$$\tilde{R}(f) = \mathbb{E}_{(X,Y)\sim\widetilde{\mathcal{P}}}\left[l(f(X), Y)\right], \tag{1}$$

where l is a loss function and it is set to the squared loss function $x \mapsto x^2$ if not specified otherwise. Similarly, we can define the m-th *source population risk* as

$$R^{(m)}(f) = \mathbb{E}_{(X,Y)\sim\mathcal{P}^{(m)}}\left[l(f(X), Y)\right]. \tag{2}$$

In addition to the risk on an absolute scale, one can quantify the target population risk achieved by an arbitrary estimator on a relative scale by comparing it with the oracle target population risk $\widetilde{R}(f_{\beta_{\text{oracle}}})$, where β_{oracle} is defined as

$$\beta_{\text{oracle}} \in \arg\min_{\beta\in\Theta} \mathbb{E}_{(X,Y)\sim\widetilde{\mathcal{P}}}\left[l(f_\beta(X), Y)\right].$$

If we don't assume any relationship between the source distribution $\mathcal{P}^{(m)}$ and the target distribution $\widetilde{\mathcal{P}}$, the target population risk of an estimator learned from the source and unlabeled target data can be arbitrarily larger than the oracle target population risk. Assumptions on the relationship between source and target distribution are needed to make the DA problem tractable. In this work, we consider the DA setting where source and target data are both generated through similar linear SCMs with additional structural assumptions on the interventions.

Domain adaptation under linear SCM with noise interventions: For $m \in \{1, 2, \cdots, M\}$, the data distribution $\mathcal{P}^{(m)}$ of the m-th source environment is specified by the following data generation equations on $\left(X^{(m)}, Y^{(m)}\right)$ from $\mathcal{P}^{(m)}$,

$$\begin{bmatrix} X^{(m)} \\ Y^{(m)} \end{bmatrix} = \begin{bmatrix} \mathbf{B} & b \\ \omega^\top & 0 \end{bmatrix} \begin{bmatrix} X^{(m)} \\ Y^{(m)} \end{bmatrix} + g(a^{(m)}, \varepsilon^{(m)}), \tag{3}$$

and the target data distribution $\widetilde{\mathcal{P}}$ is specified via the same equation except for the noise distribution,

$$\begin{bmatrix} \widetilde{X} \\ \widetilde{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{B} & b \\ \omega^\top & 0 \end{bmatrix} \begin{bmatrix} \widetilde{X} \\ \widetilde{Y} \end{bmatrix} + g(\widetilde{a}, \widetilde{\varepsilon}). \tag{4}$$

Here $\mathbf{B} \in \mathbb{R}^{d\times d}$ is an unknown constant matrix with zero diagonal such that $\mathbb{I}_d - \mathbf{B}$ is invertible, $b \in \mathbb{R}^d$ and $\omega \in \mathbb{R}^d$ are unknown constant vectors; $\varepsilon^{(m)}$ and $\widetilde{\varepsilon}$ are $d+1$ dimensional random vectors drawn from the same noise distribution $\mathcal{E}$; g is a fixed function to model the change (or intervention) across source and target environments; $a^{(m)} \in \mathbb{R}^{d+1}$ is an unknown (random or non-random) intervention that changes from one environment to another. Note that the assumption on the invertibility of $\mathbb{I}_d - \mathbf{B}$ ensures the uniqueness of the data generation given a draw of the noise $\varepsilon^{(m)}$. This assumption is in general weaker than requiring the corresponding causal graph to be directed acyclic (Rothenhäusler et al., 2019).

The only difference between the source and target data distribution is due to the difference in intervention $a^{(m)}$ and $\widetilde{a}$. According to the SCM, the way we specify the difference between source and target distribution is through the term $g(\widetilde{a}, \widetilde{\varepsilon})$. This type of intervention is often called noise intervention or soft intervention (Eberhardt and Scheines, 2007; Peters et al., 2016). As a concrete example, if the mean shift noise intervention is considered, then

we specify the function $g : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to \mathbb{R}^{d+1}$ as $(a, \varepsilon) \mapsto a+\varepsilon$, and define the intervention term with a deterministic vector in $\mathbb{R}^{d+1}$. As another example, if the variance shift noise intervention is considered, then we specify g as $(a, \varepsilon) \mapsto a \odot \varepsilon$, where $\odot$ is the element-wise product. We focus our theoretical results on the mean shift noise intervention. Other types of noise interventions are discussed in numerical experiments. The linear SCM with noise interventions clearly does not cover all kinds of perturbations to the data. For example, our theory does not apply when the matrix $\mathbf{B}$ is no longer invariant across source and target environments. We show via numerical experiments that assuming that the perturbations are due to noise interventions is plausible in many settings.

3.3 Oracle and baseline DA methods

As we aim to compare existing and new DA methods rigorously under our theoretical framework, it is useful to start with several estimators to establish the basis of comparison.

First, we introduce two oracle estimators. They are defined using the unobserved information such as target labels or SCM parameters.

- **OLSTar:** the population ordinary least squares (OLS) estimator on the target data.

$$\begin{aligned} f_{\text{OLSTar}}(x) &:= x^\top \beta_{\text{OLSTar}} + \beta_{\text{OLSTar},0} \\ \beta_{\text{OLSTar}}, \beta_{\text{OLSTar},0} &:= \underset{\beta,\beta_0}{\arg\min}\, \mathbb{E}_{(X,Y)\sim\widetilde{\mathcal{P}}} \left(Y - X^\top \beta - \beta_0\right)^2. \end{aligned} \tag{5}$$

  This is the oracle target population estimator when we restrict the function class to be linear. Hence the target risk of OLSTar defines the lowest target risk that any linear DA estimator can achieve.

- **Causal:** the population causal estimator via the linear SCM

$$\begin{aligned} f_{\text{Causal}}(x) &:= x^\top \beta_{\text{Causal}} \\ \beta_{\text{Causal}} &:= \omega, \end{aligned} \tag{6}$$

  where ω appeared in the last row of the SCM matrix in Equation (3). Note that this formulation of the causal estimator assumes that there is no intervention on Y and the intercept is also zero. The Causal estimator is closely related to distributional robust estimators. That is, the Causal estimator is the robust estimator which achieves the minimum worse-case risk when the perturbations on the covariates are allowed to be arbitrary (Bühlmann, 2020). However, in our DA setting where observed target covariates provide additional information, it is no longer clear whether Causal still achieves the lowest target risk.

Second, we introduce two population estimators that only use the source data.

- **OLSSrc$^{(m)}$:** the population OLS estimator on the single m-th source environment.

$$\begin{aligned} f^{(m)}_{\text{OLSSrc}}(x) &:= x^\top \beta^{(m)}_{\text{OLSSrc}} + \beta^{(m)}_{\text{OLSSrc},0} \\ \beta^{(m)}_{\text{OLSSrc}}, \beta^{(m)}_{\text{OLSSrc},0} &:= \underset{\beta,\beta_0}{\arg\min}\, \mathbb{E}_{(X,Y)\sim\mathcal{P}^{(m)}} \left(Y - X^\top \beta - \beta_0\right)^2. \end{aligned} \tag{7}$$

- **SrcPool:** the population OLS estimator by pooling all source data together.

$$f_{\text{SrcPool}}(x) := x^\top \beta_{\text{SrcPool}} + \beta_{\text{SrcPool},0}$$
$$\beta_{\text{SrcPool}}, \beta_{\text{SrcPool},0} := \arg\min_{\beta,\beta_0} \mathbb{E}_{(X,Y)\sim\mathcal{P}^{\text{allsrc}}} \left(Y - X^\top\beta - \beta_0\right)^2, \tag{8}$$

  where $\mathcal{P}^{\text{allsrc}}$ is the uniform mixture over M source distributions $\left\{\mathcal{P}^{(m)}\right\}_{m=\{1,\cdots,M\}}$. We omitted the SrcPool formulation with weighted mixtures because it is not the main focus of our study.

These two estimators are natural estimators in the classical statistical learning setting when the source and target data share the same distribution. A DA method that has larger target risk than SrcPool is clearly unfavorable. One goal throughout our paper is to understand under which conditions DA methods are guaranteed to outperform OLSSrc$^{(m)}$ and SrcPool.

3.4 Advanced DA methods

In this subsection, we introduce three advanced DA methods. The first two have been introduced previously and the third one is our new DA method.

First, we consider the DA method called *domain invariant projection* (DIP). DIP is a subspace-based DA method which aims at learning a common intermediate subspace that relates the source and target domains. The specific form of DIP we study follows from Baktashmotlagh et al. (2013). Specifically, the population DIP estimator involves the following optimization problem

$$f_{\text{DIP}}(x) := u_{\text{DIP}} \circ v_{\text{DIP}}(x)$$
$$u_{\text{DIP}}, v_{\text{DIP}} := \arg\min_{u\in\mathcal{U}, v\in\mathcal{V}} \mathbb{E}_{(X,Y)\sim\mathcal{P}^{(1)},\tilde{X}\sim\tilde{\mathcal{P}}} l(u\circ v(X), Y) + \lambda \cdot \mathcal{D}(v(X), v(\tilde{X})), \tag{9}$$

where $\mathcal{D}(\cdot,\cdot)$ measures the distance between two distribution, $\mathcal{U}$ and $\mathcal{V}$ are function classes that are specified through expert knowledge of the problem and λ is a positive regularization parameter. DIP, in its simple form, only uses a single source environment. Hence, without loss of generality, we used the first source data environment.

In Baktashmotlagh et al. (2013), maximum mean discrepancy (MMD) is used as the distributional distance measure and both $\mathcal{U}$ and $\mathcal{V}$ are set to be linear mappings. This DIP idea of matching the mappings of source and target data is extended later in many DA papers with various choices of distance and function classes (see e.g. Ghifary et al., 2016; Li et al., 2018). A noteworthy line of follow-up work consist of replacing the distribution distance and function classes in DIP with neural networks. For example, Ganin et al. (2016) introduced domain-adversarial neural network (DANN) which uses generative adversarial nets (GAN) in place of MMD to measure distributional distance and makes both $\mathcal{U}$ and $\mathcal{V}$ to be neural networks.

Analyzing the most generic form of DIP is out of the scope of this study. Instead, we start with a simple DIP formulation.

- **DIP$^{(m)}$-mean:** the population DIP estimator where mean squared difference is used as distributional distance, $\mathcal{V}$ is linear and $\mathcal{U}$ is the singleton of the identity mapping

and λ is chosen to be ∞. DIP, in its simple form, only uses the data from one source environment and the target covariates. This form of DIP is defined as

$$\begin{aligned} f_{\text{DIP}}^{(m)}(x) &:= x^\top \beta_{\text{DIP}}^{(m)} + \beta_{\text{DIP},0}^{(m)} \\ \beta_{\text{DIP}}^{(m)}, \beta_{\text{DIP},0}^{(m)} &:= \underset{\beta,\beta_0}{\arg\min}\ \mathbb{E}_{(X,Y)\sim\mathcal{P}^{(m)}} \left(Y - X^\top\beta - \beta_0\right)^2 \\ &\quad \text{s.t. } \mathbb{E}_{X\sim\mathcal{P}_X^{(m)}} \left[X^\top\beta\right] = \mathbb{E}_{X\sim\widetilde{\mathcal{P}}_X} \left[X^\top\beta\right]. \end{aligned} \tag{10}$$

For simplicity, we use the shorthand notation $\text{DIP}^{(m)}$ to refer to $\text{DIP}^{(m)}$-mean. The constraint in Equation (10) is called the DIP matching penalty.

Second, we introduce the *conditional invariant penalty* (CIP) estimator. Unlike DIP which projects the source and target covariates to the same subspace, CIP directly uses the label information in multiple source environments to look for the conditionally invariant components. CIP only makes use of source data.

- **CIP-mean:** the population conditional invariance penalty (CIP) estimator where the conditional mean is matched across multiple source environments.

$$\begin{aligned} f_{\text{CIP}}(x) &:= x^\top \beta_{\text{CIP}} + \beta_{\text{CIP},0} \\ \beta_{\text{CIP}}, \beta_{\text{CIP},0} &:= \underset{\beta,\beta_0}{\arg\min}\ \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{(X,Y)\sim\mathcal{P}^{(m)}} \left(Y - X^\top\beta - \beta_0\right)^2 \\ \text{s.t. } \mathbb{E}_{(X,Y)\sim\mathcal{P}^{(m)}} \left[X^\top\beta \mid Y\right] &= \mathbb{E}_{(X,Y)\sim\mathcal{P}^{(1)}} \left[X^\top\beta \mid Y\right] a.s., \forall m \in \{2,\cdots,M\}, \end{aligned} \tag{11}$$

where the equality between the conditional expectation is in the sense of almost sure equality of random variables. Note that CIP naturally requires $M \geq 2$, because when $M = 1$ the CIP constraint becomes vacuous. For simplicity, we use the shorthand notation CIP to refer to CIP-mean.

CIP puts more regression weights on the conditionally invariant components from multiple source datasets via the conditional invariance constraint in Equation (11). The idea of conditional invariance penalty in the context of anticausal learning has appeared in multiple papers with slightly different settings. Gong et al. (2016) introduced conditional transferable components which are in fact conditionally invariant features. However, unlike the formulation above, Gong et al. (2016) propose to learn the conditional transferable components with only one source environment which requires assumptions that are difficult to check. On the other hand, the algorithm from Heinze-Deml and Meinshausen (2021) learns the conditionally invariant features from a single source dataset if multiple observations of data points that share the same identifier are present. They use their conditional variance penalty to enforce their algorithm to learn the conditionally invariant features in their specific datasets with identifiers. The same idea can be applied if we replace identifiers with source environments.

Third, we introduce our new DA estimator *conditional invariant residual matching* (CIRM).

- **CIRM$^{(m)}$-mean:** the population conditional invariant residual matching estimator that uses all source environments to compute CIP and the m-th source environment to perform risk minimization.

$$
\begin{aligned}
f^{(m)}_{\mathrm{CIRM}}(x) &:= x^\top \beta^{(m)}_{\mathrm{CIRM}} + \beta^{(m)}_{\mathrm{CIRM},0} \\
\beta^{(m)}_{\mathrm{CIRM}}, \beta^{(m)}_{\mathrm{CIRM},0} &:= \operatorname*{arg\,min}_{\beta,\beta_0} \ \mathbb{E}_{(X,Y)\sim\mathcal{P}^{(m)}} \left(Y - X^\top\beta - \beta_0\right)^2 \\
&\text{s.t. } \mathbb{E}_{X\sim\mathcal{P}^{(m)}_X} \left[\beta^\top \left(X - \left(X^\top\beta_{\mathrm{CIP}}\right)\vartheta_{\mathrm{CIRM}}\right)\right] = \mathbb{E}_{X\sim\widetilde{\mathcal{P}}_X} \left[\beta^\top \left(X - \left(X^\top\beta_{\mathrm{CIP}}\right)\vartheta_{\mathrm{CIRM}}\right)\right],
\end{aligned} \tag{12}
$$

  where

$$
\vartheta_{\mathrm{CIRM}} := \frac{\mathbb{E}_{(X,Y)\sim\mathcal{P}^{\mathrm{allsrc}}}\left[X\cdot(Y - \mathbb{E}[Y])\right]}{\mathbb{E}_{(X,Y)\sim\mathcal{P}^{\mathrm{allsrc}}}\left[(X^\top\beta_{\mathrm{CIP}} - \mathbb{E}[X^\top\beta_{\mathrm{CIP}}])\cdot(Y - \mathbb{E}[Y])\right]}, \tag{13}
$$

  with $\mathcal{P}^{\mathrm{allsrc}}$ denoting the uniform mixture of all source distributions. For simplicity, we use the shorthand notation CIRM$^{(m)}$ to refer to CIRM$^{(m)}$-mean.

At first glance, the CIRM estimator is a combination of DIP and CIP. CIRM first uses CIP to compute a linear combination of conditionally invariant components to serve as a proxy of the label Y. To tackle the label distribution perturbation, CIRM then performs the DIP-type matching on the residual obtained by regressing Y on its proxy. We can show that the joint distribution of the residual together with the covariates do not suffer from the label distribution perturbation. CIRM is designed to be applied in DA scenarios with label distribution perturbation. Note that the idea of using domain-invariant features to tackle label distribution shift was proposed previously in Li et al. (2019) and in Tachet des Combes et al. (2020), while it was not clear how to formulate the exact algorithm so that its success and failure conditions can be analyzed. The intuition behind the CIRM construction becomes clearer after we state Theorem 5.

The comparison of DA methods in the following sections is centered around DIP, CIP, CIRM and their variants. To keep track of these variants, we provide a summary of all DA methods appeared in this paper in Table 5 of Appendix A.1.

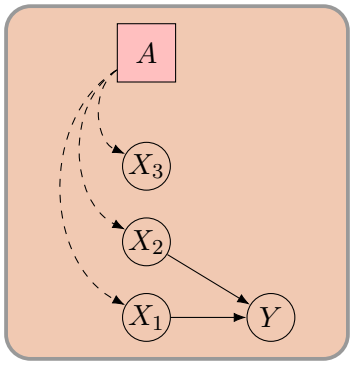
3.5 Simple motivating examples

Before we dive into theoretical comparisons of the DA methods, we go through three simple examples to illustrate the assumptions needed for DA methods to have low target risks. The simple examples have data generated via low-dimensional SCMs so that the DA estimators can be easily computed and understood.

3.5.1 Example 1: causal prediction

Example 1 has one source environment and one target environment. The data in the source and target environment are generated independently according to the following SCMs, with the causal diagram on the left and the structural equations on the right of Figure 2.

Here the noise variables follow independent Gaussian distributions with mean zero and variance 0.1 for X and variance 0.2 for Y, namely, $\varepsilon^{(1)}_{X_1}, \varepsilon^{(1)}_{X_2}, \varepsilon^{(1)}_{X_3}, \widetilde{\varepsilon}_{X_1}, \widetilde{\varepsilon}_{X_2}, \widetilde{\varepsilon}_{X_3} \sim \mathcal{N}(0, 0.1)$,



$$X_1^{(1)} = \varepsilon_{X_1}^{(1)} + 1 \qquad \widetilde{X}_1 = \widetilde{\varepsilon}_{X_1} - 1$$
$$X_2^{(1)} = \varepsilon_{X_2}^{(1)} + 1 \qquad \widetilde{X}_2 = \widetilde{\varepsilon}_{X_2} - 1$$
$$X_3^{(1)} = \varepsilon_{X_3}^{(1)} + 1 \qquad \widetilde{X}_3 = \widetilde{\varepsilon}_{X_3} + 1$$
$$Y^{(1)} = X_1^{(1)} + X_2^{(1)} + \varepsilon_Y^{(1)}, \qquad \widetilde{Y} = \widetilde{X}_1 + \widetilde{X}_2 + \widetilde{\varepsilon}_Y,$$

Figure 2: The causal diagram and structural equations for the source and target environments in Example 1: causal prediction.

$\varepsilon_Y^{(1)}, \widetilde{\varepsilon}_Y \sim \mathcal{N}(0, 0.2)$. The type of intervention is mean shift noise intervention. That is, the function g in Equation (3) is taken to be $g : (a^{(1)}, \varepsilon^{(1)}) \mapsto a^{(1)} + \varepsilon^{(1)}$. The intervention is $a^{(1)} = \begin{bmatrix} 1 & 1 & 1 & 0 \end{bmatrix}^\top$ for the source environment and it is $\widetilde{a} = \begin{bmatrix} -1 & -1 & 1 & 0 \end{bmatrix}^\top$ for the target environment. This example is called causal prediction, because the covariates X are parents of the label Y. In other words, we are predicting the effect from the causes as illustrated in the causal diagram in Figure 2.

Given the source and target distribution, the population DA estimators in the previous subsection can be computed explicitly. We obtain

$$\begin{bmatrix} \beta_{\text{OLSTar}} \\ \beta_{\text{OLSTar},0} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}^\top, \qquad \begin{bmatrix} \beta_{\text{Causal}} \\ \beta_{\text{Causal},0} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}^\top,$$
$$\begin{bmatrix} \beta_{\text{OLSSrc}}^{(1)} \\ \beta_{\text{OLSSrc},0}^{(1)} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}^\top, \qquad \begin{bmatrix} \beta_{\text{DIP}}^{(1)} \\ \beta_{\text{DIP},0}^{(1)} \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & -\frac{2}{3} & -2 \end{bmatrix}^\top.$$

Note that OLSSrc$^{(1)}$, Causal and OLSTar share the same estimate β, while DIP$^{(1)}$ does not. The corresponding population source and target risks are summarized in the first two rows of Table 1. DIP$^{(1)}$ has a larger target population risk than OLSSrc$^{(1)}$. In this example of causal prediction, using the additional target covariate information via DIP is making the target prediction performance worse. This is because DIP matching penalty in Equation (10) which is not satisfied by the oracle estimator OLSTar, makes DIP too restrictive.

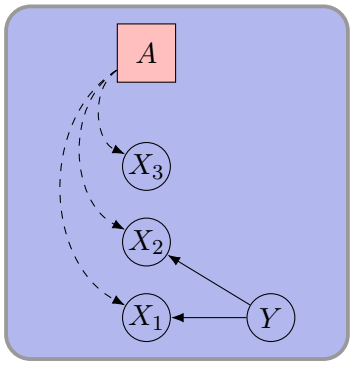
3.5.2 Example 2: anticausal prediction

Example 2 has one source environment and one target environment. The data in the source and target environment are generated independently according to the following SCMs, with the causal diagram on the left and the structural equations on the right of Figure 3.

Here the noise variables follow independent Gaussian distributions with mean zero and variance 0.1 for X and variance 0.2 for Y, namely, $\varepsilon_{X_1}^{(1)}, \widetilde{\varepsilon}_{X_1}, \varepsilon_{X_2}^{(1)}, \widetilde{\varepsilon}_{X_2}, \varepsilon_{X_3}^{(1)}, \widetilde{\varepsilon}_{X_3} \sim \mathcal{N}(0, 0.1)$, $\varepsilon_Y^{(1)}, \widetilde{\varepsilon}_Y \sim \mathcal{N}(0, 0.2)$. The type of intervention g is mean shift noise intervention, which is the same as in Example 1. The intervention is $a^{(1)} = \begin{bmatrix} 1 & 1 & 1 & 0 \end{bmatrix}^\top$ for the source environment

Risk \ Methods	**OLSTar** (oracle)	**Causal**	**OLSSrc**$^{(1)}$	**DIP**$^{(1)}$	**DIPAbs**$^{(1)}$
Ex 1, source risk $R^{(1)}$	0.200	0.200	0.200	0.333	-
Ex 1, target risk $\tilde{R}$	**0.200**	0.200	0.200	16.333	-
Ex 2, source risk $R^{(1)}$	2.600	0.200	0.040	0.086	0.044
Ex 2, target risk $\tilde{R}$	**0.040**	0.200	2.600	0.086	0.667
Ex 3, source risk $R^{(1)}$	0.200	0.200	0.040	0.066	-
Ex 3, target risk $\tilde{R}$	**0.040**	1.200	0.200	4.066	-

Table 1: Population source and target risks in two motivating examples (the lower the better). The oracle target risk is highlighted in bold. In Example 1 of causal prediction, DIP$^{(1)}$ performs worse than Causal and OLSSrc$^{(1)}$ in terms of target population risk. In Example 2 of anticausal prediction, DIP$^{(1)}$ performs better than Causal and OLSSrc$^{(1)}$, but it is still not as good as the oracle estimator. DIPAbs$^{(1)}$ is better in terms of population source risk can have worse target population risk than Causal. In Example 3 of anticausal prediction with intervention on Y, DIP$^{(1)}$ performs worse than OLSSrc$^{(1)}$ in terms of target population risk.



$$
\begin{aligned}
X_1^{(1)} &= Y^{(1)} + \varepsilon_{X_1}^{(1)} + 1 \\
X_2^{(1)} &= Y^{(1)} + \varepsilon_{X_2}^{(1)} + 1 \\
X_3^{(1)} &= \varepsilon_{X_3}^{(1)} + 1 \\
Y^{(1)} &= \varepsilon_Y^{(1)},
\end{aligned}
\qquad
\begin{aligned}
\widetilde{X}_1 &= \widetilde{Y} + \widetilde{\varepsilon}_{X_1} - 1 \\
\widetilde{X}_2 &= \widetilde{Y} + \widetilde{\varepsilon}_{X_2} - 1 \\
\widetilde{X}_3 &= \widetilde{\varepsilon}_{X_3} - 1 \\
\widetilde{Y} &= \widetilde{\varepsilon}_Y,
\end{aligned}
$$

Figure 3: The causal diagram and structural equations for the source and target environments in Example 2: anticausal prediction

and it is $\widetilde{a} = \begin{bmatrix} -1 & -1 & -1 & 0 \end{bmatrix}^\top$ for the target environment. Compared to Example 1, the main difference is that the causal direction between the covariates X and the label Y has changed. This example is called anticausal prediction, because the covariates X are descendants of the label Y. We are predicting the cause from the effects as illustrated in the causal diagram in Figure 3.

In addition to the DA estimators used in the previous example, we introduce one more variant of DIP, *DIPAbs*. It is a made-up estimator to show that arbitrary choice of the

function classes $\mathcal{U}, \mathcal{V}$ in DIP formulation (9) without consideration on the data generation process is in general a bad idea.

- **DIPAbs$^{(m)}$-mean:** the population DIP estimator where mean squared difference is used as distributional distance, $\mathcal{V}$ is element-wise absolute value followed linear mapping and $\mathcal{U}$ is singleton of identity mapping and regularization parameter λ is chosen to be ∞. For the m-th source environment, it is defined as

$$\begin{aligned} f^{(m)}_{\text{DIPAbs}}(x) &:= |x|^\top \beta^{(m)}_{\text{DIPAbs}} + \beta^{(m)}_{\text{DIPAbs},0} \\ \beta^{(m)}_{\text{DIPAbs}}, \beta^{(m)}_{\text{DIPAbs},0} &:= \underset{\beta,\beta_0}{\arg\min}\ \mathbb{E}_{(X,Y)\sim\mathcal{P}^{(m)}} \left(Y - |X|^\top \beta - \beta_0\right)^2 \\ &\qquad \text{s.t. } \mathbb{E}_{X\sim\mathcal{P}^{(m)}_X} \left[|X|^\top \beta\right] = \mathbb{E}_{X\sim\widetilde{\mathcal{P}}_X} \left[|X|^\top \beta\right]. \end{aligned} \tag{14}$$

The population estimators can be computed explicitly

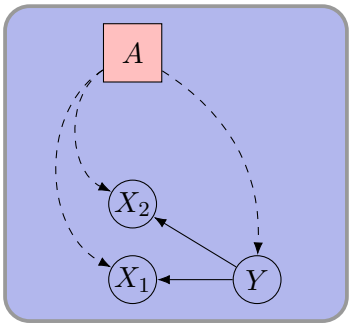
$$\begin{bmatrix} \beta_{\text{OLSTar}} \\ \beta_{\text{OLSTar},0} \end{bmatrix} = \begin{bmatrix} \frac{2}{5} & \frac{2}{5} & 0 & \frac{4}{5} \end{bmatrix}^\top, \qquad \begin{bmatrix} \beta_{\text{Causal}} \\ \beta_{\text{Causal},0} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}^\top,$$
$$\begin{bmatrix} \beta^{(1)}_{\text{OLSSrc}} \\ \beta^{(1)}_{\text{OLSSrc},0} \end{bmatrix} = \begin{bmatrix} \frac{2}{5} & \frac{2}{5} & 0 & -\frac{4}{5} \end{bmatrix}^\top, \qquad \begin{bmatrix} \beta^{(1)}_{\text{DIP}} \\ \beta^{(1)}_{\text{DIP},0} \end{bmatrix} = \begin{bmatrix} \frac{2}{7} & \frac{2}{7} & -\frac{4}{7} & 0 \end{bmatrix}^\top,$$
$$\begin{bmatrix} \beta^{(1)}_{\text{DIPAbs}} \\ \beta^{(1)}_{\text{DIPAbs},0} \end{bmatrix} = \begin{bmatrix} -\frac{2}{5} & -\frac{2}{5} & 0 & \frac{4}{5} \end{bmatrix}^\top.$$

For the other four estimators, none of the estimated β perfectly agrees with that of the oracle OLSTar. The corresponding population source and target risks are summarized in the third and fourth rows of Table 1. DIP$^{(1)}$ improves upon Causal and OLSSrc$^{(1)}$ on the target population risk. However its target population risk is not as small as that of the oracle estimator OLSTar. DIPAbs$^{(1)}$ also uses the target covariate information as DIP$^{(1)}$ does, but DIPAbs$^{(1)}$ performs worse than Causal or DIP$^{(1)}$ in terms of target population risk.

3.5.3 Example 3: anticausal prediction when Y is intervened on

Example 3 is also an anticausal prediction problem similar to Example 2, except that the label Y is also intervened on. The data in the source and target environment are generated independently according to the following SCMs, with the causal diagram on the left and the structural equations on the right of Figure 4.

Here the noise variables follow independent Gaussian distributions with $\varepsilon_{X_1}, \widetilde{\varepsilon}_{X_1} \sim \mathcal{N}(0, 0.1)$, $\varepsilon_{X_2}, \widetilde{\varepsilon}_{X_2} \sim \mathcal{N}(0, 0.1)$, $\varepsilon_Y, \widetilde{\varepsilon}_Y \sim \mathcal{N}(0, 0.2)$. The type of intervention is still mean shift noise intervention. The intervention is $a^{(1)} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^\top$ for the source environment and it is $\widetilde{a} = \begin{bmatrix} -1 & -1 & -1 \end{bmatrix}^\top$ for the target environment. Compared to Example 2, the main difference is that the intervention on Y is nonzero as illustrated in the causal diagram in Figure 4.



$$X_1^{(1)} = Y^{(1)} + \varepsilon_{X_1}^{(1)} + 1 \qquad \widetilde{X}_1 = \widetilde{Y} + \widetilde{\varepsilon}_{X_1} - 1$$
$$X_2^{(1)} = -Y^{(1)} + \varepsilon_{X_2}^{(1)} + 1 \qquad \widetilde{X}_2 = -\widetilde{Y} + \widetilde{\varepsilon}_{X_2} - 1$$
$$Y^{(1)} = \varepsilon_Y^{(1)} + 1, \qquad \widetilde{Y} = \widetilde{\varepsilon}_Y - 1,$$

Figure 4: The causal diagram and structural equations for the source and target environments in Example 3: anticausal prediction when Y is intervened on

The population estimators can be computed explicitly

$$\begin{bmatrix} \beta_{\text{OLSTar}} \\ \beta_{\text{OLSTar},0} \end{bmatrix} = \begin{bmatrix} \frac{2}{5} & -\frac{2}{5} & -\frac{1}{5} \end{bmatrix}^\top, \qquad \begin{bmatrix} \beta_{\text{Causal}} \\ \beta_{\text{Causal},0} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^\top,$$
$$\begin{bmatrix} \beta^{(1)}_{\text{OLSSrc}} \\ \beta^{(1)}_{\text{OLSSrc},0} \end{bmatrix} = \begin{bmatrix} \frac{2}{5} & -\frac{2}{5} & \frac{1}{5} \end{bmatrix}^\top, \qquad \begin{bmatrix} \beta^{(1)}_{\text{DIP}} \\ \beta^{(1)}_{\text{DIP},0} \end{bmatrix} = \begin{bmatrix} 0 & -\frac{2}{3} & 1 \end{bmatrix}^\top.$$

Note that $\text{DIP}^{(1)}$ has zero weight on the first coordinate but non-zero weight on the second because $X_2^{(1)}$ and $\widetilde{X}_2$ share the same distribution. The corresponding population source and target risks are summarized in the last two rows of Table 1. In Example 3 when Y is intervened on, $\text{DIP}^{(1)}$ is again worse than $\text{OLSSrc}^{(1)}$ on the target population risk.

3.5.4 Lessons from the simple motivating examples

The three simple motivating examples reveal three observations. First, even though DIP has a low target risk in the anticausal prediction setting (Example 2), DIP is not likely to outperform OLSSrc in the causal prediction setting (Example 1). The fact that the additional target covariate information is not useful in causal prediction problem was previously pointed out by Schölkopf et al. (2012). Consequently, DIP-type methods (Baktashmotlagh et al., 2013; Ganin et al., 2016) which make use of the additional target covariate information can make target performance worse in causal prediction problems. Second, not all DIP variants have low target risks in the anticausal prediction setting. $\text{DIPAbs}^{(1)}$, despite using the target covariate information as $\text{DIP}^{(1)}$ does, performs worse than Causal and $\text{DIP}^{(1)}$. In general, it is dangerous to treat DA methods as generic solutions that always work without consideration on the data generation process. We remark that $\text{DIPAbs}^{(1)}$ is a made-up method. But one could imagine a case where the first part $\mathcal{V}$ in $\text{DIP}^{(1)}$ is set to be a neural network that can approximate a large class of functions, then it is no longer clear the DIP matching penalty that matches the source and target covariate distributions always helps to improve the target population risk. This danger of learning blindly invariant representations was pointed out previously via a simple nonlinear data generation model by Zhao et al. (2019). Third, when Y is intervened on, DIP can perform worse than OLSSrc. Actually, none of the DA methods in Table 1 can handle the intervention on Y well. The

weakness of DIP in the presence of label perturbation has been observed previously in Zhao et al. (2019); Li et al. (2019); Tachet des Combes et al. (2020). Especially, Tachet des Combes et al. (2020) proposed the idea of using conditional invariant components for label shift correction. However, since their proposed method has no guarantees for estimating the conditional invariant components, it is not clear whether their methods have any guarantees for correcting the label shift.

The three examples are designed to demonstrate that the popular DA method DIP does not always outperform baseline methods such as OLSSrc or Causal. Besides, it also shows that the assumption of both source and data being generated from linear SCMs is not sufficient to guarantee DIP to have a low target risk. Additional assumptions on the data generation or on the intervention are needed. For example, knowing the causal direction of the data generation model or whether the label distribution is perturbed are all crucial information. Based on these examples, we ask the following questions:

1. In addition to the linear SCM assumption, what other assumptions are needed for DIP to perform better than Causal or OLSSrc? If such assumptions exist, can one quantify the gap between the target risk achieved by DIP and the oracle target risk?

2. If there is label distribution perturbation like in Example 3, are there DA methods that can outperform OLSSrc and have target risk guarantees?

3. If the prediction direction is a mix of causal and anticausal, is domain adaptation still beneficial?

In the sequel, we address these questions one by one. The first question is addressed in Section 4.1. The second question is dealt with in Section 4.2 via the introduction of CIP and CIRM. The general solution to the third question remains open. Naive applications of DIP and CIRM are not optimal. We provide a partial answer to the third question in Section 4.3 and show that the mixed-causal anticausal DA problem can be reduced to the anticausal DA problem when the causal variables have been already identified.

4. Domain adaptation with theoretical guarantees

In this section, to answer the three questions in the last section with rigor, we establish target risk guarantees for the DA estimators DIP, CIP and CIRM in three settings. Subsection 4.1 focuses on the DIP performance in the anticausal DA setting without intervention on Y. Subsection 4.2 demonstrates the difficulty of DIP in the anticausal DA setting with interventions on Y and then proves the advantage of CIP and CIRM over DIP when multiple source environments are available. Subsection 4.3 shows how domain adaptation is still possible in the mixed causal anticausal DA setting.

4.1 Anticausal domain adaptation without intervention on Y

In this subsection, we study the anticausal domain adaptation with the additional assumption that the intervention on the covariates X is a mean shift noise intervention and there is no intervention on the label Y. In the anticausal domain adaptation, all the covariates X are descendants of the label Y in the SCM. First, we derive the target risk bound for DIP

under these assumptions with a single source environment. Then we show how to make use of more source environments to improve the performance of DIP. Finally, we discuss ways to relax the mean shift noise intervention assumption.

4.1.1 Target risk guarantees of DIP with a single source environment

Before we state the main theorem, we introduce another oracle estimator to simplify the theorem statement. *DIPOracle* minimizes the mean squared error on target data while it uses the same matching penalty as DIP.

- **DIPOracle$^{(m)}$-mean:** the population DIP estimator which uses target labels and the covariate distribution from the m-th source and target environment.

$$\begin{aligned} f^{(m)}_{\text{DIPOracle}}(x) &:= x^\top \beta^{(m)}_{\text{DIPOracle}} + \beta^{(m)}_{\text{DIPOracle},0} \\ \beta^{(m)}_{\text{DIPOracle}}, \beta^{(m)}_{\text{DIPOracle},0} &:= \underset{\beta,\beta_0}{\arg\min}\ \mathbb{E}_{(X,Y)\sim\widetilde{\mathcal{P}}} \left(Y - X^\top\beta - \beta_0\right)^2 \\ &\quad \text{s.t. } \mathbb{E}_{X\sim\mathcal{P}^{(m)}_X}\left[X^\top\beta\right] = \mathbb{E}_{X\sim\widetilde{\mathcal{P}}_X}\left[X^\top\beta\right]. \end{aligned} \tag{15}$$

Next, we state the data generating assumptions needed for DIP to have target risk guarantees. Since a single source environment is considered in this subsection, without loss of generality, we assume that the first source environment is used.

Assumption 1 *Each data point in the source environment is generated i.i.d. according to distribution $\mathcal{P}^{(1)}$ specified by the following SCM*

$$\begin{bmatrix} X^{(1)} \\ Y^{(1)} \end{bmatrix} = \begin{bmatrix} \mathbf{B} & b \\ \omega^\top & 0 \end{bmatrix} \begin{bmatrix} X^{(1)} \\ Y^{(1)} \end{bmatrix} + \begin{bmatrix} a^{(1)}_X \\ a^{(1)}_Y \end{bmatrix} + \begin{bmatrix} \varepsilon^{(1)}_X \\ \varepsilon^{(1)}_Y \end{bmatrix},$$

each data point in the target environment is generated i.i.d. according to distribution $\widetilde{\mathcal{P}}$ specified with the same SCM except for the mean shift intervention term

$$\begin{bmatrix} \widetilde{X} \\ \widetilde{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{B} & b \\ \omega^\top & 0 \end{bmatrix} \begin{bmatrix} \widetilde{X} \\ \widetilde{Y} \end{bmatrix} + \begin{bmatrix} \widetilde{a}_X \\ \widetilde{a}_Y \end{bmatrix} + \begin{bmatrix} \widetilde{\varepsilon}_X \\ \widetilde{\varepsilon}_Y \end{bmatrix},$$

where $\mathbb{I}_d - \mathbf{B} \in \mathbb{R}^{d\times d}$ is invertible, the prediction problem is anticausal i.e. $\omega = 0$, $b \in \mathbb{R}^d$, the intervention terms $\begin{bmatrix} a^{(1)}_X \\ a^{(1)}_Y \end{bmatrix}$ and $\begin{bmatrix} \widetilde{a}_X \\ \widetilde{a}_Y \end{bmatrix}$ are vectors in $\mathbb{R}^{d+1}$, the noise terms $\begin{bmatrix} \varepsilon^{(1)}_X \\ \varepsilon^{(1)}_Y \end{bmatrix}$ and $\begin{bmatrix} \widetilde{\varepsilon}_X \\ \widetilde{\varepsilon}_Y \end{bmatrix}$ share the same distribution with

$$\mathbb{E}\left[\varepsilon^{(1)}_X\right] = \mathbb{E}\left[\widetilde{\varepsilon}_X\right] = 0, \quad \mathbb{E}\left[\varepsilon^{(1)}_X {\varepsilon^{(1)}_X}^\top\right] = \mathbb{E}\left[\widetilde{\varepsilon}_X \widetilde{\varepsilon}_X^\top\right] = \Sigma \in \mathbb{R}^{d\times d},$$
$$\mathbb{E}\left[\varepsilon^{(1)}_Y\right] = \mathbb{E}\left[\widetilde{\varepsilon}_Y\right] = 0, \quad \mathbb{E}\left[{\varepsilon^{(1)}_Y}^2\right] = \mathbb{E}\left[\widetilde{\varepsilon}_Y^2\right] = \sigma^2 \in \mathbb{R}.$$

Additionally, the noise terms on X and Y are uncorrelated $\mathbb{E}\left[\varepsilon^{(1)}_Y \cdot \varepsilon^{(1)}_X\right] = \mathbb{E}\left[\widetilde{\varepsilon}_Y \cdot \widetilde{\varepsilon}_X\right] = 0$.

Note that Assumption 1 does not require Σ to be diagonal, meaning that the noise terms of X can be correlated. Consequently, this assumption allows for unobserved confounders affecting the X variables to exist. Under the assumptions above and that there is no intervention on Y, we obtain the following theorem on the target risk of DIP.

Theorem 1 *Under the data generation Assumption 1 and the assumption of no intervention on Y i.e. $a_Y^{(1)} = \widetilde{a}_Y = 0$, the target population risks of OLSTar, OLSSrc$^{(1)}$ and DIP$^{(1)}$-mean satisfy*

$$\tilde{R}\left(f_{OLSTar}\right) = \frac{\sigma^2}{1+\sigma^2 b^\top \Sigma^{-1} b}, \tag{16}$$

$$\tilde{R}\left(f_{OLSSrc}^{(1)}\right) = \frac{\sigma^2}{1+\sigma^2 b^\top \Sigma^{-1} b} + \frac{\left(\sigma^2 b^\top \Sigma^{-1}\left(a_X^{(1)} - \widetilde{a}_X\right)\right)^2}{\left(1+\sigma^2 b^\top \Sigma^{-1} b\right)^2}, \tag{17}$$

$$\tilde{R}\left(f_{DIP}^{(1)}\right) = \tilde{R}\left(f_{DIPOracle}^{(1)}\right) = \frac{\sigma^2}{1+\sigma^2 b^\top \Sigma^{-\frac{1}{2}} G_{DIP}^{(1)} \Sigma^{-\frac{1}{2}} b}, \tag{18}$$

where $G_{DIP}^{(1)} = \Sigma^{1/2} Q_{DIP}^{(1)} \left(Q_{DIP}^{(1)\top} \Sigma Q_{DIP}^{(1)}\right)^{-1} Q_{DIP}^{(1)\top} \Sigma^{1/2}$ is a projection matrix with rank $d-1$; $Q_{DIP}^{(1)} \in \mathbb{R}^{d \times d-1}$ is a matrix with columns formed by the vectors that complete the vector $u^{(1)}$ to an orthonormal basis. Here $u^{(1)} = \frac{a_X^{(1)} - \widetilde{a}_X}{\left\|a_X^{(1)} - \widetilde{a}_X\right\|_2}$ when the source and target distribution is not identical i.e. $a_X^{(1)} \neq \widetilde{a}_X$. When $a_X^{(1)} = \widetilde{a}_X$, meaning that the source distribution equals the target distribution, then $u^{(1)} = 0$ and we have $\tilde{R}\left(f_{DIP}^{(1)}\right) = \tilde{R}\left(f_{OLSSrc}^{(1)}\right) = \tilde{R}\left(f_{OLSTar}\right)$.

The proof of Theorem 1 is provided in Appendix B.1. Comparing Equation (18) with Equation (16), the target population risk of DIP$^{(1)}$ is larger than that of OLSTar but it is lower than the target population risk of the Causal estimator (which equals to σ^2). The target population risk of OLSSrc$^{(1)}$ depends on the magnitude of the difference in intervention $\left\|a_X^{(1)} - \widetilde{a}_X\right\|_2$, while the target risk of DIP$^{(1)}$ is independent of that magnitude. Consequently when the difference in intervention becomes large, DIP$^{(1)}$ can outperform OLSSrc$^{(1)}$. Moreover, the first equality of Equation (18) shows that DIP$^{(1)}$ achieves the same target population risk as DIPOracle$^{(1)}$. DIPOracle$^{(1)}$ differs from OLSTar by only one linear constraint. The connection between DIP$^{(1)}$ and DIPOracle$^{(1)}$ intuitively explains why DIP$^{(1)}$ target risk should be close to that of OLSTar.

In fact, with additional assumptions on how the interventions $a_X^{(1)}$ and a_X are positioned and simplifications on the covariance matrix Σ, we obtain the following corollary which clearly highlights the difference between the target population risk of DIP and the oracle target population risk of OSLTar.

Corollary 2 *In addition to the assumptions in Theorem 1, suppose* $\Sigma = \frac{\sigma^2}{\rho}\mathbb{I}_d$ *with* $\rho > 0$, *then*

$$\tilde{R}\left(f_{OLSTar}\right) = \frac{\sigma^2}{1+\rho\left\|b\right\|_2^2}, \tag{19}$$

$$\tilde{R}\left(f^{(1)}_{OLSSrc}\right) = \frac{\sigma^2}{1+\rho\left\|b\right\|_2^2} + \frac{\left(u^{(1)\top}b\right)^2\left\|a_X^{(1)} - \widetilde{a}_X\right\|_2^2}{\left(1+\rho\left\|b\right\|_2^2\right)^2}, \tag{20}$$

$$\tilde{R}\left(f^{(1)}_{DIP}\right) = \frac{\sigma^2}{1+\rho\left\|b\right\|_2^2 - \rho\left(u^{(1)\top}b\right)^2}, \tag{21}$$

where $u^{(1)} = \frac{a_X^{(1)} - \widetilde{a}_X}{\left\|a_X^{(1)} - \widetilde{a}_X\right\|_2}$ *if* $a_X^{(1)} \neq \widetilde{a}_X$ *and* $u^{(1)} = 0$ *otherwise.*

Additionally, if $a_X^{(1)} - \widetilde{a}_X$ *is generated randomly from the Gaussian distribution* $\mathcal{N}(0, \tau\mathbb{I}_d^2)$, *then for a constant* t *satisfying* $0 < t \leq \frac{d}{2}$, *with probability at least* $1 - \exp(-t/16) - 2\exp(-t/4)$, *we have*

$$\tilde{R}\left(f^{(1)}_{OLSSrc}\right) \leq \frac{\sigma^2}{1+\rho\left\|b\right\|_2^2} + \frac{\tau^2 t\left\|b\right\|_2^2}{2\left(1+\rho\left\|b\right\|_2^2\right)^2},$$

$$\tilde{R}\left(f^{(1)}_{DIP}\right) \leq \frac{\sigma^2}{1+\rho\left(1-\frac{t}{d}\right)\left\|b\right\|_2^2}. \tag{22}$$

The proof of Corollary 2 is provided in Appendix B.2. Under the conditions of Corollary 2, Equation (21) shows that the gap between the target population risks of DIP$^{(1)}$ and OLSTar is smaller when the direction of the difference in intervention $u^{(1)}$ becomes less aligned with the vector b. When the intervention is generated randomly from a Gaussian distribution, the bound (22) shows that the gap between the target population risks of DIP$^{(1)}$ and OLSTar is small with high probability. The gap only comes from an order $\frac{1}{d}$ term in the denominator of the target risk. So when the dimension d is large, this difference in target population risk between DIP$^{(1)}$ and OLSTar becomes negligible. As for OLSSrc, it is now clear from Corollary 2 that the target risk of OLSSrc$^{(1)}$ depends on the magnitude of the difference in intervention τ. For brevity, we only provided the target risk upper bound of OLSSrc$^{(1)}$. The lower bound should hold similarly up to constant factors because the risks are derived with equality in Equation (20)-(21) and tight Gaussian concentration bounds are used to obtain the upper bounds. As the magnitude of the difference in intervention τ increases, OLSSrc$^{(1)}$ will eventually have a larger target risk than DIP$^{(1)}$.

The intuition behind the success of DIP$^{(1)}$ over OLSSrc$^{(1)}$ is that the DIP matching penalty in Equation (10) allows the DIP$^{(1)}$ estimator to equalize the source and target covariate intervention by projecting it to a common space. As a result, given the DIP matching penalty, computing least squares on the source data is the same as computing least squares on the target data. This observation also explains why the target risk of DIP$^{(1)}$ matches that of DIPOracle$^{(1)}$. Having this intuition in mind, it is not hard to

imagine extending the guarantees of DIP$^{(1)}$ to the generic form of DIP in Equation (9) under appropriate assumptions. Specifically, under Assumption 1, as long as the function class $\mathcal{V}$ is chosen such that the DIP matching penalty $\mathcal{D}(v(X), v(\tilde{X}))$ ensures the conditionals $v(X) \mid Y$ and $v(\tilde{X}) \mid Y$ to have the same distribution, then no matter how $\mathcal{U}$ is chosen we always match the target risks of DIP and DIPOracle. Moreover, we know that target risk of DIPOracle is close to the oracle target risk over the function class $\{u \circ v \mid u \in \mathcal{U}, v \in \mathcal{V}\}$. So the target risk of DIP in this extended setting is also close to the oracle one.

4.1.2 Benefit of more source environments for DIP

We show how to make use of more independent source environments to improve the performance of DIP. Here we consider M ($M \geq 2$) source environments, where for each $m \in \{1, \ldots, M\}$ the m-th source environment is generated independently according to the source environment in Assumption 1 except for the unknown interventions $a_X^{(m)}$ and we still have $a_Y^{(m)} = 0$. First, we state a corollary revealing that it is possible to pick the best source environment for DIP based on the best source risk. Second, we discuss the reason behind the performance improvement after picking the best source environment.

First, based on the proof of Theorem 1, we derive the following corollary on the source population risk.

Corollary 3 *Under the data generation Assumption 1 and the assumption of no intervention on Y i.e. $a_Y^{(1)} = \widetilde{a}_Y = 0$, the source population risk of DIP$^{(1)}$-mean as defined in Equation (2) satisfies*

$$R^{(1)}\left(f_{DIP}^{(1)}\right) = \tilde{R}\left(f_{DIP}^{(1)}\right). \tag{23}$$

The proof of Corollary 3 is provided in Appendix B.3. According to Corollary 3, one can read off the target population risk directly from the source population risk. Since the choice of source environment 1 is arbitrary, the same result holds for any of the M source environments. We have $R^{(m)}\left(f_{\mathrm{DIP}}^{(m)}\right) = \tilde{R}\left(f_{\mathrm{DIP}}^{(m)}\right)$ for any $m \in \{1, \ldots, M\}$. Consequently, having more source environments allows one to apply DIP between each source environment and the target environment one by one, and then pick the estimator $f_{\mathrm{DIP}}^{(m)}$ with the lowest source population risk to reduce the target population risk.

Second, we reason about the type of source environments that reduces the target population risk the most. According to Equation (16) and (18) in Theorem 1, the projection matrix $G_{\mathrm{DIP}}^{(1)}$ is the term that makes the target population risks of DIP$^{(1)}$ and DIP$^{(m)}$ different ($m \neq 1$). If the vector $\Sigma^{-1/2}b$ is in the span of the projection matrix $G_{\mathrm{DIP}}^{(1)}$, then DIP$^{(1)}$ achieves the oracle target population risk. Otherwise, the target population risk of DIP$^{(1)}$ depends on the norm of the component of $\Sigma^{-1/2}b$ outside the span of the projection matrix $G_{\mathrm{DIP}}^{(1)}$.

The above intuition becomes clearer under the additional assumptions of Corollary 2. There we obtain a simple form for the projection matrix $G_{\mathrm{DIP}}^{(1)} = \mathbb{I}_d - u^{(1)}{u^{(1)}}^\top$, where $u^{(1)} = \frac{a_X^{(1)} - \widetilde{a}_X}{\left\|a_X^{(1)} - \widetilde{a}_X\right\|_2}$ assuming $a_X^{(1)} \neq \widetilde{a}_X$. Under the same assumptions, the denominator in

the DIP target risk in Equation (18) becomes $1+\|b\|_2^2-\left(u^{(1)\top}b\right)^2$. If $a_X^{(1)}-\widetilde{a}_X$ is orthogonal to b then DIP$^{(1)}$ achieves the oracle target population risk. Otherwise, DIP$^{(1)}$ has larger target population risk than OLSTar. With more source environments, it is more likely to find a source environment m such that $a_X^{(m)}-\widetilde{a}_X$ is closer to be orthogonal to b.

Based on the above intuition, we propose the following DIP variant that provides a weighting choice to take advantage of multiple source environments.

- **DIPweigh-mean:** the population DIP estimator that makes use of multiple source environments based on the source risks.

$$f_{\text{DIPweigh}}(x) := \frac{1}{\sum_{m=1}^M e^{-\eta\cdot s_m}}\sum_{m=1}^M e^{-\eta\cdot s_m}\left(x^\top\beta_{\text{DIP}}^{(m)}+\beta_{\text{DIP},0}^{(m)}\right)$$
$$s_m := R^{(m)}\left(f_{\text{DIP}}^{(m)}\right). \tag{24}$$

  $\eta>0$ is a constant. Choosing η to be ∞ is equivalent to choosing the source estimator with the lowest source risk. DIPweigh weights all the source predictions based on the source risk of each source environment.

Here we introduce the weighted version of DIP rather than directly selecting the source estimator with the lowest source risk. The main intuition is that in finite sample, averaging the predictions from several source environments with low source risks can take advantage of a larger sample size. For example, in the presence of a less related source environment with very large sample size, one might still include it if the reduction in variance outweighs the increase in bias.

4.1.3 Failure scenarios of DIP

The target population risk guarantees of DIP in Theorem 1 rely on the anticausal data generation Assumption 1 and the assumption of no intervention on Y. We already showed in Example 1 that DIP cannot outperform OLSSrc in the causal prediction setting. Even in the anticausal prediction setting, we identify two scenarios where DIP$^{(1)}$-mean might have large target risk: when the DIP matching penalty is not well-suited for the underlying type of the intervention that generates the data and when there is intervention on Y.

Consider the following data generation model where the source environment is generated from the following SCM with interventions on the variance

$$\begin{bmatrix}X^{(1)}\\Y^{(1)}\end{bmatrix}=\begin{bmatrix}\mathbf{B}&b\\\omega^\top&0\end{bmatrix}\begin{bmatrix}X^{(1)}\\Y^{(1)}\end{bmatrix}+\begin{bmatrix}a_X^{(1)}\odot\varepsilon_X^{(1)}\\\varepsilon_Y^{(1)}\end{bmatrix},$$

and the target environment is generated from the same SCM except that the intervention on X takes the form $\widetilde{a}_X\odot\widetilde{\varepsilon}_X$. Here $\odot$ denotes the d-dimensional element-wise multiplication. This intervention is called variance shift noise intervention. The intervention term $a_X^{(1)}$ and $\widetilde{a}_X$ are fixed vectors in $\mathbb{R}^d$, and the noise terms are kept the same as in Assumption 1. Under the data generation model in this subsection, the matching penalty in DIP$^{(1)}$-mean (10) is always satisfied because both the left and right hand sides are zero. Consequently, DIP$^{(1)}$-mean becomes the same estimator as OLSSrc. Matching the mean between source and

target distribution in this case is no longer a good idea, because the intervention is on the variance rather than the mean.

To adapt to the new type of intervention, one can consider *DIP-std* which puts the matching penalty on the standard deviations, *DIP-std+* which puts the matching penalty on the means, standard deviations and 25% quantiles, or *DIP-MMD* which uses more generic distributional matching via MMD. The exact formulations consist of replacing the DIP mean matching penalty in Equation (10) with the appropriate matching penalties and they are formally introduced in Appendix A.1.

If the assumptions are set up such that the intervention type agrees with the matching penalty in DIP, analyzing the theoretical performance of DIP-std, DIP-std+ or DIP-MMD under linear SCMs works similarly as analyzing that of DIP-mean. Hence we leave out the theoretical guarantees of DIP-std, DIP-std+ or DIP-MMD for brevity. The empirical performance of DIP-std+ and DIP-MMD is shown through simulations and real data experiments in Section 5.

The second failure scenario of DIP is when there is intervention on Y. This failure scenario of DIP is already noticeable from Example 3 in Section 3.5.3. Here we provide a corollary that quantifies the additional target population risk made by DIP if there is intervention on Y.

Corollary 4 *Under the data generation Assumption 1, the target population risk of OLSTar and DIP$^{(1)}$-mean satisfy*

$$\begin{aligned}\tilde{R}\left(f_{OLSTar}\right) &= \frac{\sigma^2}{1+\sigma^2 b^\top \Sigma^{-1} b} \\ \tilde{R}\left(f_{DIP}^{(1)}\right) &= \frac{\sigma^2}{1+\sigma^2 b^\top \Sigma^{-\frac{1}{2}} G_2 \Sigma^{-\frac{1}{2}} b} + \left(a_Y^{(1)} - \widetilde{a}_Y\right)^2, \end{aligned} \tag{25}$$

where $G_2 = \Sigma^{1/2} Q_2 \left(Q_2^\top \Sigma Q_2\right)^{-1} Q_2^\top \Sigma^{1/2}$ *is a projection matrix with rank* $d-1$. $Q_2 \in \mathbb{R}^{d\times d-1}$ *is a matrix with columns formed by the vectors that complete the vector* $\frac{a_X^{(1)} + a_Y^{(1)} b - \widetilde{a}_X - \widetilde{a}_Y b}{\left\| a_X^{(1)} + a_Y^{(1)} b - \widetilde{a}_X - \widetilde{a}_Y b \right\|_2}$ *to an orthonormal basis.*

The proof of Corollary 4 is provided in Appendix B.3. According to Equation (25) in Corollary 4, the target population risk has an extra term that depends on the difference between target and source Y intervention. The target population risk of DIP is increases as the difference between target and source Y intervention increases. This corollary highlights the result that DIP has a large target risk when the intervention on Y is large.

4.2 Anticausal domain adaptation with intervention on Y

In this subsection, we study the anticausal domain adaptation where there is intervention on the label Y. As illustrated in Example 3 in Section 3.5.3 and in the last section, the intervention on Y can cause DIP to have a large target population risk. In fact, allowing arbitrary intervention on both X and Y can also lead to unidentifiable cases where two data generation models result in the same target covariate distribution in the anticausal domain adaptation. If it were the case, any DA estimator based solely on the target covariate

distribution will have large target risk. The following example illustrates one concrete unidentifiable case.

A simple unidentifiable example: The target data is generated form the following SCM

$$\widetilde{Y} = \widetilde{\varepsilon}_Y + \widetilde{a}_Y$$
$$\widetilde{X} = 2\widetilde{Y} + \widetilde{\varepsilon}_X + \widetilde{a}_X,$$

where $\widetilde{X}$ is 1-dimensional random variable, $\widetilde{\varepsilon}_Y \sim \mathcal{N}(0,1)$ and $\widetilde{\varepsilon}_X \sim \mathcal{N}(0,0.1)$. Assume we observe the target covariate distribution $\widetilde{X} \sim \mathcal{N}(3,1.1)$. Without observing the target label $\widetilde{Y}$ distribution, it is impossible to tell whether the intervention is $(\widetilde{a}_X, \widetilde{a}_Y) = (0,1.5)$, $(1,1)$ or $(3,0)$. This is because all three interventions result in the same target covariate distribution. However, the conditional distribution $\widetilde{Y} \mid \widetilde{X}$ are different. Consequently, any estimator of the conditional mean $\mathbb{E}\left[\widetilde{Y} \mid \widetilde{X}\right]$ without access to the target label distribution can not get it correct.

To make the anticausal domain adaptation problem with intervention on Y tractable, additional assumptions on the structure of the interventions is needed. Here we adopt one type of assumptions introduced by Gong et al. (2016); Heinze-Deml and Meinshausen (2021). This line of work assumes the existence of conditionally invariant components (CICs) across all source and target environments. That is, there exists an unknown transformation $\mathcal{T}$ of the covariates such that the conditional distribution $\mathcal{T}(X) \mid Y$ is invariant across source and target environments.

In the following, we first prove target population risk guarantees for CIP and CIRM under the conditionally invariant components assumption. We show that the CIP and CIRM target risks are much less dependent on the Y intervention than the DIP target risk. Finally we discuss extensions and variants of CIP and CIRM.

4.2.1 Target risk guarantees of CIP and CIRM with multiple source environments and conditionally invariant components

We start by stating the data generation assumptions for CIP and CIRM to have target risk guarantees, namely a SCM and the existence of conditionally invariant components.

Assumption 2 *There are M ($M \geq 2$) source environments. Each data point in the m-th source environment is generated i.i.d. according to distribution $\mathcal{P}^{(m)}$ specified by the following SCM*

$$\begin{bmatrix} X^{(m)} \\ Y^{(m)} \end{bmatrix} = \begin{bmatrix} \mathbf{B} & b \\ \omega^\top & 0 \end{bmatrix} \begin{bmatrix} X^{(m)} \\ Y^{(m)} \end{bmatrix} + \begin{bmatrix} a_X^{(m)} \\ a_Y^{(m)} \end{bmatrix} + \begin{bmatrix} \varepsilon_X^{(m)} \\ \varepsilon_Y^{(m)} \end{bmatrix}.$$

Each data point in the target environment is generated i.i.d. according to distribution $\widetilde{\mathcal{P}}$ specified with the same SCM except for the mean shift intervention term

$$\begin{bmatrix} \widetilde{X} \\ \widetilde{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{B} & b \\ \omega^\top & 0 \end{bmatrix} \begin{bmatrix} \widetilde{X} \\ \widetilde{Y} \end{bmatrix} + \begin{bmatrix} \widetilde{a}_X \\ \widetilde{a}_Y \end{bmatrix} + \begin{bmatrix} \widetilde{\varepsilon}_X \\ \widetilde{\varepsilon}_Y \end{bmatrix},$$

where $\mathbb{I}_d - \mathbf{B}$ *is invertible, the prediction problem is anticausal* $\omega = 0$*, the intervention term* $\begin{bmatrix} a_X^{(1)} \\ a_Y^{(1)} \end{bmatrix}$ *and* $\begin{bmatrix} \widetilde{a}_X \\ \widetilde{a}_Y \end{bmatrix}$ *are vectors in* $\mathbb{R}^{d+1}$*, the noise term* $\begin{bmatrix} \varepsilon_X^{(m)} \\ \varepsilon_Y^{(m)} \end{bmatrix}$ *and* $\begin{bmatrix} \widetilde{\varepsilon}_X \\ \widetilde{\varepsilon}_Y \end{bmatrix}$ *share the same distribution with*

$$\mathbb{E}\left[\varepsilon_X^{(m)}\right] = \mathbb{E}\left[\widetilde{\varepsilon}_X\right] = 0, \quad \mathbb{E}\left[\varepsilon_X^{(m)} \varepsilon_X^{(m)\top}\right] = \mathbb{E}\left[\widetilde{\varepsilon}_X \widetilde{\varepsilon}_X^{\top}\right] = \Sigma,$$
$$\mathbb{E}\left[\varepsilon_Y^{(m)}\right] = \mathbb{E}\left[\widetilde{\varepsilon}_Y\right] = 0, \quad \mathbb{E}\left[\varepsilon_Y^{(m)2}\right] = \mathbb{E}\left[\widetilde{\varepsilon}_Y^2\right] = \sigma^2.$$

In addition, the noise terms on X *and* Y *are uncorrelated* $\mathbb{E}\left[\varepsilon_Y^{(m)} \cdot \varepsilon_X^{(m)}\right] = \mathbb{E}\left[\widetilde{\varepsilon}_Y \cdot \widetilde{\varepsilon}_X\right] = 0$*. The interventions do not span the whole space (to ensure the existence of conditionally invariant components)*

$$dim\left(span\left(a_X^{(2)} - a_X^{(1)}, \ldots, a_X^{(M)} - a_X^{(1)}\right)\right) = p \leq d - 1, \tag{26}$$

and the target X *intervention is in the span of source* X *interventions*

$$\widetilde{a}_X - a_X^{(1)} \in span\left(a_X^{(2)} - a_X^{(1)}, \ldots, a_X^{(M)} - a_X^{(1)}\right). \tag{27}$$

Compared to Assumption 1, Assumption 2 involves multiple source environments instead of one. Also, the way the source and target environments are generated up to Equation (26) is the same as in Assumption 1. Equation (26) assumes that the X interventions do not span the whole covariate space $\mathbb{R}^d$. It ensures the existence of a invariant linear projection of the covariates given the label Y. More precisely, this assumption allows us to find a vector v which is orthogonal to the span $\left((\mathbb{I}_d - \mathbf{B})^{-1}\left(a_X^{(2)} - a_X^{(1)}\right), \ldots, (\mathbb{I}_d - \mathbf{B})^{-1}\left(a_X^{(M)} - a_X^{(1)}\right)\right)$ and the projection of the covariates $X^{\top}v$ has the same intervention term across source environments. This particular linear projection of the covariates becomes one conditionally invariant component. To make sure the same conditionally invariant component is also valid in the target environment, Equation (27) requires that the target X intervention is in the span of source X interventions.

Given the data generation assumption above, we have the following theorem on the target risk of CIP and CIRM.

Theorem 5 *Under the data generation Assumption 2, the population target risks of CIP-mean and CIRM*$^{(m)}$*-mean for any* $m \in \{1, \ldots, M\}$ *satisfy*

$$\tilde{R}\left(f_{OLSTar}\right) = \frac{\sigma^2}{1 + \sigma^2 b^{\top}\Sigma^{-1}b}, \tag{28}$$

$$\tilde{R}\left(f_{CIP}\right) = \frac{\sigma^2 + \Delta_Y}{1 + \left(\sigma^2 + \Delta_Y\right) b^{\top}\Sigma^{-\frac{1}{2}} G_{CIP} \Sigma^{-\frac{1}{2}} b} + \frac{\left(\widetilde{a}_Y - \bar{a}_Y\right)^2 - \Delta_Y}{\left(1 + \left(\sigma^2 + \Delta_Y\right) b^{\top}\Sigma^{-\frac{1}{2}} G_{CIP} \Sigma^{-\frac{1}{2}} b\right)^2}, \tag{29}$$

$$\tilde{R}\left(f_{CIRM}^{(m)}\right) = \frac{\sigma^2}{1 + \sigma^2 b^{\top}\Sigma^{-\frac{1}{2}} G_{DIP}^{(m)} \Sigma^{-\frac{1}{2}} b} + \frac{\left(a_Y^{(m)} - \widetilde{a}_Y\right)^2}{\left(1 + \sigma^2 b^{\top}\Sigma^{-\frac{1}{2}} G_{DIP}^{(m)} \Sigma^{-\frac{1}{2}} b\right)^2}, \tag{30}$$

where $\bar{a}_Y = \frac{1}{M}\sum_{j=1}^{M} a_Y^{(j)}$, $\Delta_Y = \frac{1}{M}\sum_{j=1}^{M}\left(a_Y^{(j)} - \bar{a}_Y\right)^2$. $G_{CIP} = \Sigma^{1/2} Q_{CIP}\left(Q_{CIP}^\top \Sigma Q_{CIP}\right)^{-1} Q_{CIP}^\top \Sigma^{1/2}$ *is a projection matrix with rank* $d-p$ *where* $Q_{CIP} \in \mathbb{R}^{d\times(d-p)}$ *is a matrix with columns formed by an orthonormal basis of the orthogonal complement of span*$(a_X^{(2)} - a_X^{(1)}, \ldots, a_X^{(M)} - a_X^{(1)})$. $G_{DIP}^{(m)}$ *is a projection matrix defined in the same way as* $G_{DIP}^{(1)}$ *in Theorem 1.*

The proof of Theorem 5 is provided in Appendix C.1. Comparing the target population risk of CIRM$^{(m)}$ in Equation (30) with that of DIP$^{(m)}$ in Equation (25), CIRM$^{(m)}$ reduces the dependency on the difference in Y intervention $a_Y^{(m)} - \widetilde{a}_Y$. This is because we always have $1 + \sigma^2 b^\top \Sigma^{-\frac{1}{2}} G_{\text{DIP}}^{(m)} \Sigma^{-\frac{1}{2}} b > 1$. The target population risk of CIRM$^{(m)}$ when there is intervention on Y in Equation (30) has an additional term depending on $a_Y^{(m)} - \widetilde{a}_Y$ when compared to that of DIP$^{(m)}$ when there is no intervention on Y in Equation (18). The additional term becomes close to zero when $\sigma^2 b^\top \Sigma^{-\frac{1}{2}} G_{\text{DIP}}^{(m)} \Sigma^{-\frac{1}{2}} b$ is large.

In fact, without additional assumptions to Assumption 2, the dependence on $a_Y^{(m)} - \widetilde{a}_Y$ or similar terms is unavoidable for any DA estimator. Because the Assumption 2 does not prevent b from being in the span $\left(a_X^{(2)} - a_X^{(1)}, \ldots, a_X^{(M)} - a_X^{(1)}\right)$, it can still result in a slightly unidentifiable example similar to the one presented at the beginning of Section 4.2. See Appendix D for ways to get rid of the $a_Y^{(m)} - \widetilde{a}_Y$ dependence at the cost of small additional target risk.

In the same spirit of Corollary 2, we present a corollary that puts additional assumptions on the positions of the interventions $a_X^{(m)}$ and $\widetilde{a}_X$ to make the results in Theorem 5 easier to understand.

Corollary 6 *In addition to Assumption 2, suppose* $\Sigma = \frac{\sigma^2}{\rho}\mathbb{I}_d$ *with* $\rho > 0$, *then*

$$\tilde{R}\left(f_{OLSTar}\right) = \frac{\sigma^2}{1 + \rho\left\|b\right\|_2^2}, \tag{31}$$

$$\tilde{R}\left(f_{CIP}\right) = \frac{\sigma^2}{1 + \rho\left\|b\right\|_2^2 - \rho\left(P_{CIP}^\top b\right)^2} + \frac{\left(\widetilde{a}_Y - \bar{a}_Y\right)^2 - \Delta_Y}{\left[1 + \rho\left\|b\right\|_2^2 - \rho\left(P_{CIP}^\top b\right)^2\right]^2}, \tag{32}$$

$$\tilde{R}\left(f_{CIRM(m)}\right) = \frac{\sigma^2}{1 + \rho\left\|b\right\|_2^2 - \rho\left(u^{(m)\top} b\right)^2} + \frac{\left(\widetilde{a}_Y - a_Y^{(m)}\right)^2}{\left[1 + \rho\left\|b\right\|_2^2 - \rho\left(u^{(m)\top} b\right)^2\right]^2}, \tag{33}$$

where $\bar{a}_Y = \frac{1}{M}\sum_{m=1}^{M} a_Y^{(m)}$, $\Delta_Y = \left(\widetilde{a}_Y - \bar{a}_Y\right)^2 - \frac{1}{M}\sum_{m=1}^{M}\left(a_Y^{(m)} - \bar{a}_Y\right)^2$, $P_{CIP} = Q_{CIP}^\perp \in \mathbb{R}^{d\times p}$ *is the matrix with columns formed by the orthonormal basis of span* $\left(a_X^{(2)} - a_X^{(1)}, \ldots, a_X^{(M)} - a_X^{(1)}\right)$ *and* $u^{(m)} = \frac{a_X^{(m)} - \widetilde{a}_X}{\left\|a_X^{(m)} - \widetilde{a}_X\right\|_2}$. *Additionally, if* $a_X^{(2)} - a_X^{(1)}, \ldots, a_X^{(M)} - a_X^{(1)}, \xi$ *are generated independently from Gaussian distribution* $\mathcal{N}\left(0, \tau^2\mathbb{I}_d\right)$ *and* $\widetilde{a}_X - a_X^{(1)} = P_{CIP}\xi$, $d \geq 6M$, *a constant* $0 < t \leq d/2$, *then with probability at least* $(1 - 2\exp(-d/32) - 2\exp(-M/32)) \cdot (1 - \exp(-t/16) - 2\exp(-t/4))$, *we have dim* $\left(\text{span}\left(a_X^{(2)} - a_X^{(1)}, \ldots, a_X^{(M)} - a_X^{(1)}\right)\right) = M - 1$

and additionally

$$\tilde{R}\left(f_{CIP}\right) \leq \frac{\sigma^2}{1+\rho\left(1-\frac{3(M-1)}{d}\right)\|b\|_2^2} + \frac{\left(\widetilde{a}_Y-\bar{a}_Y\right)^2-\Delta_Y}{\left[1+\rho\left(1-\frac{3(M-1)}{d}\right)\|b\|_2^2\right]}, \tag{34}$$

$$\tilde{R}\left(f_{CIP}\right) \geq \frac{\sigma^2}{1+\rho\left(1-\frac{(M-1)}{3d}\right)\|b\|_2^2} + \frac{\left(\widetilde{a}_Y-\bar{a}_Y\right)^2-\Delta_Y}{\left[1+\rho\left(1-\frac{(M-1)}{3d}\right)\|b\|_2^2\right]}, \tag{35}$$

$$\tilde{R}\left(f_{CIRM(m)}\right) \leq \frac{\sigma^2}{1+\rho\left(1-\frac{t}{d}\right)\|b\|_2^2} + \frac{\left(\widetilde{a}_Y-a_Y^{(m)}\right)^2}{\left[1+\rho\left(1-\frac{t}{d}\right)\|b\|_2^2\right]}. \tag{36}$$

The proof of Corollary 6 is provided in Appendix C.2. Under the assumptions of Corollary 6, it is clear that for M and d sufficiently large, with high probability, CIRM$^{(m)}$ has smaller target population risk than CIP when all Y interventions are equal $a_Y^{(1)} = \ldots = a_Y^{(M)} = \widetilde{a}_Y$. This is because the last term in Equation (32) become zero can be ignored and $\frac{M-1}{3d} \geq \frac{t}{d}$ when $M \geq 3t+1$. Intuitively, CIP only uses the conditionally invariant components (roughly $d-p$ coordinates) to build the estimator, while CIRM$^{(m)}$ takes advantage of the other coordinates of X (roughly $d-1$ coordinates).

The intuition behind CIRM becomes apparent after the theorem statement. CIRM$^{(m)}$ is indeed a combination of DIP$^{(m)}$ and CIP: it applies CIP in the first step to obtain the conditionally invariant component which serves as a proxy of the label Y to correct for the intervention on Y, then it applies DIP$^{(m)}$ with the corrected target covariate distribution to improve the target prediction performance. The conditionally invariant components identified by CIP are useful to constitute a proxy of Y but they alone do not predict Y very well especially when there are only a few of them. DIP has good target risk guarantees when there is no Y intervention. When there is Y intervention, CIRM uses CIP to correct for the Y distribution perturbation and to create a residual which does not suffer from intervention, then apply DIP on the residual.

Since CIRM can be seen as applying DIP on the label-distribution-corrected source and target data, it is natural to introduce a weighted version of CIRM called CIRMweigh similar to DIPweigh (24). A precise formulation of CIRMweigh can be found in Appendix A.1.

Under the assumptions of Corollary 6, we summarize the target risk upper bounds of all DA methods in this paper in Table 2. For brevity, we only present the target population risk upper bounds under high probability when the interventions are generated i.i.d. Gaussian $\mathcal{N}\left(0, \tau^2 \mathbb{I}_d\right)$. The upper bounds are tight up to constant factors in the denominators because we first proved the exact target risks with equality and then applied Gaussian concentration to obtain the upper bounds. Consequently, the upper bounds serve the comparison purpose.

4.2.2 On relaxing the assumptions needed for CIP and CIRM

We discuss two different relaxations of Assumption 2.

First, the two assumptions (26) and (27) in Assumption 2 can be replaced by a single assumption, namely there exists a non-zero vector $v \in \mathbb{R}^d$ such that $v^\top X^{(1)}, \ldots, v^\top X^{(M)}$ and $v^\top \widetilde{X}$ all have the same distribution. This is also equivalent to saying that the projection of the covariates on the direction v is a conditionally invariant component. Stating the CIC

	Estimator	**Target population risk upper bound** interventions under general position
No intervention on Y (Corollary 2)	OLSTar (oracle)	$\frac{\sigma^2}{1+\rho\left\|b\right\|_2^2}$
	$\text{OLSSrc}^{(1)}$	$\frac{\sigma^2}{1+\rho\left\|b\right\|_2^2}+\frac{c\cdot\tau^2\left\|b\right\|_2^2}{\left[1+\rho\left\|b\right\|_2^2\right]^2}$
	$\text{DIP}^{(1)}$	$\frac{\sigma^2}{1+\rho(1-\frac{c}{d})\left\|b\right\|_2^2}$
Intervention on Y with CICs M sources (Corollary 6)	$\text{DIP}^{(m)}$	$\frac{\sigma^2}{1+\rho(1-\frac{c}{d})\left\|b\right\|_2^2}+\left(\widetilde{a}_Y-a_Y^{(m)}\right)^2$
	CIP	$\frac{\sigma^2}{1+\rho(1-\frac{c(M-1)}{d})\left\|b\right\|_2^2}+\frac{(\widetilde{a}_Y-\bar{a}_Y)^2-\Delta_Y}{\left[1+\rho(1-\frac{c(M-1)}{d})\left\|b\right\|_2^2\right]^2}$
	$\text{CIRM}^{(m)}$	$\frac{\sigma^2}{1+\rho(1-\frac{c}{d})\left\|b\right\|_2^2}+\frac{\left(\widetilde{a}_Y-a_Y^{(m)}\right)^2}{\left[1+\rho(1-\frac{c}{d})\left\|b\right\|_2^2\right]^2}$

Table 2: Summary of target population risk for DA estimators for anticausal domain adaptation under the assumptions of Corollary 2 and Corollary 6. Here $c>0$ is a generic constant to improve readability, which can be different for different methods. For simplicity, we only present the target population risk upper bounds holding with high probability when the interventions are generated i.i.d. Gaussian $\mathcal{N}(0,\tau^2\mathbb{I}_d)$ and when both M and d are large. Under the assumptions of Corollary 2, the target risk of $\text{OLSSrc}^{(1)}$ depends on the magnitude of the difference in intervention τ, while the other DA methods in the table do not: $\text{DIP}^{(1)}$ outperforms $\text{OLSSrc}^{(1)}$. Under the assumptions of Corollary 6, when intervention on Y becomes large, $\text{DIP}^{(m)}$ is worse than CIP or $\text{CIRM}^{(m)}$. $\text{CIRM}^{(m)}$ slightly outperforms CIP.

assumption as we did in Assumption 2 has the advantage of separating the assumptions on the source interventions from those on the target intervention.

When the dimension d is larger than the number of source environments M, the assumption (26) on the dimension of the span of the source X interventions is always satisfied. In the case of $M<d$, the more source environments there are, the more likely that the target X intervention is in the span of the source X interventions.

Second, the mean shift noise intervention in Assumption 2 can be relaxed to other types of noise interventions. For example, the noise intervention can be the standard deviation shift as we did for DIP in Section 4.1.3. We may consider CIP-std which puts the conditional invariance penalty on the standard deviations, CIP-std+ which puts the conditional invariance penalty on the means, standard deviations and 25% quantiles or CIP-MMD which uses generic distributional matching to make sure the conditional distribution of $v^\top X$ is invariant across source environments. CIRM can be extended similarly. We do not provide

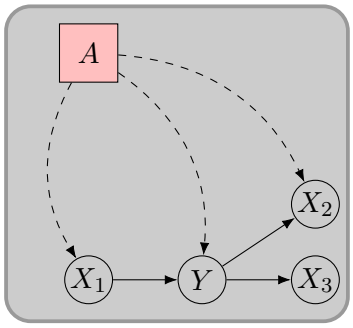
theoretical guarantees of these extensions and we only show their empirical performance through simulation and data experiments in Section 5.

4.3 Mixed-causal-anticausal domain adaptation

In this subsection, we consider the mixed-causal-anticausal DA setting where both causal and anticausal variables are present. In terms of the SCMs, neither the vector b nor ω in Equation (3) is assumed to be zero. In terms of the graph associated with the SCM, some variables are descendants of the label Y and some variables are ancestors of Y. As we have seen in Example 1 and Example 2 in Section 3.5, DA methods such as DIP perform worse than Causal or OLSSrc in the causal setting, while DIP can have smaller target population risk than Causal or OLSSrc in the anticausal setting. When both causal and anticausal variables are present, it is no longer clear whether more sophisticated DA methods would be better than Causal or OLSSrc. To gain some intuition, we first show a simple mixed-causal-anticausal DA example to illustrate how the standard DIP, CIP and CIRM can have worse target risk than Causal. Then we introduce new algorithms to tackle the mixed-causal-anticausal DA problem.

4.3.1 A mixed-causal-anticausal DA example to illustrate the failure of the standard DIP, CIP and CIRM

Here we provide a simple mixed-causal-anticausal DA problem to illustrate the difficulty of the mixed-causal-anticausal setting. Example 4 is a DA problem with a large number of source environments ($M \geq 4$) and one target environment. The data in the source and target environment are generated independently according to the following SCMs, with the causal diagram on the left and the structural equations on the right of Figure 5.



$$
\begin{aligned}
X_1^{(m)} &= \varepsilon_{X_1}^{(m)} + a_{X_1}^{(m)} & \widetilde{X}_1 &= \widetilde{\varepsilon}_{X_1} + 0 \\
X_2^{(m)} &= Y^{(m)} + \varepsilon_{X_2}^{(m)} + a_{X_2}^{(m)} & \widetilde{X}_2 &= \widetilde{Y} + \widetilde{\varepsilon}_{X_2} + 0 \\
X_3^{(m)} &= Y^{(m)} + \varepsilon_{X_3}^{(m)} & \widetilde{X}_3 &= \widetilde{Y} + \widetilde{\varepsilon}_{X_3} \\
Y^{(m)} &= X_1^{(m)} + 0 + a_Y^{(m)} & \widetilde{Y} &= \widetilde{X}_1 + 0 + 0
\end{aligned}
$$

Figure 5: The causal diagram and structural equations for the m-th source ($m \in \{1, \ldots, M\}$) and target environments in Example 4

Here the noise variables follow independent Gaussian distributions with $\varepsilon_{X_i}^{(m)}, \widetilde{\varepsilon}_{X_i} \sim \mathcal{N}(0, 0.1)$, $i \in \{1, \ldots, 3\}$. The noise variable of Y is set to zero. The type of intervention is mean shift noise intervention. The variable X_3 is never intervened on, so X_3 is conditionally invariant. For the first source environment, the intervention $a^{(1)} = \begin{bmatrix} 1 & 1 & 0 & 1 \end{bmatrix}^\top$. For $m \geq 2$, $a^{(m)} = \begin{bmatrix} a_{X_1}^{(m)} & a_{X_2}^{(m)} & 0 & a_Y^{(m)} \end{bmatrix}^\top$ where each coordinate is generated i.i.d. from $\mathcal{N}(0, 10)$ and we ensure that the $M - 1$ interventions span a subspace of dimension 3. For

the target environment, the intervention is $\begin{bmatrix}0 & 0 & 0 & 0\end{bmatrix}^\top$. Compared to Example 3, the inclusion of the causal covariates X_1 and X_2 makes the DA problem more complicated.

The population estimators when M tends to ∞ can be computed explicitly

$$\begin{bmatrix}\beta_{\text{OLSTar}}\\ \beta_{\text{OLSTar},0}\end{bmatrix} = \begin{bmatrix}1 & 0 & 0 & 0\end{bmatrix}^\top, \qquad \begin{bmatrix}\beta_{\text{Causal}}\\ \beta_{\text{Causal},0}\end{bmatrix} = \begin{bmatrix}1 & 0 & 0 & 0\end{bmatrix}^\top,$$
$$\begin{bmatrix}\beta_{\text{SrcPool}}\\ \beta_{\text{SrcPool},0}\end{bmatrix} = \begin{bmatrix}0 & 0 & 1 & 0\end{bmatrix}^\top, \qquad \begin{bmatrix}\beta^{(1)}_{\text{DIP}}\\ \beta^{(1)}_{\text{DIP},0}\end{bmatrix} = \begin{bmatrix}\frac{4}{3} & -\frac{1}{3} & -\frac{1}{6} & 2\end{bmatrix}^\top,$$
$$\begin{bmatrix}\beta_{\text{CIP}}\\ \beta_{\text{CIP},0}\end{bmatrix} = \begin{bmatrix}0 & 0 & 1 & 0\end{bmatrix}^\top, \qquad \begin{bmatrix}\beta^{(1)}_{\text{CIRM}}\\ \beta^{(1)}_{\text{CIRM},0}\end{bmatrix} = \begin{bmatrix}1 & 0 & 0 & 1\end{bmatrix}^\top.$$

Note that with a large number of source environments, CIP and SrcPool both identify the conditional invariant component X_3. The Causal estimator is the best because we made the noise variable on Y to be 0 and the intervention on Y in the target to be 0. DIP$^{(1)}$ does not find the Causal estimator because the Y intervention makes it difficult to align the covariate interventions between source and target. CIRM$^{(1)}$ does not find the Causal estimator because even if it can correct for the label intervention, matching on the causal covariate X_1 is not a good idea due to similar reason we have seen applying DIP for causal prediction in Example 1. The corresponding population source and target risks are summarize in Table 3. We conclude that unlike in the anticausal DA setting, in the presence of both causal and anticausal covariates, CIRM which is based on the idea of domain invariant projection after label correction no longer outperforms SrcPool.

Risk \ **Methods**	**OLSTar** (oracle)	**Causal**	**SrcPool**	**DIP**$^{(1)}$	**CIP**	**CIRM**$^{(1)}$
Ex 4, source risk $R^{(1)}$	1.000	1.000	0.100	0.000	0.100	0.000
Ex 4, target risk $\tilde{R}$	**0.000**	0.000	0.100	4.000	0.100	1.000

Table 3: Population source and target risks in Example 4 (the lower the better). The oracle target risk is highlighted in bold. In Example 3 of mixed-causal-anticausal prediction with intervention on Y, DIP$^{(1)}$, CIP and CIRM$^{(1)}$ all perform worse than Causal in terms of target population risk.

One might argue that the superior performance of Causal over SrcPool is made up because we set the target label intervention $\widetilde{a}_Y$ to be zero. The concern is valid in general. However, it is not difficult to observe that no matter how we set $\widetilde{a}_Y$ there exist better estimators than SrcPool. In fact, if we know X_1 is the only causal variable, we can regress $Y, X_2, X3$ on X_1 and consider the prediction problem on the residuals. The resulting prediction problem is anticausal with intervention on Y. Consequently, it can be solved with DA methods discussed in Section 4.2. We formulate this idea precisely in the next subsection.

4.3.2 New DA methods for mixed-causal-anticausal domain adaptation

Instead of studying the most general mixed-causal-anticausal domain adaptation, we first restrict ourselves to the setting where the "rough causal structure around Y" is known. That is, we know which covariates are the descendants of Y, denoted as X_{D}, and we also know which covariates are the parents of Y or the parents of X_{D} (denoted as X_{P}) as shown in Figure 6. Assuming the "rough causal structure", we show that this mixed-causal-anticausal problem can be reduced to the familiar problem in the previous subsections. Assumption 3 makes data generation requirements precise.

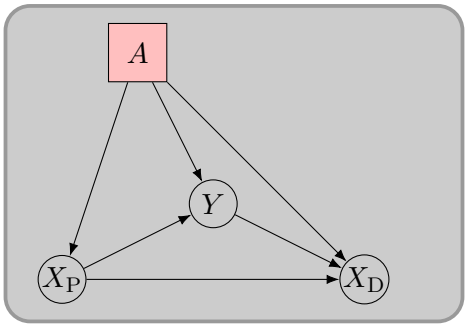


Figure 6: The causal diagram for our mixed-causal-anticausal domain adaptation setup.

Assumption 3 *There are M ($M \geq 1$) source environments. Each data point in the m-th source environment is generated i.i.d. according to distribution $\mathcal{P}^{(m)}$ specified by the following SCM*

$$\begin{bmatrix} X_P^{(m)} \\ X_D^{(m)} \\ Y^{(m)} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_P & 0 & 0 \\ \mathbf{B}_{D\text{-}P} & \mathbf{B}_D & b_D \\ \omega_P^\top & 0 & 0 \end{bmatrix} \begin{bmatrix} X_P^{(m)} \\ X_D^{(m)} \\ Y^{(m)} \end{bmatrix} + \begin{bmatrix} a_{X,P}^{(m)} \\ a_{X,D}^{(m)} \\ a_Y^{(m)} \end{bmatrix} + \begin{bmatrix} \varepsilon_{X,P}^{(m)} \\ \varepsilon_{X,D}^{(m)} \\ \varepsilon_Y^{(m)} \end{bmatrix}.$$

Each data point in the target environment is generated i.i.d. according to distribution $\widetilde{\mathcal{P}}$ specified with the same SCM except for the mean shift intervention term

$$\begin{bmatrix} \widetilde{X}_P \\ \widetilde{X}_D \\ \widetilde{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_P & 0 & 0 \\ \mathbf{B}_{D\text{-}P} & \mathbf{B}_D & b_D \\ \omega_P^\top & 0 & 0 \end{bmatrix} \begin{bmatrix} \widetilde{X}_P \\ \widetilde{X}_D \\ \widetilde{Y} \end{bmatrix} + \begin{bmatrix} \widetilde{a}_{X,P} \\ \widetilde{a}_{X,D} \\ \widetilde{a}_Y \end{bmatrix} + \begin{bmatrix} \widetilde{\varepsilon}_{X,P} \\ \widetilde{\varepsilon}_{X,D} \\ \widetilde{\varepsilon}_Y \end{bmatrix},$$

where $\mathbf{B}_P \in \mathbb{R}^{r\times r}$, $\mathbf{B}_D \in \mathbb{R}^{(d-r)\times(d-r)}$ and $\mathbf{B}_{D\text{-}P} \in \mathbb{R}^{(d-r)\times r}$ are unknown constant matrices such that $\mathbb{I}_d - \begin{bmatrix} \mathbf{B}_P & 0 \\ \mathbf{B}_{D\text{-}P} & \mathbf{B}_D \end{bmatrix}$ is invertible, $\omega_P \in \mathbb{R}^r$ and $b_D \in \mathbb{R}^{d-r}$ are unknown constant vectors, the intervention terms $\begin{bmatrix} a_{X,P}^{(m)} \\ a_{X,D}^{(m)} \\ a_Y^{(m)} \end{bmatrix}$ and $\begin{bmatrix} \widetilde{a}_{X,P} \\ \widetilde{a}_{X,D} \\ \widetilde{a}_Y \end{bmatrix}$ are vectors in $\mathbb{R}^{d+1}$, the noise terms

$\begin{bmatrix} \varepsilon_{X,P}^{(m)} \\ \varepsilon_{X,D}^{(m)} \\ \varepsilon_{Y}^{(m)} \end{bmatrix}$ *and* $\begin{bmatrix} \widetilde{\varepsilon}_{X,P} \\ \widetilde{\varepsilon}_{X,D} \\ \widetilde{\varepsilon}_{Y} \end{bmatrix}$ *share the same distribution with*

$$\mathbb{E}\begin{bmatrix} \varepsilon_{X,P}^{(m)} \\ \varepsilon_{X,D}^{(m)} \end{bmatrix} = \mathbb{E}\begin{bmatrix} \widetilde{\varepsilon}_{X,P} \\ \widetilde{\varepsilon}_{X,D} \end{bmatrix} = 0, \quad \mathbb{E}\left[\begin{bmatrix} \varepsilon_{X,P}^{(m)} \\ \varepsilon_{X,D}^{(m)} \end{bmatrix}\begin{bmatrix} \varepsilon_{X,P}^{(m)} \\ \varepsilon_{X,D}^{(m)} \end{bmatrix}^\top\right] = \mathbb{E}\left[\begin{bmatrix} \widetilde{\varepsilon}_{X,P} \\ \widetilde{\varepsilon}_{X,D} \end{bmatrix}\begin{bmatrix} \widetilde{\varepsilon}_{X,P} \\ \widetilde{\varepsilon}_{X,D} \end{bmatrix}^\top\right] = \begin{bmatrix} \Sigma_P & 0 \\ 0 & \Sigma_D \end{bmatrix},$$
$$\mathbb{E}\left[\varepsilon_Y^{(1)}\right] = \mathbb{E}\left[\widetilde{\varepsilon}_Y\right] = 0, \quad \mathbb{E}\left[{\varepsilon_Y^{(1)}}^2\right] = \mathbb{E}\left[\widetilde{\varepsilon}_Y^2\right] = \sigma^2.$$

Additionally, the noise terms on X *and* Y *are uncorrelated, namely,* $\mathbb{E}\left[\varepsilon_Y^{(m)} \cdot \begin{bmatrix} \varepsilon_{X,P}^{(m)} \\ \varepsilon_{X,D}^{(m)} \end{bmatrix}\right] = \mathbb{E}\left[\widetilde{\varepsilon}_Y \cdot \begin{bmatrix} \widetilde{\varepsilon}_{X,P} \\ \widetilde{\varepsilon}_{X,D} \end{bmatrix}\right] = 0.$

Compared to the SCMs in Assumption 1 and 2, Assumption 3 introduces additional covariates X_{P} which are the parents of Y or X_{D} and are located at the first r coordinates. In the most general DA problem, this information of which covariates are parents may not be available. The information about causal relationships can be obtained for example from domain experts or from the causal discovery literature. The causal discovery problem is well studied and we refer the readers to Glymour et al. (2019) for a review of the causal discovery methods. We leave the most general mixed-causal-anticausal domain adaptation and the design of new DA methods to future work.

Based on Assumption 3, we can introduce the intermediate random variables

$$\begin{aligned} X_{\mathcal{I}}^{(m)} &:= X_{\mathrm{D}}^{(m)} - (\mathbb{I}_{d-r} - \mathbf{B}_{\mathrm{D}})^{-1}\mathbf{B}_{\mathrm{D\text{-}P}}X_{\mathrm{P}}^{(m)} - (\mathbb{I}_{d-r} - \mathbf{B}_{\mathrm{D}})^{-1} b_{\mathrm{D}}\omega_{\mathrm{P}}^\top X_{\mathrm{P}}^{(m)} \\ Y_{\mathcal{I}}^{(m)} &:= Y^{(m)} - \omega_{\mathrm{P}}^\top X_{\mathrm{P}}^{(m)} \\ \widetilde{X}_{\mathcal{I}} &:= \widetilde{X}_{\mathrm{D}} - (\mathbb{I}_{d-r} - \mathbf{B}_{\mathrm{D}})^{-1}\mathbf{B}_{\mathrm{D\text{-}P}}\widetilde{X}_{\mathrm{P}} - (\mathbb{I}_{d-r} - \mathbf{B}_{\mathrm{D}})^{-1} b_{\mathrm{D}}\omega_{\mathrm{P}}^\top \widetilde{X}_{\mathrm{P}} \\ \widetilde{Y}_{\mathcal{I}} &:= \widetilde{Y} - \omega_{\mathrm{P}}^\top \widetilde{X}_{\mathrm{P}}. \end{aligned} \tag{37}$$

The intermediate random variable can be seen as the residuals after regressing on the parent variables X_{P}. We observe that the intermediate random variables satisfy the following SCMs

$$\begin{bmatrix} X_{\mathcal{I}}^{(m)} \\ Y_{\mathcal{I}}^{(m)} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_{\mathrm{D}} & b_{\mathrm{D}} \\ 0 & 0 \end{bmatrix}\begin{bmatrix} X_{\mathcal{I}}^{(m)} \\ Y_{\mathcal{I}}^{(m)} \end{bmatrix} + \begin{bmatrix} a_{X,\mathrm{D}}^{(m)} \\ a_Y^{(m)} \end{bmatrix} + \begin{bmatrix} \varepsilon_{X,\mathrm{D}}^{(m)} \\ \varepsilon_Y^{(m)} \end{bmatrix}$$
$$\begin{bmatrix} \widetilde{X}_{\mathcal{I}} \\ \widetilde{Y}_{\mathcal{I}} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_{\mathrm{D}} & b_{\mathrm{D}} \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \widetilde{X}_{\mathcal{I}} \\ \widetilde{Y}_{\mathcal{I}} \end{bmatrix} + \begin{bmatrix} \widetilde{a}_{X,\mathrm{D}} \\ \widetilde{a}_Y \end{bmatrix} + \begin{bmatrix} \widetilde{\varepsilon}_{X,\mathrm{D}} \\ \widetilde{\varepsilon}_Y \end{bmatrix}.$$

Using the intermediate random variables $(X_{\mathcal{I}}^{(m)}, Y_{\mathcal{I}}^{(m)})$, the mixed-causal-anticausal DA problem can be reduced to the anticausal DA problem. Intuitively, it works as follows. First, regress Y and X_{D} on the first r covariates X_{P}. Second, create a transformed dataset with new covariates and new labels from the corresponding residuals like how we define the intermediate random variables. Third, apply the DA methods in the anticausal DA setting

to the transformed dataset. Finally, bring back the original covariates and labels in the final estimator.

Based on the above intuition, we introduce the following estimators for the mixed-causal-anticausal DA setting.

- **DIP$\diamondsuit^{(m)}$-mean:** the population domain invariant projection estimator for the mixed-causal-anticausal DA setting.

$$f^{(m)}_{\mathrm{DIP}\diamondsuit}(x) := x^\top \begin{bmatrix} \gamma^{(m)} - \Gamma^{(m)} \beta^{(m)}_{\mathrm{DIP}\diamondsuit} \\ \beta^{(m)}_{\mathrm{DIP}\diamondsuit} \end{bmatrix} + \beta^{(m)}_{\mathrm{DIP}\diamondsuit,0}$$

$$\beta^{(m)}_{\mathrm{DIP}\diamondsuit}, \beta^{(m)}_{\mathrm{DIP}\diamondsuit,0} := \underset{\beta,\beta_0}{\arg\min}\ \mathbb{E}_{(X_\mathrm{P},X_\mathrm{D},Y)\sim\mathcal{P}^{(m)}} \left(Y_\mathcal{I} - X_\mathcal{I}^\top \beta - \beta_0\right)^2$$
$$\text{s.t. } \mathbb{E}_{(X_\mathrm{P},X_\mathrm{D})\sim\mathcal{P}^{(m)}_X} \left[X_\mathcal{I}^\top \beta\right] = \mathbb{E}_{(X_\mathrm{P},X_\mathrm{D})\sim\widetilde{\mathcal{P}}_X} \left[X_\mathcal{I}^\top \beta\right], \tag{38}$$

where $Y_\mathcal{I} := Y - X_\mathrm{P}^\top \gamma^{(m)}$, $X_\mathcal{I} := X_\mathrm{D} - {\Gamma^{(m)}}^\top X_\mathrm{P}$, and the regression weights $\gamma^{(m)}$ and $\Gamma^{(m)}$ are defined as

$$\gamma^{(m)}, \gamma^{(m)}_0 := \underset{\gamma,\gamma_0\in\mathbb{R}^r\times\mathbb{R}}{\arg\min}\ \mathbb{E}_{(X_\mathrm{P},X_\mathrm{D},Y)\sim\mathcal{P}^{(m)}} \left(Y - X_\mathrm{P}^\top \gamma - \gamma_0\right)^2, \tag{39}$$

$$\Gamma^{(m)}, \Gamma^{(m)}_0 := \underset{\Gamma,\Gamma_0\in\mathbb{R}^{r\times(d-r)}\times\mathbb{R}^{d-r}}{\arg\min}\ \mathbb{E}_{(X_\mathrm{P},X_\mathrm{D})\sim\mathcal{P}^{(m)}_X} \left\|X_\mathrm{D} - \Gamma^\top X_\mathrm{P} - \Gamma_0\right\|_2^2. \tag{40}$$

Corollary 7 *Under the data generation Assumption 3, $M = 1$ and the assumption of no intervention on Y i.e. $a_Y^{(1)} = \widetilde{a}_Y = 0$, the target population risks of OLSTar and DIP$\diamondsuit^{(1)}$-mean satisfy*

$$\tilde{R}\left(f_{OLSTar}\right) = \frac{\sigma^2}{1 + \sigma^2 b_D^\top \Sigma_D^{-1} b_D}, \tag{41}$$

$$\tilde{R}\left(f^{(1)}_{DIP\diamondsuit}\right) = \frac{\sigma^2}{1 + \sigma^2 b_D^\top \Sigma_D^{-\frac{1}{2}} G^{(1)}_{DIP\diamondsuit} \Sigma_D^{-\frac{1}{2}} b_D}, \tag{42}$$

where $G^{(1)}_{DIP\diamondsuit}$ is a projection matrix with rank $d-1$.

The proof of Corollary 7 is provided in Appendix C.4. The target risk results in Corollary 7 is very similar to those in Theorem 1. This is because we reduce the mixed-causal-anticausal DA problem without intervention on Y under Assumption 3 to the anticausal DA problem without intervention on Y under Assumption 1. According to Equation (42), the target population risk of DIP$\diamondsuit^{(1)}$ is lower than the target risk of Causal (which equals to σ^2).

Based on the intuition of DIP$\diamondsuit$, a similar strategy can be applied to extend CIP and CIRM to the mixed-causal-anticausal DA problems. The precise formulations of the extensions CIP$\diamondsuit$ and CIRM$\diamondsuit$ are introduced in Appendix A.1.

Just as Corollary 7 serves as the equivalent of Theorem 1 in the mixed-causal-anticausal DA setting, the equivalent of Theorem 5 can be established for CIP$\diamondsuit$ and CIRM$\diamondsuit$. Additionally, we can also introduce the weighted version of CIRM$\diamondsuit$, called CIRM$\diamondsuit$weigh, following the discussion of CIRMweigh. The details are omitted.

5. Numerical experiments

In this section, we numerically compare DA methods in simulated, synthetic and real datasets. The experiments are used to validate our theoretical results in finite sample data, to illustrate DA failure modes when assumptions are violated and to provide ideas of adapting DA methods to scenarios where not all assumptions are satisfied. Section 5.1 formulates the finite-sample DA estimators from the population DA ones to make sure our implementations are reproducible. Section 5.2 contains simulation experiments from deterministic or randomly generated linear SCMs. Section 5.3 discusses the performance of DA estimators on the MNIST dataset with synthetic interventions. Finally, Section 5.4 illustrates through a real data experiment that DA can be difficult in practice when not much domain knowledge about the data generating model is available.

Our code to reproduce all the numerical experiments is publicly available in the Github repository https://github.com/yuachen/CausalDA.

5.1 Finite-sample formulations and hyperparameter choices

Here we introduce the regularized formulations of the finite-sample DIP, CIP and CIRM. The DIP matching penalty in Equation (10) and the conditional invariance penalty in Equation (11) are enforced via regularization terms. The finite-sample versions of their variants can be formulated similarly after translating the constraints to regularization terms. For the sake of space, they are presented in Appendix A.2.

- **DIP$^{(m)}$-mean-finite:** the finite-sample formulation of the DIP$^{(m)}$-mean estimator (10). The mean squared difference is used as distributional distance and is enforced via a regularization term,

$$\begin{aligned} \hat{f}^{(m)}_{\mathrm{DIP}}(x) &:= x^\top \hat{\beta}^{(m)}_{\mathrm{DIP}} + \hat{\beta}^{(m)}_{\mathrm{DIP},0} \\ \hat{\beta}^{(m)}_{\mathrm{DIP}}, \hat{\beta}^{(m)}_{\mathrm{DIP},0} &:= \underset{\beta,\beta_0}{\arg\min}\ \frac{1}{n_m}\sum_{i=1}^{n_m}\left(y_i^{(m)} - {x_i^{(m)}}^\top \beta - \beta_0\right)^2 + \\ &\qquad \lambda_{\mathrm{match}}\left(\frac{1}{n_m}\sum_{i=1}^{n_m} {x_i^{(m)}}^\top \beta - \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \tilde{x}_i^\top \beta\right)^2, \end{aligned} \tag{43}$$

  where λ_{match} is a positive regularization parameter that controls the match between the covariate mean of the source and target environment.

- **CIP-mean-finite:** the finite sample formulation of the CIP-mean estimator (11). The conditional mean is matched across source environments and is enforced via a

regularization term,

$$\begin{aligned}\hat{f}_{\text{CIP}}(x) &:= x^\top \hat{\beta}_{\text{CIP}} + \hat{\beta}_{\text{CIP},0} \\ \hat{\beta}_{\text{CIP}}, \hat{\beta}_{\text{CIP},0} &:= \arg\min_{\beta,\beta_0} \frac{1}{M}\sum_{m=1}^{M}\frac{1}{n_m}\sum_{i=1}^{n_m}\left(y_i^{(m)} - {x_i^{(m)}}^\top \beta - \beta_0\right)^2 + \\ &\qquad \frac{\lambda_{\text{CIP}}}{M^2}\sum_{m,k=1}^{M}\left(\frac{1}{n_m}\sum_{i=1}^{n_m}{z_{\text{cond},i}^{(m)}}^\top \beta - \frac{1}{n_k}\sum_{i=1}^{n_k}{z_{\text{cond},i}^{(k)}}^\top \beta\right)^2,\end{aligned} \tag{44}$$

where $z_{\text{cond},i}^{(k)} = x_i^{(k)} - \frac{y_i^{(k)}\bar{y}^{(k)}}{\left(\frac{1}{n_k}\sum_{j=1}^{n_k} {y_j^{(k)}}^2\right)} x_i^{(k)}$, $\bar{y}^{(k)} = \frac{1}{n_k}\sum_{j=1}^{n_k} y_j^{(k)}$ and λ_{CIP} is a positive regularization parameter that controls the strength of the conditional invariant penalty. In the finite-sample formulation, the conditional expectation in the constraint of population formulation (11) is computed via regressing $X^\top\beta$ on Y. As a result, $z_{\text{cond},i}^{(k)}$'s are the residuals after regressing $x_i^{(k)}$ on $y_i^{(k)}$.

- **CIRM$^{(m)}$-mean-finite:** the finite-sample formulation of the CIRM$^{(m)}$-mean estimator (12). The residual after removing conditionally invariant components is matched between source and target environments. The matching is enforced via a regularization term.

$$\begin{aligned}\hat{f}_{\text{CIRM}}^{(m)}(x) &:= x^\top \hat{\beta}_{\text{CIRM}}^{(m)} + \hat{\beta}_{\text{CIRM},0}^{(m)} \\ \hat{\beta}_{\text{CIRM}}^{(m)}, \hat{\beta}_{\text{CIRM},0}^{(m)} &:= \arg\min_{\beta,\beta_0} \frac{1}{n_m}\sum_{i=1}^{n_m}\left(y_i^{(m)} - {x_i^{(m)}}^\top \beta - \beta_0\right)^2 \\ &\qquad + \lambda_{\text{match}}\left(\frac{1}{n_m}\sum_{i=1}^{n_m}{z_{\text{res},i}^{(m)}}^\top \beta - \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\tilde{z}_{\text{res},i}^\top \beta\right)^2,\end{aligned} \tag{45}$$

where $z_{\text{res},i}^{(m)} = x_i^{(m)} - \left({x_i^{(m)}}^\top \hat{\beta}_{\text{CIP}}\right)\hat{\vartheta}_{\text{CIRM}}$ with $\hat{\vartheta}_{\text{CIRM}}$ defined as

$$\hat{\vartheta}_{\text{CIRM}} = \frac{\frac{1}{M}\sum_{m=1}^{M}\frac{1}{n_m}\sum_{i=1}^{n_m}\left(y_i^{(m)} - \bar{y}^{(m)}\right)x_i^{(m)}}{\frac{1}{M}\sum_{m=1}^{M}\frac{1}{n_m}\sum_{i=1}^{n_m}\left(y_i^{(m)} - \bar{y}^{(m)}\right)\left({x_i^{(m)}}^\top \hat{\beta}_{\text{CIP}} - \frac{1}{n_m}\sum_{j=1}^{n_m}{x_j^{(m)}}^\top \hat{\beta}_{\text{CIP}}\right)},$$

and λ_{match} is a positive regularization parameter similar to the one define in DIP. Just like the population CIRM depends on the population CIP, the finite-sample CIRM also depends on the finite-sample CIP with the same regularization parameter λ_{CIP}.

Regularization parameter choices: Choosing the regularization parameters in finite-sample DA formulations such as λ_{match} in DIP formulation (43) is a difficult subject. Because the DA setting here assumes no access to any target labels, one cannot get good estimates of the target performance easily. The classical model selection strategies based on a validation set are no longer applicable in domain adaptation.

We make use of the fact that when DIP works as Theorem 1 predicts, the source population risk of DIP equals to the target population risk as shown in Corollary 3. The source finite-sample risk is close to the target finite-sample risk up to finite sample errors. So we would like to choose λ_{match} large enough so that the DIP matching penalty takes effect, but not too large so that the source finite risk remains reasonably small. We propose to choose the largest λ_{match} so that the source finite-sample risk is less than two times of the source finite-sample risk when λ_{match} is set to zero. The choice of "two times" is arbitrary here, the precise amount depends on the desired target finite-sample risk, the sample size and the variance of the source finite risk estimate. The precise amount is specified for each DA dataset separately. The regularization parameter λ_{match} in DIPweigh is chosen similarly.

Next, we consider the λ_{CIP} choice in CIP formulation (44). Since CIP only uses source data and never touches the target data, we leave a small part of each source data out for validation and choose λ_{CIP} based on the validation source data. We choose λ_{CIP} so that the average source risk across source environment is small.

Finally, the finite-sample CIRM formulation (45) requires the choice of both λ_{match} and λ_{CIP}. The λ_{CIP} in CIRM is the same as the λ_{CIP} in CIP. With λ_{CIP} fixed, we choose λ_{match} in CIRM as we did for DIP.

5.2 Linear SCM simulations

In this section, we numerically compare DA estimators via simulations from linear SCMs. First, we consider seven simulations with data generated via linear SCMs with mean shift noise interventions. The first three simulations aim at illustrating the results in Theorem 1 and Theorem 5. The last four are to show the performance of DA estimators when at least one assumption is misspecified. Second, we consider two simulations where the interventions are variance shift noise interventions. Through these two simulations, we demonstrate the necessity of adapting the DIP matching penalties according to the type of interventions. The simulation settings are summarized in Table 4.

5.2.1 Linear SCM with mean shift noise interventions

We consider seven simulations with data generated via linear SCMs with mean shift noise interventions. In all experiments in this subsection, we use large sample size $n = 5000$ for each environment, $n = 5000$ target test data for the evaluation of the target risk and we fix the regularization parameters $\lambda_{\text{match}} = 10.0$ and $\lambda_{\text{CIP}} = 1.0$ to focus the discussions on the choice of DA estimators.

(i) Single source anticausal DA without Y intervention: We consider three datasets of dimension $d = 3, 10, 20$. In each dataset, there is one source environment and one target environment generated according to Assumption 1. The matrix $\mathbf{B}$ and the vector b is specified as follows

- For $d = 3$, the matrix $\mathbf{B} = 0$, $b = [1, -1, 3]^\top$ and $\omega = 0$.
- For $d = 10, 20$, the matrix $\mathbf{B}$ is upper triangular with diagonal zero and with each entry drawn i.i.d. from $\mathcal{N}(0, 0.25)$. Each entry of b i.i.d. from $\mathcal{N}(0, 1)$. $\omega = 0$.

Sim Num	**# Src envs**	**Causal Direction**	**Interv X type**	**Interv on Y?**	**Has CIC?**	**Better** estimator(s)	**Baseline** estimator(s)
(i)	single	anticausal	mean shift	N	-	DIP$^{(1)}$	OLSSrc$^{(1)}$
(ii)	multiple	anticausal	mean shift	N	-	DIPweigh	OLSPool DIP$_{(1)}$
(iii)	multiple	anticausal	mean shift	Y	Y	CIRMweigh	OLSPool DIPweigh
(iv)	single	causal	mean shift	N	-	-	OLSSrc$^{(1)}$
(v)	single	mixed	mean shift	N	-	DIP$\diamondsuit^{(1)}$	OLSSrc$^{(1)}$ DIP$^{(1)}$
(vi)	multiple	anticausal	mean shift	Y	N	-	OLSPool
(vii)	multiple	mixed	mean shift	Y	Y	CIRM$\diamondsuit$weigh	OLSPool CIRMweigh
(viii)	single	anticausal	var shift	N	-	DIP-std+ DIP-MMD	OLSSrc$^{(1)}$ DIP
(ix)	multiple	anticausal	var shift	Y	Y	CIRMweigh-std+ CIRMweigh-MMD	OLSPool

Table 4: List of linear SCM simulations. The first three simulations are done under the correct assumptions according to the main theorems. Starting from Simulation (iv), some assumptions are modified to show the performance of DA methods under misspecified assumptions (highlighted in red). The last two simulations are done with variance shift noise intervention. "Y" means yes, "N" means no, "-" means the information is irrelevant in that setting.

The interventions $a_X^{(1)}$ and $\widetilde{a}_X$ are generated i.i.d. from $\mathcal{N}(0, \mathbb{I}_d)$ and stay the same for all data points in one environment. For each data point, $\varepsilon^{(1)}$ or ε is generated i.i.d. from the standard Gaussian distribution $\mathcal{N}(0, \mathbb{I}_{d+1})$. For each dataset, the matrix $\mathbf{B}$ and the vector b of the SCM are generated once; the noise and interventions are generated i.i.d. 10 times with source and target sample size $n = \tilde{n} = 5000$. The boxplot of the 10 target risks is reported for OLSTar, OLSSrc$^{(1)}$, DIP$^{(1)}$ and DIPOracle$^{(1)}$ in Figure 7. The target risk of OLSTar is highlighted in red dashed horizontal line. The target risk of OLSSrc$^{(1)}$ is highlighted in blue dashed horizontal line. The first three plots in Figure 7 show that the target risk of DIP$^{(1)}$ is similar to that of DIPOracle$^{(1)}$ and it is very close to OLSTar as predicted by Theorem 1.

The target risk performance of DA estimators depends on the how the matrix $\mathbf{B}$ and the vector b are generated. To make the comparison less dependent on the randomness in the matrix $\mathbf{B}$ and the vector b, we complement the boxplot with a scatterplot that compares the target risks of DIP$^{(1)}$ and OLSSrc$^{(1)}$ over 100 random generations of the matrix $\mathbf{B}$ for

$d = 10$. The last plot in Figure 7 shows that in 97 out of 100 random data generations, $\text{DIP}^{(1)}$ has lower target risk than $\text{OLSSrc}^{(1)}$.

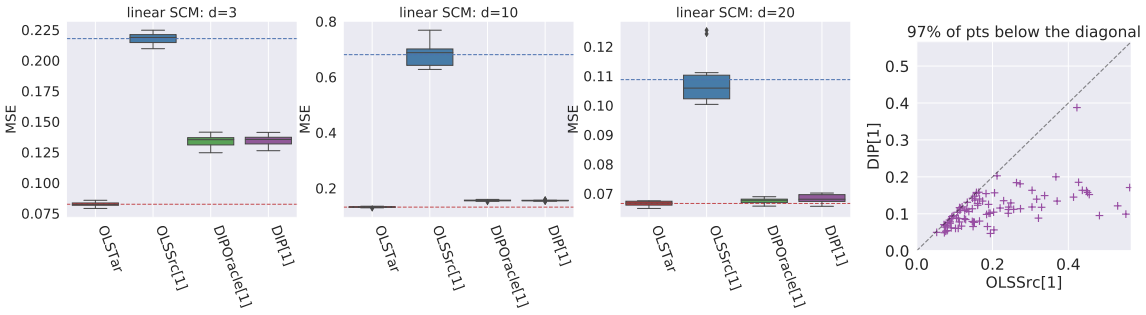


Figure 7: The target risk comparison in simulation (i) single source anticausal DA without Y intervention (the lower the better). In all three datasets ($d = 3, 10, 20$) in the left three plots, $\text{DIP}^{(1)}$ has lower target risk than $\text{OLSSrc}^{(1)}$. The target risk of $\text{DIP}^{(1)}$ is close to that of $\text{DIPOracle}^{(1)}$. The last plot shows that in 98 out of 100 random coefficient $\mathbf{B}$ data generations ($d = 10$), $\text{DIP}^{(1)}$ has lower target risk than $\text{OLSSrc}^{(1)}$.

(ii) Multiple source anticausal DA without Y intervention: We consider three datasets of covariate dimension $d = 3, 10, 20$. In each dataset, there is $M \geq 1$ source environments and one target environment generated according to Assumption 1. The matrix $\mathbf{B}$, the vector b, the interventions a and noises ε are generated as in simulation (i). The only difference is the availability of multiple ($M \geq 1$) source environments.

The boxplot of the 10 target risks are reported for OLSTar, $\text{DIP}^{(1)}$ and DIPweigh with increasing number of source environments in Figure 8. The constant ρ in DIPweigh formulation (24) is fixed to be 1000. To make the comparison less dependent on the randomness in the matrix $\mathbf{B}$, we complement the boxplot with a scatterplot that compares the target risks of DIPweigh($M = 8$) and $\text{DIP}^{(1)}$ over 100 random generations of the matrix $\mathbf{B}$ for $d = 10$. Figure 8 shows that in the anticausal DA setting without Y intervention, the more source environments the lower the target risk DIPweigh can achieve.

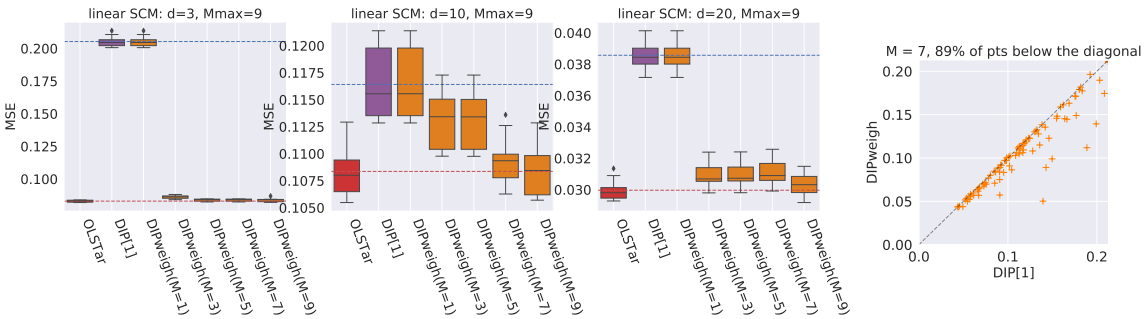


Figure 8: The target risk comparison in simulation (ii) multiple source anticausal DA without Y intervention (the lower the better). The left three plots show that with more number of source environments, DIPweigh perform better. The last plot shows that in 89 out of 100 simulations, DIPweigh with $M = 8$ has lower target risk than $\text{DIP}^{(1)}$.

(iii) Multiple source anticausal DA with Y intervention and with CICs: We fix the dimension $d = 20$ and the number of source environments $M = 14$. The source and target datasets are generated similarly to simulation (ii) with two exceptions: 1. there are interventions on Y for the target environment: $\widetilde{a}_Y$ follows $\mathcal{N}(0, 1)$ and it is generated once then it is fixed for the target environment. 2. there are conditionally invariant components (CICs): the interventions on X only apply to first 10 coordinates. That is, $a_X^{(m)}$ and $\widetilde{a}_X$ have the last 10 coordinates equal to zero. Note that we don't explicitly assume Equation (27) in Assumption 2. Instead we expect that a large number of source environments will span the vector space with last 10 coordinates zero and make the assumption (27) satisfied.

The left most plot in Figure 9 compares the boxplot of the target risks of OLSTar, SrcPool, DIPweigh, CIP and CIRMweigh with one time generation of $\mathbf{B}$ and 10 random generations of the interventions and the noises. CIRMweigh has lower target risk than SrcPool and DIPweigh. The right three plots in Figure 9 show the scatterplots comparing the pair-wise target risks in 100 random generations of $\mathbf{B}$ and the interventions for the following pairs: CIP vs DIPweigh, CIRMweigh vis SrcPool. CIRMweigh vs DIPweigh. The number of points below the diagonal are reported in the titles.

Both CIP and CIRM have lower target risk than DIPweigh. DIPweigh loses target risk guarantees because of the intervention on Y. CIRMweigh also outperforms SrcPool. Note that due to finite sample errors, there are rare cases (about 5%) where CIRM does not outperform SrcPool or DIPweigh.

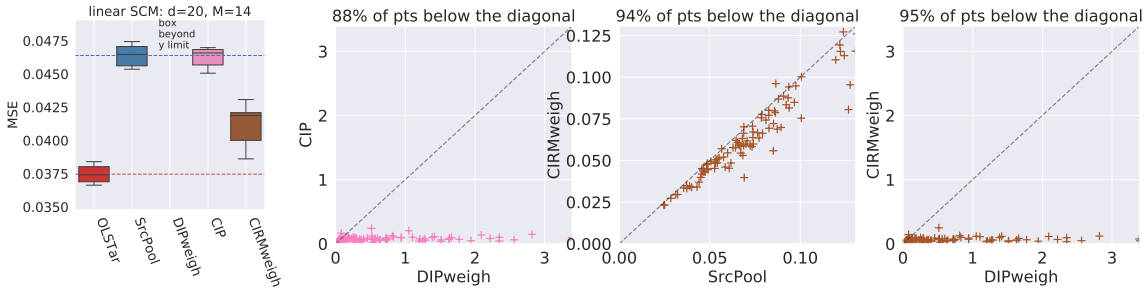


Figure 9: Target risk comparison in simulation (iii) multiple source anticausal DA with Y intervention and with CICs (the lower the better). CIRMweigh has smaller target risk than SrcPool. In the presence of Y intervention, the target risk of DIPweigh can be much larger than that of SrcPool. The scatterplots are for 100 random coefficient $\mathbf{B}$ data generations.

(iv) Single source causal DA without Y intervention: We consider three datasets of covariate dimension $d = 3, 10, 20$ similar to simulation (i) except that the prediction direction is changed from anticausal to causal: $\omega \neq 0$ and $b = 0$. Figure 10 shows that DIP$^{(1)}$ can have larger target risk than OLSSrc$^{(1)}$. This simulation result makes it clear that DIP is not useful for causal DA in general. Example 3 in Section 3.5.3 is not just a pathological failure example of DIP.

(v) Single source mixed DA without Y intervention: We consider two datasets of covariate dimension $d = 10, 20$ similar to simulation (i) except that the prediction direction is changed from anticausal to mixed-causal-anticausal: The odd coordinates of b are set to

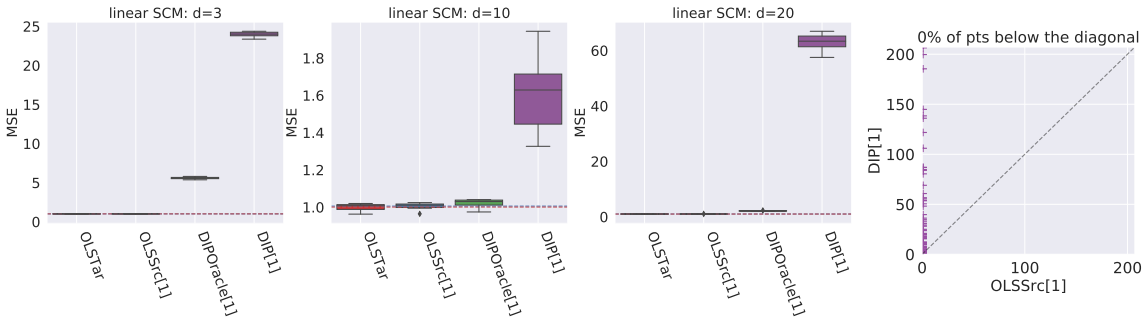


Figure 10: Target risk comparison in simulation (iv) single source causal domain adaptation without Y intervention (the lower the better). $\text{DIP}^{(1)}$ always has larger target risk than $\text{OLSSrc}^{(1)}$ in the causal domain adaptation problem over 100 random coefficient B data generations ($d = 10$).

zero and the even coordinates are nonzero. For ω, it is the contrary: the odd coordinates of ω are nonzero and the even ones are zero. The matrix $\mathbf{B}$ is generated according to Assumption 3 with one block zero. The left two plots in Figure 11 show boxplots of target risks of OLSTar, $\text{OLSSrc}^{(1)}$, $\text{DIPOracle}^{(1)}$, $\text{DIP}^{(1)}$ and $\text{DIP}\diamondsuit^{(1)}$ for random generation of the matrix $\mathbf{B}$. $\text{DIP}\diamondsuit^{(1)}$ uses the ground-truth knowledge of the nonzero coordinates of b. In mixed causal-anticausal DA, $\text{DIP}^{(1)}$ has larger risk than $\text{OLSSrc}^{(1)}$. With the ground-truth knowledge of the causal variables, $\text{DIP}\diamondsuit^{(1)}$ still have lower risk than $\text{OLSSrc}^{(1)}$. The right two plots in Figure 11 confirms the result via scatterplot comparisons of $\text{OLSSrc}^{(1)}$, $\text{DIPOracle}^{(1)}$, $\text{DIP}^{(1)}$ and $\text{DIP}\diamondsuit^{(1)}$ after 100 runs.

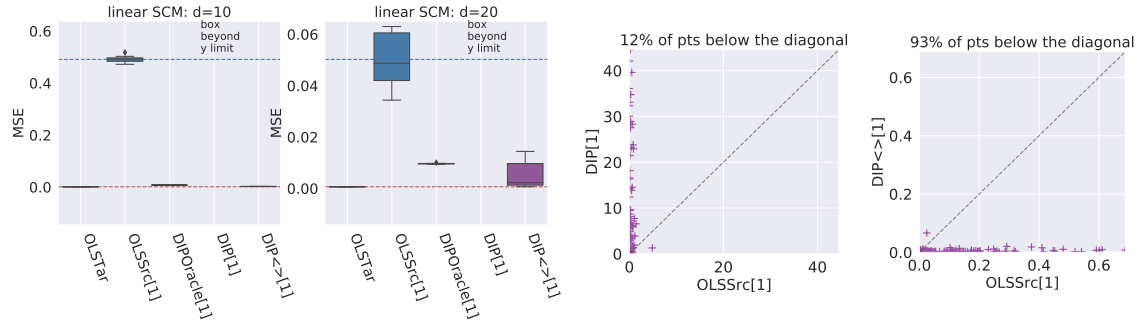


Figure 11: Target risk comparison in simulation (v) single source mixed DA without Y intervention (the lower the better). $\text{DIP}^{(1)}$ has larger target risk than $\text{OLSSrc}^{(1)}$ in the mixed causal anticausal DA setting. $\text{DIP}\diamondsuit^{(1)}$ with ground-truth causal variables outperforms $\text{OLSSrc}^{(1)}$. The two scatterplots are over 100 random coefficient $\mathbf{B}$ data generations ($d = 10$).

(vi) Multiple source anticausal DA with Y intervention and without CICs: We consider a simulation ($d = 20$, $M = 14$) similar to simulation (iii) except that there are no conditionally invariant components (CICs). Figure 12 shows that without CICs, CIRMweigh no longer outperforms SrcPool. Only in 64 of 100 simulations CIRMweigh has

lower target risk than SrcPool. Interestingly, CIRMweigh still has a large chance of having smaller target risk than DIPweigh. This may be due to the fact that even though there are no pure conditional invariant components, CIP may still be able to pick a combination of covariates that is relatively less sensitive to X interventions.

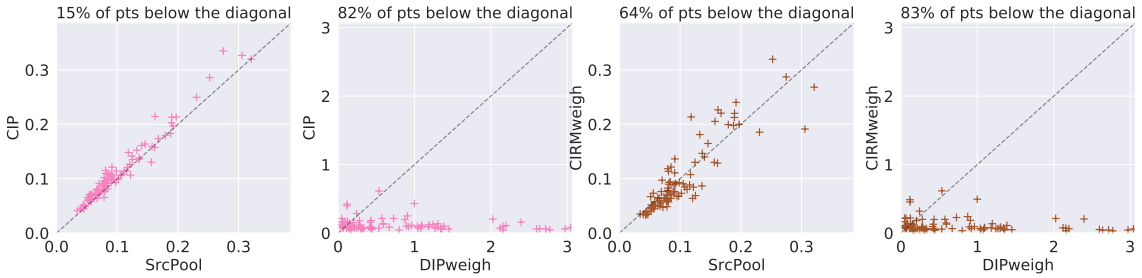


Figure 12: Target risk comparison in simulation (vi) multiple source anticausal DA with Y intervention and without CICs (the lower the better). Without CICs, CIRMweigh no longer outperforms SrcPool.

(vii) Multiple source mixed DA with Y intervention and with CICs: We consider a simulation ($d = 20$, $M = 14$) similar to simulation (iii) except that the prediction direction is changed from anticausal to mixed causal anticausal. The odd coordinates of b is set to zero and the even ones are nonzero. The even coordinates of ω is set to zero and the odd ones are nonzero. The first five even coordinates are set to be anticausal CICs. Additionally, we tuned down the variance of $\varepsilon_Y^{(m)}$ to be 0.01 to make the causal part important for prediction. Figure 13 shows that in mixed causal and anticausal DA, CIRMweigh has lower target risk than SrcPool only in 64 of 100 simulations. However, with the true causal covariates, the oracle estimator CIRM$\diamondsuit$weigh has lower target risk than SrcPool in 96 of 100 simulations.

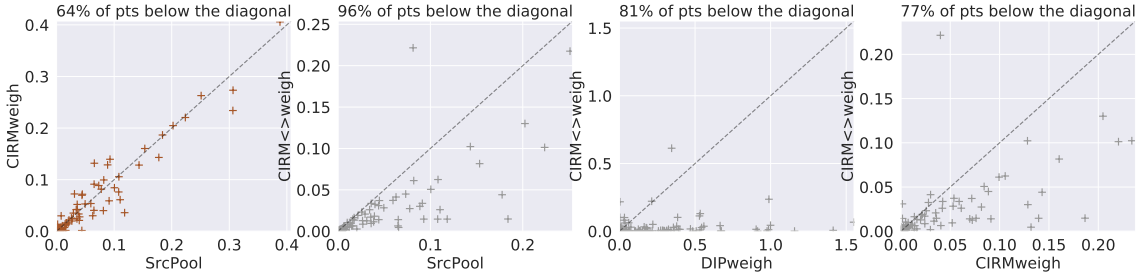


Figure 13: Target risk comparison in simulation (vii) multiple source mixed DA with Y intervention and with CICs (the lower the better). The scatterplots are over 100 random coefficient $\mathbf{B}$ data generations. CIRMweigh has lower target risk than SrcPool only in 64 of 100 simulations. However, knowing the indexes of the true causal covariates, the oracle estimator CIRM$\diamondsuit$weigh has lower target risk than SrcPool in 96 of 100 simulations. CIRM$\diamondsuit$weigh also outperforms DIPweigh and CIRMweigh. The scatterplots are over 100 random coefficient $\mathbf{B}$ data generations.

5.2.2 Linear SCM with variance shift noise interventions

We consider two simulations with data generated via linear SCMs with variance shift noise interventions. Since DIP-std+ and DIP-MMD are involved, we have to adapt the regularization parameter choice strategy described at the beginning of Section 5. We apply all DIP variants with the regularization parameter λ_{match} ranging in the set $\{10^{-5}, 10^{-4}, \cdots, 10^{4}, 10^{5}\}$. We choose the largest λ_{match} such that the source risk is smaller than $\min(2r, r + 0.01)$ as the final parameter. Here r is the source risk of OLSSrc. Based on the average source risk, λ_{CIP} is still chosen to be 1.0.

Since DIP-std+, DIP-MMD, CIRMweigh-std+ and CIRMweigh-MMD do not have closed form solutions, we use Pytorch's gradient descent to optimize these methods. Specifically, Adam's optimizer is used with step-size 10^{-4} and iteration number 2000 epochs to ensure convergence.

(viii) Single source anticausal DA without Y intervention + variance shift noise intervention: We consider a simulation ($d = 10$) similar to simulation (i) except that the type of intervention is changed from mean shift noise intervention to variance shift noise intervention. The intervention affects the variance of the noise $g(a_x^{(1)}, \varepsilon_X^{(1)}) = a_X^{(1)} \odot \varepsilon_X^{(1)}$. The interventions $a_X^{(1)}$ and $\widetilde{a}_X$ are still generated i.i.d. from $\mathcal{N}(0, \mathbb{I}_d)$ and stay the same for all data points in one environment. The left-most plot in Figure 14 shows the boxplot of the target risks of OLSTar, OLSSrc$^{(1)}$, DIP$^{(1)}$-mean, DIP$^{(1)}$-std+, DIP$^{(1)}$-MMD. DIP$^{(1)}$-mean has the same target risk and OLSSrc$^{(1)}$, because the DIP matching penalty on mean has no effect when the intervention is variance shift noise intervention. DIP$^{(1)}$-std+ and DIP$^{(1)}$-MMD improves upon DIP$^{(1)}$-mean. The right three plots in Figure 14 show scatterplots of 100 random coefficient $\mathbf{B}$ data generations to confirm this observation.

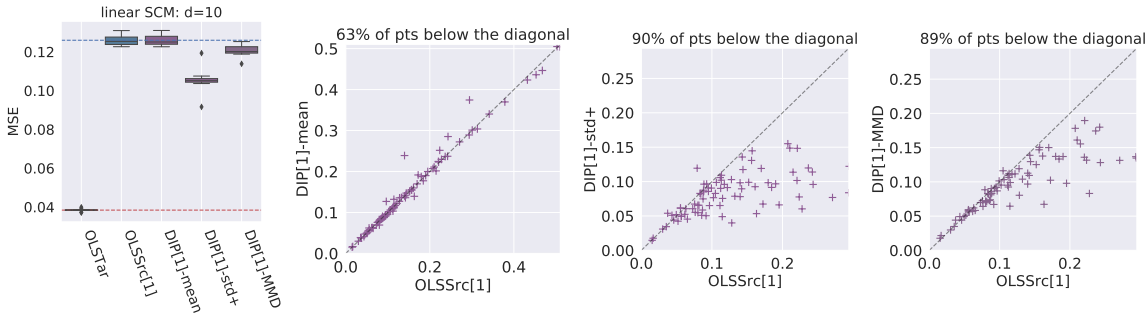


Figure 14: Target risk comparison in simulation (viii) single source causal domain adaptation with variance intervention without Y intervention (the lower the better). Left: boxplots of the target risks. DIP-mean has the same target risk than Src[1]. DIP-std+ and DIP-MMD outperform Src[1]. Right: three scatterplots of 100 runs comparing DIP-mean, DIP-std+ and DIP-MMD with Src[1].

(ix) Multiple source anticausal DA with Y intervention + variance shift noise intervention: We consider a simulation (d = 20, M = 14) similar to simulation (iii) except that the type of intervention is changed from mean shift noise intervention to variance shift noise intervention. The variance shift noise intervention on X is as specified simulation

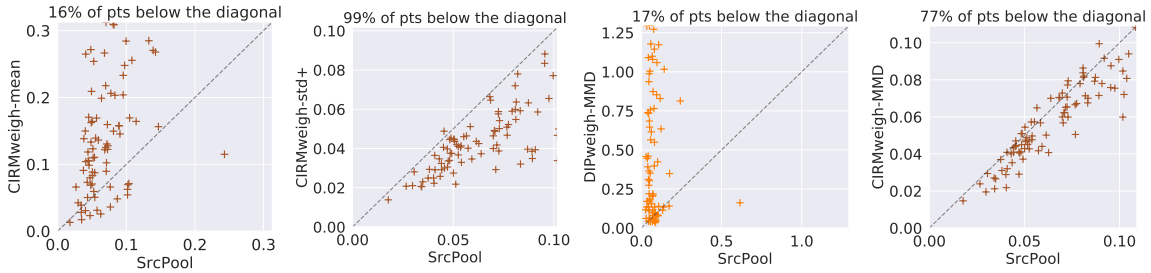


Figure 15: Target risk comparison in simulation (ix) multiples source causal domain adaptation with variance intervention with Y intervention and with conditionally invariant components (CICs) (the lower the better). CIRMweigh-std+ and CIRMweigh-MMD outperform SrcPool. CIRMweigh-mean does not outperform SrcPool, because it can not handle the variance shift noise intervention. DIPweigh-MMD performs much worse than SrcPool because of the intervention on Y.

(viii). The intervention on Y is still mean shift noise intervention as in simulation (iii). The scatterplots in Figure 15 compares CIRMweigh-mean, CIRMweigh-std+, DIPweigh-MMD, CIRMweigh-MMD with SrcPool in 100 runs. CIRMweigh-std+ and CIRMweigh-MMD outperform SrcPool. CIRMweigh-mean does not outperform SrcPool, because it can not handle the variance shift noise intervention. DIPweigh-MMD performs much worse than SrcPool because of the intervention on Y.

5.3 MNIST experiments with synthetic image interventions

In this section, we generate synthetic image classification datasets by modifying the images from the MNIST dataset (LeCun, 1998). Then we compare the performance of various DA methods on this digit classification task. The interventions are added directly on the image pixels. The DA methods are applied to the last-layer features of a pre-trained CNN on the original MNIST dataset. Since our DA methods are only defined for regression problems in the previous sections, for this classification problem, we adapt the loss function to softmax loss and we use one-hot encoding of the label Y in the DIP, CIP and CIRM formulations.

For the MNIST experiments with synthetic image interventions, a priori the linear SCM assumption is no longer valid and the interventions are not necessarily noise interventions. Here we show that the DA methods such as DIP, CIRM still have target performance as predicted by our theorems despite the likely violation of several assumptions.

All the methods in this subsection are implemented via Pytorch (Paszke et al., 2019) and are optimized with Pytorch's stochastic gradient descent. Specifically, stochastic gradient descent (SGD) optimizer is used with step-size (or learning rate) 10^{-4}, batchsize 500 and number of epochs 100.

5.3.1 MNIST with patch intervention

MNIST DA with patch intervention without Y intervention: We take the original MNIST dataset with 60000 training samples and create two synthetic datasets (source and target) as follows. For the source dataset, each training image is masked by the mask (a) in

Figure 16. That is, for each training image, the pixels at the white region of the mask (a) are set to white (maximum pixel value). For the target dataset environment, each training image is masked by the mask (b) in Figure 16. For each experiment, we take random 20% of samples from the source dataset as the source environment and random 20% of samples from the target dataset as the target environment. The task is to predict the labels of the images in the target environment without observing any target labels. This experiment is repeated 10 times and we report the boxplot of the 10 target classification accuracies for each DA method.

We apply the DA methods on the last-layer features of a pre-trained convolutional neural network (CNN). The CNN has two convolutional layers and two fully connected layers similar to the LeNet-5 architecture. It is pre-trained on the original MNIST dataset with test accuracy on the separate 10000 test images from the MNIST dataset being 98.8%. A priori, it is not clear what type of interventions on the last-layer features happen when we only know the intervention is on the pixels. To fit into our theoretical and conceptual framework, these induced interventions on the last-layer features would need to be approximated by shift noise interventions. The following DA methods are applied:

- Original: CNN model pre-trained on the original MNIST dataset without any modification.
- Tar: oracle CNN model trained only on the target environment. It is similar to OLSTar because only the weights in the last layer are trained. We changed the name to Tar because the full model is a neural network.
- $\text{Src}^{(1)}$: CNN model trained only on the source environment.
- $\text{DIP}^{(1)}$: CNN model where $\text{DIP}^{(1)}$-mean-finite is applied using the source label and the last-layer features of source and target environments.
- $\text{DIP}^{(1)}$-MMD: oracle CNN model where where $\text{DIP}^{(1)}$-MMD-finite is applied using the target label and the last-layer features of source and target environments.

Based on the discussion on the regularization parameter choice at the beginning of Section 5, we vary the regularization parameter λ_{match} from the set $\left\{10^k\right\}_{k=-5,\cdots,4}$, and we choose the largest λ_{match} such that the source accuracy has not dropped too much (in this case no more than 1% of the source accuracy of Src) as the final regularization parameter.

The left plot in Figure 17 shows that both $\text{DIP}^{(1)}$ and $\text{DIP}^{(1)}$-MMD achieve better target accuracy than Original or $\text{Src}^{(1)}$.

MNIST DA with patch intervention with Y intervention: We take the original MNIST dataset with 60000 training samples and create 12 synthetic datasets ($M = 11$ source environments and one target environment) as follows. For the m-th source dataset, each training image is masked by the m-th image (from left to right) in Figure 18. For the target dataset, each train or test image is masked by the right most image in Figure 18. The target dataset suffers additional Y intervention: for digits (3, 4, 5, 6, 8, 9) in the target dataset, 80% of the images in the MNIST dataset are removed from the target dataset. For experiment, we take random 20% of samples from the source datasets as the 11 source environments and 20% of samples from the target dataset as the target environment.

(a) mask (a) on 32×32 image

(b) mask (b) on 32×32 image

(c) data samples masked by mask (a)

(d) data samples masked by mask (b)

Figure 16: The two interventions in the MNIST single source domain adaptation without Y intervention. From left to right, mask (a), mask (b) and the corresponding data samples.

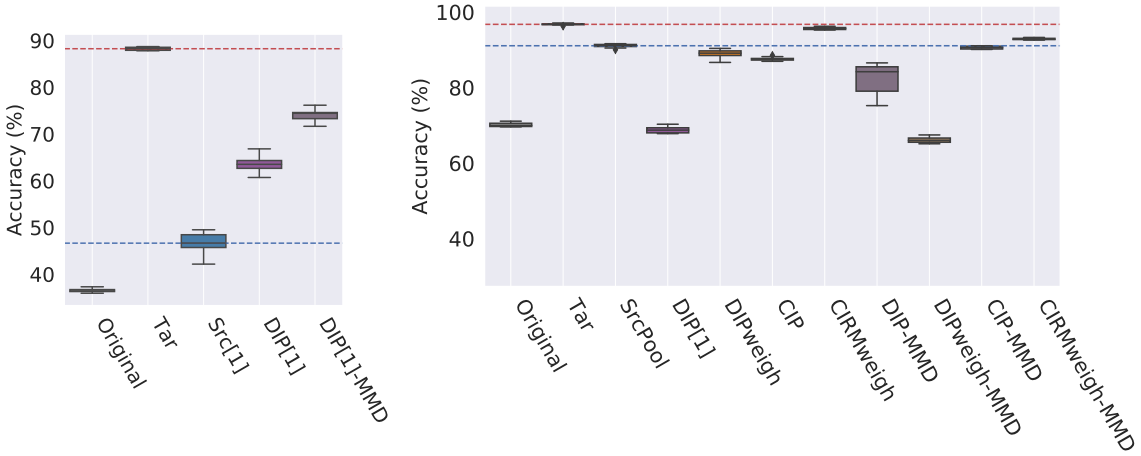


Figure 17: Left: Target accuracy comparison in MNIST experiment with patch intervention and single source without Y intervention. Right: Target accuracy comparison in MNIST experiment with patch intervention and multiple sources with Y intervention.

This experiment is repeated 10 times and we report the boxplot of 10 target classification accuracies for each DA method.

As the MNIST experiments above, we apply the DA methods on the last-layer features of the same pre-trained convolutional neural network (CNN). In addition to the methods in the MNIST experiments above, we also consider the following methods and their MMD variants:

- DIPweigh: CNN model where DIPweigh-mean-finite is applied on the last-layer features.
- CIP: CNN model where CIP-mean-finite is applied on the last-layer features.
- CIRMweigh: CNN model where CIRM-mean-finite is applied on the last-layer features.

The regularization parameter λ_{CIP} is chosen to be 0.1 based on source risk. For the regularization parameter λ_{match} in DIPweigh and CIRMweigh, we first output the source envi-

ronment index with the largest weight as "the best source". Then we vary it from the set $\{10^k\}_{k=-5,\cdots,4}$, and we choose the largest λ_{match} such that the source accuracy of "the best source" is not dropped too much (in this case not more than 1% of the source accuracy of Src applied to "the best source") as the final regularization parameter.

The right plot in Figure 17 compares the target accuracies of the DA methods in this setting. We observe that CIRMweigh and CIRMweigh-MMD outperforms SrcPool and Original. Due to the intervention on Y, the matching penalty of DIPweigh and DIPweigh-MMD is not useful and the two methods perform worse than SrcPool.

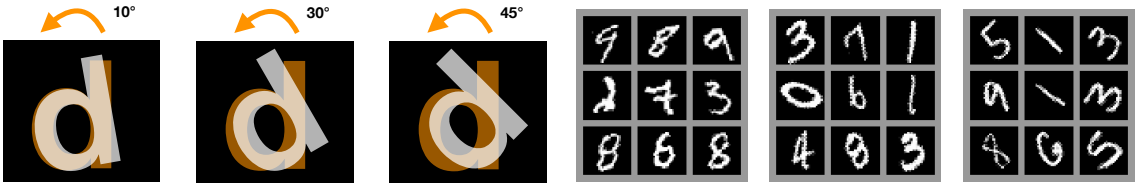
Figure 18: The 12 X interventions in the MNIST multiple source domain adaptation with Y intervention. From left to right, the first 11 are the source X interventions, the last is the target X intervention. The target environment suffers additional Y intervention.

5.3.2 MNIST with rotation intervention

MNIST DA with rotation intervention without Y intervention: We take the original MNIST dataset with 60000 training samples and create three synthetic datasets (2 sources and 1 target) as follows. For each dataset, each training image is rotated by one of the angles $\{10, 30, 45\}$ anti-clock-wise as shown in Figure 19. The three datasets are named Rotation 10°, Rotation 30° and Rotation 45° respectively. For the first experiment, we use Rotation 10° as the source environment and Rotation 45° as the target environment. For the second experiment, we use Rotation 30° as the source environment and Rotation 45° as the target environment. For each experiment, we take random 20% of samples from the source dataset as the source environment and random 20% of samples from the target dataset as the target environment. The task is to predict the labels of the images in the target environment without observing any target labels. Except for the change in the type of intervention, the other experimental settings are the same as in MNIST DA with patch intervention without Y intervention in Section 5.3.1.

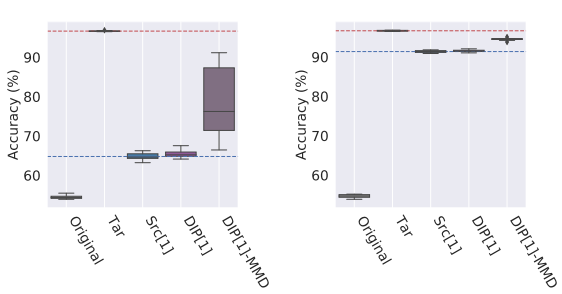
Figure 20a shows the boxplot of the target accuracies of Original, Tar, $\text{Src}^{(1)}$, $\text{DIP}^{(1)}$ and $\text{DIP}^{(1)}$-MMD for the first experiment. $\text{DIP}^{(1)}$-MMD achieves higher target accuracy than $\text{Src}^{(1)}$. Figure 20b shows the boxplot of 10 runs of Original, Tar, $\text{Src}^{(1)}$, $\text{DIP}^{(1)}$ and $\text{DIP}^{(1)}$-MMD for the second experiment. $\text{DIP}^{(1)}$-MMD achieves higher target accuracy than $\text{Src}^{(1)}$. Comparing Figure 20a and 20b, we also observe that the first experiment is a more difficult classification task than the second experiment as the accuracies achieved by our DA methods are lower in the first experiment.

MNIST DA with rotation intervention with Y intervention: We take the original MNIST dataset with 60000 training samples and create 5 synthetic datasets ($M = 4$ source environments and one target environment) as follows. For $m \in \{1, \cdots, 4\}$, for the m-th source dataset, each training image is rotated by $(m \times 15 - 30)^\circ$ clock-wise. Each image in the target dataset is rotated by 30° clock-wise. The target dataset suffers additional Y



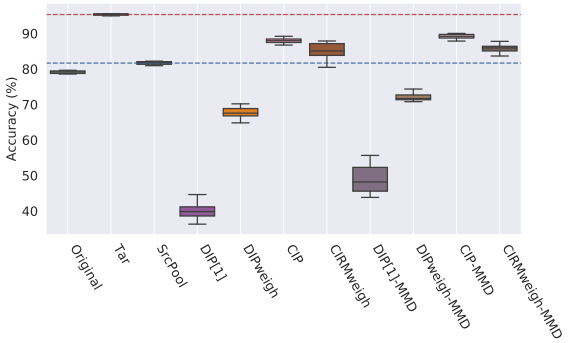
(a) Rotation10° (b) Rotation30° (c) Rotation45° (d) samples 10° (e) samples 30° (f) samples 45°

Figure 19: (a)(b)(c): the three interventions in the MNIST rotation intervention domain adaptation without Y intervention. (d)(e)(f): the corresponding data samples under rotation intervention.



(a) src: Rotation 10° tar: Rotation 45°

(b) src: Rotation 30° tar: Rotation 45°



(c) MNIST rotation interv. with Y interv.

Figure 20: (a) Target accuracy comparison in MNIST experiment with rotation intervention and single source without Y intervention from Rotation 10% to Rotation 45%. (b) Target accuracy comparison in MNIST experiment with rotation intervention and single source without Y intervention from Rotation 30% to Rotation 45%. (c) Target accuracy comparison MNIST experiment with rotation intervention and multiple source with Y intervention.

intervention: for digits (3, 4, 5, 6, 8, 9) in the target dataset, 80% of the images in the MNIST dataset are removed from the target dataset. For experiment, we take random 20% of samples from the source datasets as the 4 source environments and 20% of samples from the target dataset as the target environment. Except for the change in the type of intervention, the other experimental settings are the same as in MNIST DA with patch intervention with Y intervention in Section 5.3.1.

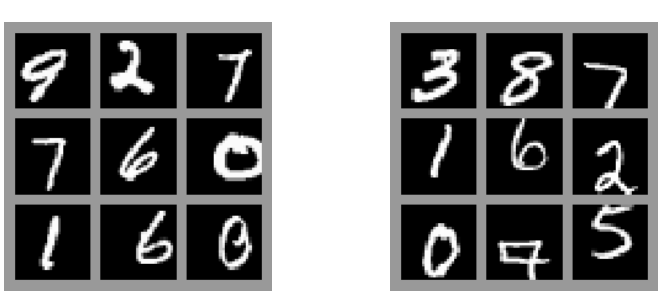
Figure 20c shows the boxplot of the target accuracies of Original, Tar, SrcPool, DIP$^{(1)}$, DIPweigh, CIP, CIRMweigh, DIP-MMD, CIP-MMD and CIRMweigh-MMD. CIP and CIP-MMD outperforms SrcPool and Original. Due to the intervention on Y, the matching penalty of DIPweigh and DIPweigh-MMD is not useful and the two methods perform worse than SrcPool. CIP, CIRMweigh and the corresponding MMD variants outperform SrcPool.

5.3.3 MNIST with random translation intervention

We take the original MNIST dataset with 60000 training samples and create two synthetic datasets (source and target) as follows. For the source dataset, each training image is

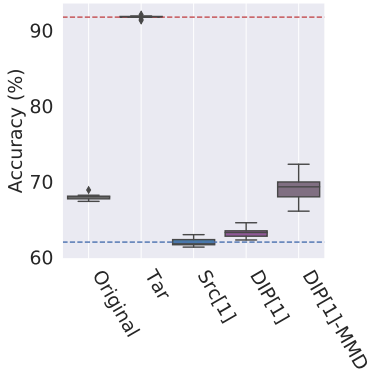
translated horizontally with a distance randomly selected from $-0.2 \times$ image width to $0.2 \times$ image width as shown in Figure 21a. For the target dataset environment, each training image is translated vertically with a distance randomly selected from $-0.2 \times$ image height to $0.2 \times$ image height as shown in Figure 21b. For each experiment, we take random 20% of samples from the source dataset as the source environment and random 20% of samples from the target dataset as the target environment. The task is to predict the labels of the images in the target environment without observing any target labels. Except for the change in the type of intervention, the other experimental settings are the same as in MNIST DA with patch intervention without Y intervention in Section 5.3.1.

Figure 21c shows the boxplot of the target accuracies of Original, Tar, $\text{Src}^{(1)}$, $\text{DIP}^{(1)}$ and $\text{DIP}^{(1)}$-MMD. $\text{DIP}^{(1)}$-MMD still performs better than $\text{Src}^{(1)}$, but it barely improves over Original. Since the intervention is random rather than fixed for each environment, intuitively this experiment is more difficult for our DA methods to have high accuracy than the first two experiments with fixed intervention for each environment.



(a) source samples after random horizontal translation

(b) target samples after random vertical translation



(c) MNIST translation interv.

Figure 21: (a) Source samples after random horizontal translation. (b) Target samples after random vertical translation. (c) Target accuracy comparison in MNIST experiment with random translation intervention and single source without Y intervention.

5.4 Experiments on Amazon review dataset with unknown interventions

In this subsection, we compare DA methods on the regression dataset Amazon Review Data (Ni et al., 2019). Unlike the simulated experiments or the MNIST experiment, neither the data generation process nor the type of interventions is known.

This dataset contains product reviews and metadata from Amazon in the date range May 1996 - Oct 2018. In our experiment, we consider the task of predicting review rating from review text from various product categories (Automotive, Digital Music, Office Products etc.). For each review, the covariates X are the TF-IDF features generated from raw review text; the label Y is the review rating (score from $1-5$). The different product categories are used as source and target environments. We take 15 product categories with the largest sample sizes, use 14 of them as source environments and leave the last one as the target environment.

The samples size n is 10000 in each source or target environment. The dimension depends on the TF-IDF feature extractor. Here we use both unigrams and bigrams, and build the vocabulary with terms that have a document frequency not smaller than 0.008. This results in feature dimension $d = 482$. A linear model with ℓ_2 regularization is used on top of the TF-IDF features to predict ratings.

Without explicit knowledge about the causal structure or the type of interventions, a priori it is no longer clear which DA method is the best. We apply SrcPool and the advanced DA methods, DIPweigh, CIP, CIRMweigh on the linear model. Figure 22 shows that the advanced DA methods do not outperform SrcPool except for target environment number 4.

We did an additional experiment with a small portion of target labels revealed. We use the small portion of target labels to choose the best method out of SrcPool, DIPweigh, CIP and CIRMweigh based on the small portion of labeled target data. The last three boxes (labeled as best20, best40 and best60) in each subplot of Figure 22 show that with 20, 40, 60 target labels revealed, the best method based on the small portion of labeled target data is always not worse than SrcPool and can sometimes outperform SrcPool.

We arrive at a conclusion that in general, without explicit knowledge about causal structure or the type of interventions, it is ambitious to expect advanced DA methods to always outperform SrcPool. However, with a small portion of labeled target data, we show that DA methods still lead to substantial improvements. Our empirical observation agrees with the concurrent empirical study of DA methods by Wiles et al. (2021). In a larger-scale experimental framework with many real and synthetic datasets, they find that DA methods can outperform SrcPool in some settings but there is no single best method across all settings. Hence, knowledge about the distribution shift is necessary for successful DA with guarantees.

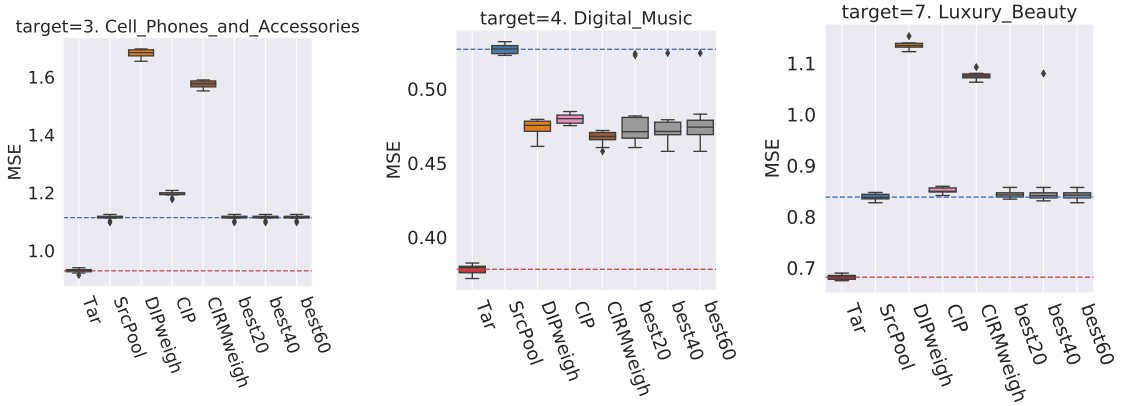


Figure 22: Target risk in Amazon review data prediction experiment (the lower the better). Without explicit knowledge about the type of interventions, it is no longer clear which domain adaptation method is the best. From left to right, depending on the target environment choice, the best methods are CIP, DIPweigh, CIRMweigh accordingly.

6. Discussion

In this paper, we propose a theoretical framework using structural causal models (SCMs) to analyze and obtain insights on prediction performance of several popular domain adaptation methods. First, we show that under the assumption of anticausal prediction, linear SCM and no intervention on Y, the popular DA method DIP achieves a low target risk. This theoretical result is compatible with many previous empirical results on the usefulness of DIP for domain adaptation. Second, we derive several conditions where DIP fails to achieve a low target risk: when the prediction problem is causal or mixed-causal-anticausal; when there is label distribution perturbation. To tackle these difficult DA scenarios, we design a new DA method called CIRM, among other modifications. The theoretical analysis is complemented with simulation analysis and real data experiments. The theoretical extension to nonlinear SCMs is a challenging future direction. However our empirical results suggest that the linear SCM assumption provides useful insights even for cases where a linear SCM does not hold.

The real data experiments show that it can be beneficial to use DA methods when the anticausal prediction assumption is satisfied but it can also be dangerous to blindly use DA methods when little is known about the data generation process. Thus, prior knowledge of the underlying causal structure and the types of interventions are often crucial. The prior knowledge of the causal structure can come from domain expert knowledge or causal studies on related datasets with the same variables. How to seamlessly combine causal studies about the causal structure and domain adaptation with appropriate uncertainty quantification is one promising future direction.

In absence of prior information about the causal structure and the types and locations of interventions, one needs to select good DA methods. The assessment of the goodness of a model or algorithm is difficult in general but it can be done when having access to a small fraction of labeled target data. We show empirically that a small fraction of labeled target data substantially helps to select the best DA method. It remains an open question what is the minimal amount of labeled target data points in order to guarantee the best DA method selection.

Acknowledgments

Y. Chen and P. Bühlmann have received funding from the European ResearchCouncil under the Grant Agreement No 786461 (CausalStats - ERC-2017-ADG). They both acknowledge scientific interaction and exchange at “ETH Foundations of Data Science”. They also thank Domagoj Cevid, Christina Heinze-Deml, Wooseok Ha, Jinzhou Li, Nicolai Meinshausen, Armeen Taeb, Fanny Yang and Andrii Zadaianchuk for fruitful discussions and for helpful suggestions on presentation and writing.

Appendix A. Summary of DA methods in this paper

In this section, we provide a summary of the DA methods presented in this paper. Methods that do not fit in the main paper due to the space limitation are formally introduced here.

A.1 Population DA methods

First, we formulate the DIP variants that adapt to other types of intervention as mentioned in Section 4.1.3.

- **DIP$^{(m)}$-std:** the population DIP estimator where the difference between the source and target standard deviations is used as distributional distance. For the m-th source environment, it is defined as

$$\begin{aligned} f^{(m)}_{\text{DIP-std}}(x) &:= x^\top \beta^{(m)}_{\text{DIP-std}} + \beta^{(m)}_{\text{DIP-std},0} \\ \beta^{(m)}_{\text{DIP-std+}}, \beta^{(m)}_{\text{DIP-std},0} &:= \underset{\beta,\beta_0}{\arg\min}\ \mathbb{E}_{(X,Y)\sim\mathcal{P}^{(m)}} \left(Y - X^\top\beta - \beta_0\right)^2 \\ &\quad \text{s.t. } \operatorname{Var}_{X\sim\mathcal{P}^{(m)}_X}\left[X^\top\beta\right] = \operatorname{Var}_{X\sim\widetilde{\mathcal{P}}_X}\left[X^\top\beta\right]. \end{aligned} \tag{46}$$

- **DIP$^{(m)}$-std+:** the population DIP estimator where the differences between the source and target means, standard deviations and 25% quantiles are used as distributional distance

$$\begin{aligned} f^{(m)}_{\text{DIP-std+}}(x) &:= x^\top \beta^{(m)}_{\text{DIP-std+}} + \beta^{(m)}_{\text{DIP-std+},0} \\ \beta^{(m)}_{\text{DIP-std+}}, \beta^{(m)}_{\text{DIP-std+},0} &:= \underset{\beta,\beta_0}{\arg\min}\ \mathbb{E}_{(X,Y)\sim\mathcal{P}^{(m)}} \left(Y - X^\top\beta - \beta_0\right)^2 \\ &\quad \text{s.t. } \mathbb{E}_{X\sim\mathcal{P}^{(m)}_X}\left[X^\top\beta\right] = \mathbb{E}_{X\sim\widetilde{\mathcal{P}}_X}\left[X^\top\beta\right] \\ &\quad \operatorname{Var}_{X\sim\mathcal{P}^{(m)}_X}\left[X^\top\beta\right] = \operatorname{Var}_{X\sim\widetilde{\mathcal{P}}_X}\left[X^\top\beta\right] \\ &\quad \psi_{25\%}\left({X^{(m)}}^\top\beta\right) = \psi_{25\%}\left(\widetilde{X}^\top\beta\right), \end{aligned} \tag{47}$$

  where $\psi_{25\%}$ is the 25% quantile function which takes a random variable and returns its 25% quantile.

- **DIP$^{(m)}$-MMD:** the population DIP estimator where the maximum mean discrepancy (MMD) is used as distributional distance.

$$\begin{aligned} f^{(m)}_{\text{DIP-MMD}}(x) &:= x^\top \beta^{(m)}_{\text{DIP-MMD}} + \beta^{(m)}_{\text{DIP-MMD},0} \\ \beta^{(m)}_{\text{DIP-MMD}}, \beta^{(m)}_{\text{DIP-MMD},0} &:= \underset{\beta,\beta_0}{\arg\min}\ \mathbb{E}_{(X,Y)\sim\mathcal{P}^{(m)}} \left(Y - X^\top\beta - \beta_0\right)^2 \\ &\quad \text{s.t. } \mathcal{D}_{\text{MMD},\mathcal{H}}\left({X^{(m)}}^\top\beta, \widetilde{X}^\top\beta\right) = 0, \end{aligned} \tag{48}$$

  where the maximum mean discrepancy (MMD) Gretton et al. (2012) with respect to the reproducing kernel Hilbert space (RHKS) $\mathcal{H}$ between two random variable Z_1 and

Z_2 is

$$\mathcal{D}_{\mathrm{MMD},\mathcal{H}}(Z_1, Z_2) = \sup_{h\in\mathcal{H}} \left|\mathbb{E}\left[h(Z_1)\right] - \mathbb{E}\left[h(Z_2)\right]\right|.$$

By default, the RKHS with Gaussian kernel is used throughout this paper.

Second we introduce the weighted version of CIRM following DIPweigh (24).

- **CIRMweigh-mean:** the population CIRM estimator that weights the source environments based on the source risks. It is defined as follows

$$f_{\mathrm{CIRMweigh}}(x) := \frac{1}{\sum_{m=1}^{M} e^{-\eta\cdot s_m}} \sum_{m=1}^{M} e^{-\eta\cdot s_m} \left(x^\top \beta_{\mathrm{CIRM}}^{(m)} + \beta_{\mathrm{CIRM},0}^{(m)}\right)$$
$$s_m := R^{(m)}\left(f_{\mathrm{CIRM}}^{(m)}\right). \tag{49}$$

  Here $\eta > 0$ is a constant. Choosing η to be ∞ is equivalent to choosing the source estimator with the lowest source risk.

The rational behind the use of source risks to weigh the environments follows from the corollary below.

Corollary 8 *Under the data generation Assumption 2, the m-th source population risk* (2) *of CIRM$^{(m)}$-mean satisfies*

$$R^{(m)}\left(f_{CIRM}^{(m)}\right) = \tilde{R}\left(f_{CIRM}^{(m)}\right) + \frac{\left(a_Y^{(m)} - \widetilde{a}_Y\right)^2}{\left(1 + \sigma^2 b^\top \Sigma^{-\frac{1}{2}} G_{DIP}^{(m)} \Sigma^{-\frac{1}{2}} b\right)^2} \tag{50}$$

The proof of Corollary 8 is provided in Appendix C.3. Comparing Corollary 8 with Corollary 3, the source risk of CIRM is no longer exactly equal to its target risk. The source risk has an additional term that depends on the intervention on Y. However, in the case where Σ has eigenvalues much smaller than σ, this additional term is negligible. In these scenarios, the source risk of CIRM still constitutes a good approximation of the target risk of CIRM. It is still possible to apply CIRM for each $m \in \{1, \ldots, M\}$ and pick the source environment with the lowest source population risk in order to reduce the target population risk. Based on the above ideas, we introduce the following weighted version of CIRM.

Third we introduce the CIP and CIRM extensions to deal with mixed-causal-anticausal DA problems in Section 4.3.

- **CIP$\diamondsuit$-mean:** the population conditional invariance penalty estimator for the mixed causal anticausal DA setting.

$$f_{\mathrm{CIP}\diamondsuit}(x) := x^\top \begin{bmatrix} \gamma^* - \Gamma^* \beta_{\mathrm{CIP}\diamondsuit} \\ \beta_{\mathrm{CIP}\diamondsuit} \end{bmatrix} + \beta_{\mathrm{CIP}\diamondsuit,0}^{(m)}$$
$$\beta_{\mathrm{CIP}\diamondsuit}, \beta_{\mathrm{CIP}\diamondsuit,0} := \arg\min_{\beta,\beta_0} \frac{1}{M} \sum_{k=1}^{M} \mathbb{E}_{(X_{\mathrm{P}},X_{\mathrm{D}},Y)\sim\mathcal{P}^{(k)}} \left(Y_{\mathcal{I}} - X_{\mathcal{I}}{}^\top \beta - \beta_0\right)^2$$
$$\text{s.t. } \mathbb{E}_{(X_{\mathrm{P}},X_{\mathrm{D}},Y)\sim\mathcal{P}^{(m)}} \left[X_{\mathcal{I}}{}^\top \beta \mid Y_{\mathcal{I}} = y\right] = \mathbb{E}_{(X_{\mathrm{P}},X_{\mathrm{D}},Y)\sim\mathcal{P}^{(1)}} \left[X_{\mathcal{I}}{}^\top \beta \mid Y_{\mathcal{I}} = y\right],$$
$$\forall y \in \mathbb{R}, \forall m \in \{1, \ldots, M\}, \tag{51}$$

where $Y_{\mathcal{I}} := Y - X_{\mathrm{P}}^\top \gamma^*$, $X_{\mathcal{I}} := X_{\mathrm{D}} - {\Gamma^*}^\top X_{\mathrm{P}}$,

$$\gamma^*, \gamma_0^* := \operatorname*{arg\,min}_{\gamma, \gamma_0 \in \mathbb{R}^r \times \mathbb{R}} \mathbb{E}_{(X_{\mathrm{P}}, X_{\mathrm{D}}, Y) \sim \mathcal{P}^{\mathrm{allsrc}}} \left(Y - X_{\mathrm{P}}^\top \gamma - \gamma_0 \right)^2,$$
$$\Gamma^*, \Gamma_0^* := \operatorname*{arg\,min}_{\Gamma, \Gamma_0 \in \mathbb{R}^{r \times (d-r)} \times \mathbb{R}^{d-r}} \mathbb{E}_{(X_{\mathrm{P}}, X_{\mathrm{D}}) \sim \mathcal{P}_X^{\mathrm{allsrc}}} \left\| X_{\mathrm{D}} - \Gamma^\top X_{\mathrm{P}} - \Gamma_0 \right\|_2^2, \tag{52}$$

and $\mathcal{P}^{\mathrm{allsrc}}$ denotes the uniform mixture of all source distribution.

- **CIRM$\diamondsuit^{(m)}$-mean:** the population conditional invariant residual matching estimator using m-th source environment for the mixed causal anticausal DA setting.

$$f^{(m)}_{\mathrm{CIRM}\diamondsuit}(x) := x^\top \begin{bmatrix} \gamma^* - \Gamma^* \beta^{(m)}_{\mathrm{CIRM}\diamondsuit} \\ \beta^{(m)}_{\mathrm{CIRM}\diamondsuit} \end{bmatrix} + \beta^{(m)}_{\mathrm{CIRM}\diamondsuit,0}$$
$$\beta^{(m)}_{\mathrm{CIRM}\diamondsuit}, \beta^{(m)}_{\mathrm{CIRM}\diamondsuit,0} := \operatorname*{arg\,min}_{\beta, \beta_0} \mathbb{E}_{(X_{\mathrm{P}}, X_{\mathrm{D}}, Y) \sim \mathcal{P}^{(m)}} \left(Y_{\mathcal{I}} - X_{\mathcal{I}}^\top \beta - \beta_0 \right)^2$$
$$\text{s.t. } \mathbb{E}_{(X_{\mathrm{P}}, X_{\mathrm{D}}) \sim \mathcal{P}_X^{(m)}} \left[\beta^\top \left(X_{\mathcal{I}} - \left(X_{\mathcal{I}}^\top \beta_{\mathrm{CIP}\diamondsuit} \right) \vartheta_{\mathrm{CIRM}\diamondsuit} \right) \right]$$
$$= \mathbb{E}_{(X_{\mathrm{P}}, X_{\mathrm{D}}) \sim \widetilde{\mathcal{P}}_X} \left[\beta^\top \left(X_{\mathcal{I}} - \left(X_{\mathcal{I}}^\top \beta_{\mathrm{CIP}\diamondsuit} \right) \vartheta_{\mathrm{CIRM}\diamondsuit} \right) \right], \tag{53}$$

where $Y_{\mathcal{I}} := Y - X_{\mathrm{P}}^\top \gamma^*$, $X_{\mathcal{I}} := X_{\mathrm{D}} - {\Gamma^*}^\top X_{\mathrm{P}}$ with γ^* and Γ^* defined in Equation (52) and

$$\vartheta_{\mathrm{CIRM}\diamondsuit} := \frac{\mathbb{E}_{(X_{\mathrm{P}}, X_{\mathrm{D}}, Y) \sim \mathcal{P}^{\mathrm{allsrc}}} \left[X_{\mathcal{I}} \cdot (Y_{\mathcal{I}} - \mathbb{E}[Y_{\mathcal{I}}]) \right]}{\mathbb{E}_{(X_{\mathrm{P}}, X_{\mathrm{D}}, Y) \sim \mathcal{P}^{\mathrm{allsrc}}} \left[\left(X_{\mathcal{I}}^\top \beta_{\mathrm{CIP}\diamondsuit} - \mathbb{E}[X_{\mathcal{I}}^\top \beta_{\mathrm{CIP}\diamondsuit}] \right) \cdot (Y_{\mathcal{I}} - \mathbb{E}[Y_{\mathcal{I}}]) \right]}, \tag{54}$$

with $\mathcal{P}^{\mathrm{allsrc}}$ denote the uniform mixture of all source distributions.

Finally, we summarize in Table 5 all the population DA methods introduced in this paper.

A.2 Finite-sample formulation of DA methods

We introduce the finite-sample formulations of the population DA methods in the previous subsection so that they can be implemented to reproduce the results in the numerical experiments in Section 5. The finite-sample DA formulations are summarized in Table 6.

The finite-sample formulations of DIP-mean, CIP-mean and CIRM-mean are introduced in Section 5.1. Here we state the finite-sample formulations of weighted variants, matching penalty and mixed-causal-anticausal variants.

To formulate the finite-sample versions of DIPweigh-mean (24) and CIRMweigh-mean (49), it is sufficient to replace the corresponding population estimators and the population source risks with finite-sample ones. The mixed-causal-anticausal variants of DIP, CIP and CIRM only requires two additional regressions. They are omitted for the sake of space.

Next, we introduce the finite-sample formulations of the matching penalty variants for DIP.

Original	**Made-ups to assist explanation**	**Weighted variants**	**Matching penalty variants**	**Mixed variants**
DIP$^{(m)}$ (10)	DIPAbs$^{(m)}$ (14) DIPOracle$^{(m)}$ (15)	DIPweigh (24)	DIP-std (46) DIP-std+ (47) DIP-MMD (48)	DIP$\diamondsuit^{(m)}$ (38) DIP$\diamondsuit$weigh
CIP (11)	N/A	N/A	CIP-std CIP-std+ CIP-MMD	CIP$\diamondsuit$ (51)
CIRM$^{(m)}$ (12)	N/A	CIRMweigh (49)	CIRM-std CIRM-std+ CIRM-MMD	CIRM$\diamondsuit^{(m)}$ (53) CIRM$\diamondsuit$weigh

Table 5: Summary of population DA methods introduced in this paper. The matching penalty variants of CIP and CIRM can be formulated similarly as those for DIP. They are omitted for the sake of space.

- **DIP$^{(m)}$-std-finite:** the finite-sample formulation of the DIP$^{(m)}$-std estimator (55). The difference between the source and target standard deviations is used as distributional distance,

$$\begin{aligned} \hat{f}^{(m)}_{\text{DIP-std}}(x) &:= x^\top \hat{\beta}^{(m)}_{\text{DIP-std}} + \hat{\beta}^{(m)}_{\text{DIP-std},0} \\ \hat{\beta}^{(m)}_{\text{DIP-std}}, \hat{\beta}^{(m)}_{\text{DIP-std},0} &:= \underset{\beta,\beta_0}{\arg\min}\ \frac{1}{n_m}\sum_{i=1}^{n_m}\left(y_i^{(m)} - x_i^{(m)\top}\beta - \beta_0\right)^2 + \lambda_{\text{match}}\left(\delta^{(m)} - \tilde{\delta}\right)^2, \end{aligned} \tag{55}$$

  where $\delta^{(m)}$ is the standard deviation of $\left(x_1^{(m)\top}\beta, x_2^{(m)\top}\beta, \ldots, x_{n_m}^{(m)\top}\beta\right)$, and $\tilde{\delta}$ is the standard deviation of $\left(\tilde{x}_1^\top\beta, \tilde{x}_2^\top\beta, \ldots, \tilde{x}_{\tilde{n}}^\top\beta\right)$.

- **DIP$^{(m)}$-std+-finite:** the finite-sample DIP estimator where the differences between the source and target means, standard deviations and 25% quantiles are used as distributional distance, $\mathcal{V}$ is linear and $\mathcal{U}$ is the singleton of identity mapping. For the

m-th source environment, it is defined as

$$\begin{aligned}\hat{f}^{(m)}_{\text{DIP-std+}}(x) &:= x^\top \hat{\beta}^{(m)}_{\text{DIP-std+}} + \hat{\beta}^{(m)}_{\text{DIP-std+},0} \\ \hat{\beta}^{(m)}_{\text{DIP-std+}}, \hat{\beta}^{(m)}_{\text{DIP-std+},0} &:= \underset{\beta,\beta_0}{\arg\min}\ \frac{1}{n_m}\sum_{i=1}^{n_m}\left(y_i^{(m)} - {x_i^{(m)}}^\top\beta - \beta_0\right)^2 + \lambda_{\text{match}}\left(\mu^{(m)} - \tilde{\mu}\right)^2 \\ &\qquad + \lambda_{\text{match}}\left(\delta^{(m)} - \tilde{\delta}\right)^2 + \lambda_{\text{match}}\left(\psi^{(m)}_{25\%} - \tilde{\psi}_{25\%}\right)^2,\end{aligned} \tag{56}$$

where $\mu^{(m)}, \delta^{(m)}, \psi^{(m)}_{25\%}$ are respectively the mean, standard deviation and 25% quantile of $\left({x_1^{(m)}}^\top\beta, {x_2^{(m)}}^\top\beta, \ldots, {x_{n_m}^{(m)}}^\top\beta\right)$, and $\tilde{\mu}, \tilde{\delta}, \tilde{\psi}_{25\%}$ are respectively the mean, standard deviation and 25% quantile of $\left(\tilde{x}_1^\top\beta, \tilde{x}_2^\top\beta, \ldots, \tilde{x}_{\tilde{n}}^\top\beta\right)$.

- **DIP$^{(m)}$-MMD-finite:** the finite-sample DIP estimator where mean squared difference is used as distributional distance, $\mathcal{V}$ is linear and $\mathcal{U}$ is the singleton of identity mapping. For the m-th source environment, it is defined as

$$\begin{aligned}\hat{f}^{(m)}_{\text{DIP-MMD}}(x) &:= x^\top \hat{\beta}^{(m)}_{\text{DIP-MMD}} + \hat{\beta}^{(m)}_{\text{DIP-MMD},0} \\ \hat{\beta}^{(m)}_{\text{DIP-MMD}}, \hat{\beta}^{(m)}_{\text{DIP-MMD},0} &:= \underset{\beta,\beta_0}{\arg\min}\ \frac{1}{n_m}\sum_{i=1}^{n_m}\left(y_i^{(m)} - {x_i^{(m)}}^\top\beta - \beta_0\right)^2 + \lambda_{\text{match}}\mathcal{D}_{\text{MMD},\mathcal{H}}\left(Z^{(m)}, \tilde{Z}\right),\end{aligned} \tag{57}$$

where $Z^{(m)}$ is the set of predicted responses in the m-th source environment $\left\{{x_1^{(m)}}^\top\beta, \ldots, {x_{n_m}}^\top\beta\right\}$, similarly $\tilde{Z} = \left\{\tilde{x}_1^\top\beta, \ldots, \tilde{x}_{\tilde{n}}^\top\beta\right\}$ and the maximum mean discrepancy (MMD) Gretton et al. (2012) with respect to the reproducing kernel Hilbert space (RHKS) $\mathcal{H}$ between these two sets is

$$\mathcal{D}_{\text{MMD},\mathcal{H}}\left(Z^{(m)}, \tilde{Z}\right) = \sup_{h\in\mathcal{H}}\left(\frac{1}{n_m}\sum_{i=1}^{n_m} h({x_i^{(m)}}^\top\beta) - \frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} h({\tilde{x}_i}^\top\beta)\right).$$

In all experiments, RKHS with Gaussian kernel is used.

Appendix B. Proofs related to Theorem 1

In this section, we prove Theorem 1 and related corollaries.

B.1 Proof of Theorem 1

At a high level, the proof of Theorem 1 goes by connecting the target population risk of DIP$^{(1)}$ with that of DIPOracle$^{(1)}$, and then bounding the target population risk of DIPOracle$^{(1)}$ as a regularized version of OLSTar.

Original (finite)	**Weighted variants**	**Matching penalty variants**	**Mixed variants**
DIP-mean (43)	DIPweigh-mean	DIP-std (55) DIP-std+ (56) DIP-MMD (57)	DIP$\diamondsuit^{(m)}$-mean DIP$\diamondsuit$weigh-mean
CIP-mean (44)	N/A	CIP-std CIP-std+ CIP-MMD	CIP$\diamondsuit$-mean
CIRM-mean (45)	CIRMweigh-mean	CIRM-std CIRM-std+ CIRM-MMD	CIRM$\diamondsuit^{(m)}$-mean CIRM$\diamondsuit$weigh-mean

Table 6: Summary of finite-sample formulations of the DA methods discussed in this paper. The suffixes "-finite" of the finite-sample DA methods are omitted in the table for brevity. The matching penalty variants of CIP and CIRM can be formulated similarly as those for DIP. The mixed-causal-anticausal variants of DIP, CIP and CIRM only requires two additional regressions. They are omitted for the sake of space.

Using the linear SCM assumption in Assumption 1, each data point in the source environment is generated i.i.d. from the following equation

$$\begin{aligned} X^{(1)} &= \mathbf{B}X^{(1)} + bY^{(1)} + a_X^{(1)} + \varepsilon_X^{(1)}, \\ Y^{(1)} &= \varepsilon_Y^{(1)}. \end{aligned} \tag{58}$$

Define $H = (\mathbb{I}_d - \mathbf{B})^{-1}$. Solving $X^{(1)}$ from Equation (58), we have

$$X^{(1)} = Hb\varepsilon_Y^{(1)} + Ha_X^{(1)} + H\varepsilon_X^{(1)}. \tag{59}$$

For $(\beta, \beta_0) \in \mathbb{R}^d \times \mathbb{R}$, the residual takes the following form

$$Y^{(1)} - \beta^\top X^{(1)} - \beta_0 = \left(1 - \beta^\top Hb\right)\varepsilon_Y^{(1)} - \left(\beta^\top Ha_X^{(1)} + \beta_0\right) - \beta^\top H\varepsilon_X^{(1)}.$$

Taking expectation and using the fact that noise has zero mean, we obtain

$$\mathbb{E}\left[Y^{(1)} - \beta^\top X^{(1)} - \beta_0\right]^2 = \left(1 - \beta^\top Hb\right)^2\sigma^2 + \left(\beta^\top Ha_X^{(1)} + \beta_0\right)^2 + \beta^\top H\Sigma H^\top\beta. \tag{60}$$

Similarly, we obtain the target expected residual

$$\mathbb{E}\left[\widetilde{Y} - \beta^\top \widetilde{X} - \beta_0\right]^2 = \left(1 - \beta^\top Hb\right)^2\sigma^2 + \left(\beta^\top H\widetilde{a}_X + \beta_0\right)^2 + \beta^\top H\Sigma H^\top\beta. \tag{61}$$

Risk of OLSTar: Using the expression for the target residual, the OLSTar estimator in Equation (5) becomes the solution of the following quadratic program

$$\min_{\beta,\beta_0} \left(1 - \beta^\top H b\right)^2 \sigma^2 + \left(\beta^\top H \widetilde{a}_X + \beta_0\right)^2 + \beta^\top H \Sigma H^\top \beta. \tag{62}$$

Solving the quadratic program and with matrix inversion lemma, we obtain

$$\begin{aligned}
\beta_{\text{OLSTar}} &= \sigma^2 \left(\mathbb{I}_d - \mathbf{B}\right)^\top \left(\Sigma + \sigma^2 b b^\top\right)^{-1} b \\
&= \left(\mathbb{I}_d - \mathbf{B}\right)^\top \frac{\sigma^2 \Sigma^{-1} b}{1 + \sigma^2 b^\top \Sigma^{-1} b} \\
\beta_{\text{OLSTar},0} &= -\beta_{\text{OLSTar}}^\top H \widetilde{a}_X.
\end{aligned}$$

The target population loss of OLSTar is

$$\begin{aligned}
\tilde{R}\left(f_{\text{OLSTar}}\right) &= \sigma^2 - \sigma^4 b^\top \left(\Sigma + \sigma^2 b b^\top\right)^{-1} b \\
&= \frac{\sigma^2}{1 + \sigma^2 b^\top \Sigma^{-1} b},
\end{aligned} \tag{63}$$

where the last equality follows from matrix inversion lemma.

Risk of OLSSrc$^{(1)}$: The minimization problem of OLSSrc can be similarly solved by changing the target variables to source ones in Equation (62). We obtain

$$\begin{aligned}
\beta_{\text{OLSSrc}}^{(1)} &= \sigma^2 \left(\mathbb{I}_d - \mathbf{B}\right)^\top \left(\Sigma + \sigma^2 b b^\top\right)^{-1} b \\
&= \left(\mathbb{I}_d - \mathbf{B}\right)^\top \frac{\sigma^2 \Sigma^{-1} b}{1 + \sigma^2 b^\top \Sigma^{-1} b} \\
\beta_{\text{OLSSrc},0}^{(1)} &= -{\beta_{\text{OLSSrc}}^{(1)}}^\top H a_X^{(1)}.
\end{aligned}$$

Because of the difference in the intercept term, the target population risk of OLSSrc has one additional term

$$\tilde{R}\left(f_{\text{OLSSrc}}^{(1)}\right) = \frac{\sigma^2}{1 + \sigma^2 b^\top \Sigma^{-1} b} + \frac{\left(\sigma^2 b^\top \Sigma^{-1} \left(a_X^{(1)} - \widetilde{a}_X\right)\right)^2}{\left(1 + \sigma^2 b^\top \Sigma^{-1} b\right)^2}. \tag{64}$$

Risk of DIP$^{(1)}$: Using Equation (59), for any $\beta \in \mathbb{R}^d$, we have

$$\beta^\top X^{(1)} = \beta^\top H b + \beta^\top H a_X^{(1)} + \beta^\top H \varepsilon_X^{(1)}.$$

Together with the similar equation on target, we obtain

$$\mathbb{E}\left[\beta^\top X^{(1)}\right] = \mathbb{E}\left[\beta^\top \widetilde{X}\right] + \beta^\top H \left(a_X^{(1)} - \widetilde{a}_X\right). \tag{65}$$

Combining Equation (60) and (65), we observe that the DIP$^{(1)}$ problem (10) becomes a constrained quadratic program

$$\begin{aligned}&\min_{\beta,\beta_0}\left(1-\beta^\top Hb\right)^2\sigma^2+\left(\beta^\top Ha_X^{(1)}+\beta_0\right)^2+\beta^\top H\Sigma H^\top\beta\\&\text{s.t. }\beta^\top H\left(a_X^{(1)}-\widetilde{a}\right)=0.\end{aligned}\tag{66}$$

Note that because of the constraint, this quadratic program is the same for the source data and for the target data. As a consequence, we have

$$\tilde{R}\left(f^{(1)}_{\text{DIP}}\right)=\tilde{R}\left(f^{(1)}_{\text{DIPOracle}}\right).$$

Now we solve the constrained quadratic program in Equation (66). First, we observe that the minimization on β_0 can be easily solved. Second, since H is invertible, we can re-parametrize the original problem and obtain the following problem

$$\begin{aligned}&\min_{\gamma}\left(1-\gamma^\top b\right)^2\sigma^2+\gamma^\top\Sigma\gamma\\&\text{s.t. }\gamma^\top\left(a_X^{(1)}-\widetilde{a}\right)=0.\end{aligned}\tag{67}$$

Define $u=\frac{a^{(1)}-\widetilde{a}}{\left\|a^{(1)}-\widetilde{a}\right\|_2}$. Using Gram-Schmidt orthogonalization, we can complete the vector u to form an orthonormal basis $(u,q_1,\ldots,q_{d-1})$. Let $Q_{\text{DIP}}\in\mathbb{R}^{d\times(d-1)}$ be the matrix formed with i-th column being q_i. Then the mapping

$$\begin{aligned}\mathbb{R}^{d-1}&\to\mathbb{R}^d\\\zeta&\mapsto Q_{\text{DIP}}\zeta\end{aligned}$$

constitute a bijection between $\mathbb{R}^{d-1}$ and the set $\left\{\gamma\in\mathbb{R}^d\mid\gamma^\top u=0\right\}$. This bijection allows us to transform the constrained quadratic program (67) to the following unconstrained one

$$\min_{\zeta\in\mathbb{R}^{d-1}}\left(1-\zeta^\top Q_{\text{DIP}}^\top b\right)^2\sigma^2+\zeta^\top Q_{\text{DIP}}^\top\Sigma Q_{\text{DIP}}\zeta.$$

Solving the quadratic program by setting gradient to zero, we obtain the minimizer

$$\zeta^*=\sigma^2\left(Q_{\text{DIP}}^\top\left(\sigma^2bb^\top+\Sigma\right)Q_{\text{DIP}}\right)^{-1}Q_{\text{DIP}}^\top b$$

and

$$\begin{aligned}\beta^{(1)}_{\text{DIP}}&=\sigma^2\left(\mathbb{I}_d-\mathbf{B}\right)^\top Q_{\text{DIP}}\left(Q_{\text{DIP}}^\top(\Sigma+\sigma^2bb^\top)Q_{\text{DIP}}\right)^{-1}Q_{\text{DIP}}^\top b\\&=\left(\mathbb{I}_d-\mathbf{B}\right)^\top\frac{\sigma^2Q_{\text{DIP}}\left(Q_{\text{DIP}}^\top\Sigma Q_{\text{DIP}}\right)^{-1}Q_{\text{DIP}}^\top b}{1+\sigma^2b^\top Q_{\text{DIP}}\left(Q_{\text{DIP}}^\top\Sigma Q_{\text{DIP}}\right)^{-1}Q_{\text{DIP}}^\top b}\\\beta^{(1)}_{\text{DIP},0}&=-{\beta^{(1)}_{\text{DIP}}}^\top Ha_X^{(1)},\\\beta^{(1)}_{\text{DIPOracle}}&=\beta^{(1)}_{\text{DIP}}\\\beta^{(1)}_{\text{DIPOracle},0}&=-{\beta^{(1)}_{\text{DIPOracle}}}^\top H\widetilde{a}_X.\end{aligned}$$

Consequently, the target population can be obtained by replacing b with $Q_{\text{DIP}}^\top b$ and Σ with $Q_{\text{DIP}}^\top \Sigma$ in Equation (63)

$$\tilde{R}\left(f_{\text{DIP}(1)}\right) = \tilde{R}\left(f_{\text{DIPOracle}(1)}\right) = \frac{\sigma^2}{1+\sigma^2 b^\top \Sigma^{-\frac{1}{2}} G_{\text{DIP}} \Sigma^{-\frac{1}{2}} b}.$$

where $G_{\text{DIP}} = \Sigma^{1/2} Q_{\text{DIP}} \left(Q_{\text{DIP}}^\top \Sigma Q_{\text{DIP}}\right)^{-1} Q_{\text{DIP}}^\top \Sigma^{1/2}$ is a projection matrix with rank $d-1$.

B.2 Proof of Corollary 2

Equation (19), (20) and (21) follow directly by plugging in $\Sigma = \frac{\sigma^2}{\rho}\mathbb{I}_d$ in the corresponding equations in Theorem 1. For the high probability bound, we use several tail inequalities. Let $a = a_X^{(1)} - \widetilde{a}_X$. Since a is generated randomly from $\mathcal{N}(0, \tau^2 \mathbb{I}_d)$, then for a fixed vector $v \in \mathbb{R}^d$, $a^\top v$ follows $\mathcal{N}(0, \tau^2 v^\top v)$. The standard Gaussian tail bound on $a^\top v$ gives, for $t > 0$,

$$\mathbb{P}\left(\left|\frac{a^\top v}{\tau \left(v^\top v\right)^{1/2}}\right| \geq t\right) \leq 2\exp\left(-t^2/2\right). \tag{68}$$

Since $\|a\|_2^2$ follows Chi-square distribution with d-degree of freedom, the standard chi-square tail bound (see e.g. Laurent and Massart (2000)) gives, for $t > 0$,

$$\mathbb{P}\left(\frac{\|a\|_2^2}{\tau^2 d} \leq 1 - \frac{t}{\sqrt{d}}\right) \leq \exp\left(-t^2/8\right). \tag{69}$$

Combining Equation (68) and (69), with probability at least $1 - \exp(-t^2/8) - 2\exp(-t^2/2)$, we have

$$\frac{\left(a^\top v\right)^2}{\|a\|_2^2} \leq \frac{t^2}{d - \sqrt{d}t}\left(v^\top v\right).$$

For t constant satisfying $0 < t \leq \frac{\sqrt{d}}{2}$, we have

$$\frac{t^2}{d - \sqrt{d}t} \leq \frac{2t^2}{d}.$$

Plugging the above high probability bound into Equation (20) and (21) with $v = b$, with probability at least $1 - \exp(-t^2/8) - 2\exp(-t^2/2)$, we have

$$\tilde{R}\left(f_{\text{OLSSrc}}^{(1)}\right) \leq \frac{\sigma^2}{1+\rho\|b\|_2^2} + \frac{\tau^2 t^2 \|b\|_2^2}{\left(1+\rho\|b\|_2^2\right)^2}, \text{ and}$$

$$\tilde{R}\left(f_{\text{DIP}}^{(1)}\right) \leq \frac{\sigma^2}{1+\rho\left(1-\frac{2t^2}{d}\right)\|b\|_2^2}.$$

We conclude and obtain the form needed in the corollary by a change of variable from $2t^2$ to t.

B.3 Proof of Corollary 3

Corollary 3 shows that the source population risk of $\text{DIP}^{(1)}$ is the same as target population risk of $\text{DIP}^{(1)}$. For this, it suffices to observe that the source expected residual (60) and the target expected residual (61) only differ by the term $\beta^\top H a_X^{(1)}$ and the term $\beta^\top H \widetilde{a}_X$. Since the DIP constraint enforces $\beta^\top H \left(a_X^{(1)} - \widetilde{a}_X\right) = 0$ as shown in Equation (66), we obtain that

$$R^{(1)}\left(f_{\text{DIP}}^{(1)}\right) = \tilde{R}\left(f_{\text{DIP}}^{(1)}\right). \tag{70}$$

B.4 Proof of Corollary 4

Using the linear SEM assumption in Assumption 1 and the additional intervention on Y, each data point in the source environment is generated i.i.d. from the following equation

$$\begin{aligned} X^{(1)} &= \mathbf{B}X^{(1)} + bY^{(1)} + a_X^{(1)} + \varepsilon_X^{(1)}, \\ Y^{(1)} &= a_Y^{(1)} + \varepsilon_Y^{(1)}. \end{aligned} \tag{71}$$

For $(\beta, \beta_0) \in \mathbb{R}^d \times \mathbb{R}$, the source residual takes the following form

$$Y^{(1)} - \beta^\top X^{(1)} - \beta_0 = \left(1 - \beta^\top H b\right)\varepsilon_Y^{(1)} + \left(\left(1 - \beta^\top H b\right) a_Y^{(1)} - \beta^\top H a_X^{(1)} - \beta_0\right) - \beta^\top H \varepsilon_X^{(1)}.$$

The $\text{DIP}^{(1)}$ problem (10) becomes a constrained quadratic program

$$\begin{aligned} &\min_{\beta,\beta_0} \left(1 - \beta^\top H b\right)^2 \sigma^2 + \left(\left(1 - \beta^\top H b\right) a_Y^{(1)} - \beta^\top H a_X^{(1)} - \beta_0\right)^2 + \beta^\top H \Sigma H^\top \beta \\ &\text{s.t. } \beta^\top H \left(a_X^{(1)} + a_Y^{(1)} b\right) = \beta^\top H \left(\widetilde{a}_X + \widetilde{a}_Y b\right). \end{aligned} \tag{72}$$

Note that because of the intervention on Y, unlike in Theorem 1, this quadratic program is no longer the same for $\text{DIP}^{(1)}$ and $\text{DIPOracle}^{(1)}$. However, the constrained quadratic program (72) can be solved similarly as we did in Appendix B.1 around Equation (66) by introducing

$$u = \frac{a_X^{(1)} + a_Y^{(1)} b - \widetilde{a}_X - \widetilde{a}_Y b}{\left\| a_X^{(1)} + a_Y^{(1)} b - \widetilde{a}_X - \widetilde{a}_Y b \right\|_2}$$

and $Q_2 \in \mathbb{R}^{d \times d-1}$ is the matrix with columns formed by the vectors that complete the vector u to an orthonormal basis of $\mathbb{R}^d$. Following the rest of the proof in Appendix B.1, we obtain

$$\begin{aligned} \beta_{\text{DIP}}^{(1)} &= \left(\mathbb{I}_d - \mathbf{B}\right)^\top \frac{\sigma^2 Q_2 \left(Q_2^\top \Sigma Q_2\right)^{-1} Q_2^\top b}{1 + b^\top Q_2 \left(Q_2^\top \Sigma Q_2\right)^{-1} Q_2^\top b} \\ \beta_{\text{DIP},0}^{(1)} &= \left(1 - {\beta_{\text{DIP}}^{(1)}}^\top H b\right) a_Y^{(1)} - {\beta_{\text{DIP}}^{(1)}}^\top H a_X^{(1)}. \end{aligned} \tag{73}$$

Note that the intercept $\beta^{(1)}_{\mathrm{DIP},0}$ has an extra term due to the intervention on $a^{(1)}_Y$. For $(\beta, \beta_0) \in \mathbb{R}^d \times \mathbb{R}$, the target residual takes the following form

$$\widetilde{Y} - \beta^\top \widetilde{X} - \beta_0 = \left(1 - \beta^\top H b\right) \widetilde{\varepsilon}_Y + \left(\left(1 - \beta^\top H b\right) \widetilde{a}_Y - \beta^\top H \widetilde{a}_X - \beta_0\right) - \beta^\top H \widetilde{\varepsilon}_X.$$

Plugging $(\beta^{(1)}_{\mathrm{DIP}}, \beta^{(1)}_{\mathrm{DIP},0})$ of Equation (73) into the above residual, we obtain that the population target risk which has an extra term that depends on $a^{(1)}_Y - \widetilde{a}_Y$

$$\tilde{R}\left(f_{\mathrm{DIP}(1)}\right) = \frac{\sigma^2}{1 + \sigma^2 b^\top \Sigma^{-\frac{1}{2}} G_2 \Sigma^{-\frac{1}{2}} b} + \left(a^{(1)}_Y - \widetilde{a}_Y\right)^2,$$

where $G_2 = \Sigma^{1/2} Q_2 \left(Q_2^\top \Sigma Q_2\right)^{-1} Q_2^\top \Sigma^{1/2}$ is a projection matrix with rank $d - 1$.

Appendix C. Proofs related to Theorem 5

In this section, we prove Theorem 5 and related corollaries.

C.1 Proof of Theorem 5

Using the linear SEM assumption in Assumption 2, each data point in the m-th source environment is generated i.i.d. from the following equation

$$\begin{aligned} X^{(m)} &= \mathbf{B} X^{(m)} + b Y^{(m)} + a^{(m)}_X + \varepsilon^{(m)}_X, \\ Y^{(m)} &= a^{(m)}_Y + \varepsilon^{(m)}_Y. \end{aligned} \tag{74}$$

Define $H = (\mathbb{I}_d - \mathbf{B})^{-1}$. For $(\beta, \beta_0) \in \mathbb{R}^d \times \mathbb{R}$, the residual takes the following form

$$Y^{(m)} - \beta^\top X^{(m)} - \beta_0 = \left(1 - \beta^\top H b\right) \left(a^{(m)}_Y + \varepsilon^{(m)}_Y\right) - \left(\beta^\top H a^{(m)}_X + \beta_0\right) - \beta^\top H \varepsilon^{(m)}_X. \tag{75}$$

The target residual has a similar form

$$\widetilde{Y} - \beta^\top \widetilde{X} - \beta_0 = \left(1 - \beta^\top H b\right) \left(\widetilde{a}_Y + \widetilde{\varepsilon}_Y\right) - \left(\beta^\top H \widetilde{a}_X + \beta_0\right) - \beta^\top H \widetilde{\varepsilon}_X. \tag{76}$$

Risk of OLSTar: Using the expression for the target residual, the OLSTar estimator in Equation (5) becomes the solution of the following quadratic program

$$\min_{\beta, \beta_0} \left(1 - \beta^\top H b\right)^2 \sigma^2 + \frac{1}{M} \sum_{m=1}^{M} \left(\left(1 - \beta^\top H b\right) \widetilde{a}_Y - \beta^\top H \widetilde{a}_X - \beta_0\right)^2 + \beta^\top H \Sigma H^\top \beta.$$

Solving the quadratic program by setting the gradient to zero, we obtain

$$\begin{aligned} \beta_{\mathrm{OLSTar}} &= \sigma^2 \left(\mathbb{I}_d - \mathbf{B}\right)^\top \left(\Sigma + \sigma^2 b b^\top\right)^{-1} b \\ \beta_{\mathrm{OLSTar},0} &= \left(1 - \beta^\top_{\mathrm{OLSTar}} H b\right) \widetilde{a}_Y - \beta^\top_{\mathrm{OLSTar}} H \widetilde{a}_X. \end{aligned}$$

Despite the difference in the intercept term, the corresponding target risk is the same as in Theorem 1

$$\tilde{R}\left(f_{\mathrm{OLSTar}}\right) = \frac{\sigma^2}{1 + \sigma^2 b^\top \Sigma^{-1} b}.$$

Risk of CIP: Using the SEM (74), the constraint of CIP in Equation (11) becomes

$$\beta^\top H a_X^{(m)} = \beta^\top H a_X^{(1)}, \quad \forall m \in \{2, \dots, M\}.$$

Together with the residual expression in Equation (75), the CIP objective (11) is equivalent to

$$\begin{aligned} &\min_{\beta,\beta_0} \left(1 - \beta^\top H b\right)^2 \sigma^2 + \frac{1}{M} \sum_{m=1}^{M} \left(\left(1 - \beta^\top H b\right) a_Y^{(m)} - \beta^\top H a_X^{(m)} - \beta_0\right)^2 + \beta^\top H \Sigma H^\top \beta \\ &\text{s.t. } \beta^\top H \left(a_X^{(m)} - a_X^{(1)}\right) = 0, \quad \forall m \in \{2, \dots, M\}. \end{aligned} \tag{77}$$

First, the one-dimensional quadratic program on β_0 can be solved easily by setting derivative to zero and we obtain

$$\beta_0^* = \frac{1}{M} \sum_{m=1}^{M} \left(1 - \beta^{*\top} H b\right) a_Y^{(m)} - \beta^{*\top} H a_X^{(m)}.$$

Plugging the expression of β_0 back to Equation (77), the minimization on β becomes

$$\begin{aligned} &\min_{\beta} \left(1 - \beta^\top H b\right)^2 \left(\sigma^2 + \Delta_Y\right) + \beta^\top H \Sigma H^\top \beta \\ &\text{s.t. } \beta^\top H \left(a_X^{(m)} - a_X^{(1)}\right) = 0, \ \forall m \in \{2, \dots, M\}, \end{aligned}$$

where

$$\Delta_Y = \frac{1}{M} \sum_{m=1}^{M} \left(a_Y^{(m)} - \bar{a}_Y\right)^2 \text{ and } \bar{a}_Y = \frac{1}{M} \sum_{k=1}^{M} a_Y^{(k)}. \tag{78}$$

Second, since H is invertible, we can re-parametrize the optimization on β as follows

$$\begin{aligned} &\min_{\gamma \in \mathbb{R}^d} \left(1 - \gamma^\top b\right)^2 \left(\sigma^2 + \Delta_Y)\right) + \gamma^\top \Sigma \gamma \\ &\text{s.t. } \gamma^\top \left(a_X^{(m)} - a_X^{(1)}\right) = 0, \ \forall m \in \{2, \dots, M\}. \end{aligned} \tag{79}$$

Let $P \in \mathbb{R}^{d \times p}$ be the matrix formed with an orthonormal basis of the p-dimensional subspace span $\left(a^{(2)} - a^{(1)}, \dots, a^{(m)} - a^{(1)}\right)$. Let $Q_{\text{CIP}} \in \mathbb{R}^{d \times (d-p)}$ be the matrix with columns formed by completing the columns of P to a basis of $\mathbb{R}^d$ via Gram-Schmidt orthogonalization. The following mapping

$$\begin{aligned} \mathbb{R}^{d-p} &\to \mathbb{R}^d \\ \zeta &\mapsto Q_{\text{CIP}} \zeta \end{aligned}$$

constitutes a bijection between $\mathbb{R}^{d-p}$ and the set $\left\{\gamma \in \mathbb{R}^d \mid P^\top \gamma = 0\right\}$. With the change of variable, the constrained optimization in Equation (79) is equivalent to the unconstrained one

$$\min_{\zeta \in \mathbb{R}^{d-p}} \left(1 - \zeta^\top Q_{\text{CIP}}^\top b\right)^2 \left(\sigma^2 + \Delta_Y\right) + \zeta^\top Q_{\text{CIP}}^\top \Sigma Q_{\text{CIP}} \zeta.$$

Solving the unconstrained quadratic program by setting gradient to zero, we obtain the minimizer

$$\begin{aligned}\zeta^* &= \left(\sigma^2+\Delta_Y\right)\left(Q_{\text{CIP}}^\top\left(\left(\sigma^2+\Delta_Y\right)bb^\top+\Sigma\right)Q_{\text{CIP}}\right)^{-1}Q_{\text{CIP}}^\top b\\ &= \left(\sigma^2+\Delta_Y\right)\frac{Q_{\text{CIP}}\left({Q_{\text{CIP}}}^\top\Sigma Q_{\text{CIP}}\right)^{-1}Q_{\text{CIP}}^\top b}{1+\left(\sigma^2+\Delta_Y\right)b^\top Q_{\text{CIP}}\left({Q_{\text{CIP}}}^\top\Sigma Q_{\text{CIP}}\right)^{-1}Q_{\text{CIP}}^\top b},\end{aligned}$$

where the last equality uses the matrix inversion lemma. Finally, transforming the variables back, the CIP estimator is

$$\begin{aligned}\beta_{\text{CIP}} &= \left(\sigma^2+\Delta_Y\right)\frac{\left(\mathbb{I}_d-\mathbf{B}\right)^\top Q_{\text{CIP}}\left({Q_{\text{CIP}}}^\top\Sigma Q_{\text{CIP}}\right)^{-1}Q_{\text{CIP}}^\top b}{1+\left(\sigma^2+\Delta_Y\right)b^\top Q_{\text{CIP}}\left({Q_{\text{CIP}}}^\top\Sigma Q_{\text{CIP}}\right)^{-1}Q_{\text{CIP}}^\top b}\\ \beta_{\text{CIP},0} &= \left(1-{\beta_{\text{CIP}}}^\top Hb\right)\bar{a}_Y-{\beta_{\text{CIP}}}^\top Ha_X^{(m)}.\end{aligned}\tag{80}$$

According to the form of the target residual (76), the target population risk for (β,β_0) takes the following form

$$\left(1-\beta^\top Hb\right)^2\sigma^2+\left(\left(1-\beta^\top Hb\right)\widetilde{a}_Y-\beta^\top H\widetilde{a}_X-\beta_0\right)^2+\beta^\top H\Sigma H^\top\beta.$$

Plugging in the CIP estimator into the equation above, we obtain the target population risk of CIP

$$\tilde{R}\left(f_{\text{CIP}}\right)=\frac{\sigma^2+\Delta_Y}{1+\left(\sigma^2+\Delta_Y\right)b^\top\Sigma^{-\frac{1}{2}}G_{\text{CIP}}\Sigma^{-\frac{1}{2}}b}+\frac{\left(\widetilde{a}_Y-\bar{a}_Y\right)^2-\Delta_Y}{\left(1+\left(\sigma^2+\Delta_Y\right)b^\top\Sigma^{-\frac{1}{2}}G_{\text{CIP}}\Sigma^{-\frac{1}{2}}b\right)^2},$$

where $G_{\text{CIP}}=\Sigma^{1/2}Q_{\text{CIP}}\left(Q_{\text{CIP}}^\top\Sigma Q_{\text{CIP}}\right)^{-1}Q_{\text{CIP}}^\top\Sigma^{1/2}$ is a projection matrix with rank $d-p$.

Risk of CIRM: First we derive ϑ_{CIRM} in Equation (13). From the CIP expression in Equation (80) and the SEM (74), we obtain the following expression for ${X^{(m)}}^\top\beta_{\text{CIP}}$

$$\frac{\left(\sigma^2+\Delta_Y\right)\left(\varepsilon_Y^{(m)}b^\top\Sigma^{-1/2}G_{\text{CIP}}\Sigma^{-1/2}b+{\varepsilon_X^{(m)}}^\top\Sigma^{-1/2}G_{\text{CIP}}\Sigma^{-1/2}b+{a_X^{(m)}}^\top\Sigma^{-1/2}G_{\text{CIP}}\Sigma^{-1/2}b\right)}{1+\left(\sigma^2+\Delta_Y\right)b^\top\Sigma^{-1/2}G_{\text{CIP}}\Sigma^{-1/2}b}.$$

The deviation from its expectation is

$$\begin{aligned}&{X^{(m)}}^\top\beta_{\text{CIP}}-\mathbb{E}\left[{X^{(m)}}^\top\beta_{\text{CIP}}\right]\\ &=\frac{\left(\sigma^2+\Delta_Y\right)b^\top\Sigma^{-1/2}G_{\text{CIP}}\Sigma^{-1/2}b}{1+\left(\sigma^2+\Delta_Y\right)b^\top\Sigma^{-1/2}G_{\text{CIP}}\Sigma^{-1/2}b}\cdot\varepsilon_Y^{(m)}+\frac{\left(\sigma^2+\Delta_Y\right){\varepsilon_X^{(m)}}^\top\Sigma^{-1/2}G_{\text{CIP}}\Sigma^{-1/2}b}{1+\left(\sigma^2+\Delta_Y\right)b^\top\Sigma^{-1/2}G_{\text{CIP}}\Sigma^{-1/2}b}.\end{aligned}$$

Plugging the above equation into Equation (13), together with

$$X^{(m)}=HbY^{(m)}+Ha_X^{(m)}+H\varepsilon_X^{(m)}\tag{81}$$

and $Y^{(m)} - \mathbb{E}\left[Y^{(m)}\right] = \varepsilon_Y^{(m)}$, we obtain

$$\vartheta_{\text{CIRM}} = \left(\frac{1 + \left(\sigma^2 + \Delta_Y\right) b^\top \Sigma^{-1/2} G_{\text{CIP}} \Sigma^{-1/2} b}{\left(\sigma^2 + \Delta_Y\right) b^\top \Sigma^{-1/2} G_{\text{CIP}} \Sigma^{-1/2} b} \right) Hb. \tag{82}$$

Note that the vector ϑ_{CIRM} is co-linear with the Y component in Equation (81). We remark that the idea of CIRM is to use $\widetilde{X}^\top \beta_{\text{CIP}}$ as a proxy for the unobserved $\widetilde{Y}$ to remove the $\widetilde{Y}$ part in the covariates so that the DIP matching based ideas still can be applied.

With the ϑ_{CIRM} expression (82) and the β_{CIP} expression (80), the LHS of the CIRM constraint (12) becomes

$$\begin{aligned}
&X^{(m)} - \left({X^{(m)}}^\top \beta_{\text{CIP}} \right) \vartheta_{\text{CIRM}} \\
&= \left(Y^{(m)} - \left({X^{(m)}}^\top \beta_{\text{CIP}} \right) \cdot \frac{1 + \left(\sigma^2 + \Delta_Y\right) b^\top \Sigma^{-1/2} G_{\text{CIP}} \Sigma^{-1/2} b}{\left(\sigma^2 + \Delta_Y\right)) b^\top \Sigma^{-1/2} G_{\text{CIP}} \Sigma^{-1/2} b} \right) Hb + H a_X^{(m)} + H \varepsilon_X^{(m)} \\
&= \left(\beta_{\text{CIP}}^\top H a_X^{(m)} + \beta_{\text{CIP}}^\top H \varepsilon_X^{(m)} \right) Hb + H a_X^{(m)} + H \varepsilon_X^{(m)} \\
&\overset{(i)}{=} \left(\beta_{\text{CIP}}^\top H a_X^{(1)} + \beta_{\text{CIP}}^\top H \varepsilon_X^{(m)} \right) Hb + H a_X^{(m)} + H \varepsilon_X^{(m)},
\end{aligned} \tag{83}$$

where the last inequality (i) follows from the fact that the constraint in CIP (11) forces

$$ {\beta_{\text{CIP}}}^\top H a_X^{(m)} = {\beta_{\text{CIP}}}^\top H a_X^{(1)}, \quad \forall m \in \{1, \ldots, M\}. $$

An expression similar to Equation (83) can be obtained for the target environments by taking into account that $\widetilde{a}_X \in \text{span}\left(a_X^{(1)}, \ldots, a_X^{(M)}\right)$,

$$\widetilde{X} - \left(\widetilde{X}^\top \beta_{\text{CIP}} \right) \vartheta_{\text{CIRM}} = Hb \left(\beta_{\text{CIP}}^\top H a_X^{(1)} + \beta_{\text{CIP}}^\top H \widetilde{\varepsilon}_X \right) + H \widetilde{a}_X + H \widetilde{\varepsilon}_X. \tag{84}$$

Multiply Equation (83) and (84) by β and take expectation, we obtain a simplified form of the CIRM constraint (12) as follows

$$\beta^\top H a_X^{(m)} = \beta^\top H \widetilde{a}_X.$$

Note that the CIRM constraint is effectively matching the covariate interventions as DIP constraint did when Y is not intervened on. As a consequence, given β_{CIP} and ϑ_{CIRM}, the CIRM$^{(m)}$ estimator (12) is very similar to DIP$^{(m)}$ (66) except that the intervention on Y still appears in the residual. The CIRM$^{(m)}$ estimator (12) can be written as follows

$$\begin{aligned}
&\min_{\beta, \beta_0} \left(1 - \beta^\top Hb\right)^2 \sigma^2 + \left(\left(1 - \beta^\top Hb\right) a_Y^{(m)} - \beta^\top H a_X^{(m)} - \beta_0 \right)^2 + \beta^\top H \Sigma H^\top \beta \\
&\text{s.t. } \beta^\top H \left(a_X^{(m)} - \widetilde{a}_X \right) = 0.
\end{aligned} \tag{85}$$

After solving for β_0, the quadratic program for β is exactly the same as in the DIP proof in Appendix B.1 around Equation (66). Thus we obtain

$$\beta_{\text{CIRM}}^{(m)} = (\mathbb{I}_d - \mathbf{B})^\top \frac{\sigma^2 Q_{\text{DIP}}^{(m)} \left({Q_{\text{DIP}}^{(m)}}^\top \Sigma Q_{\text{DIP}}^{(m)} \right)^{-1} {Q_{\text{DIP}}^{(m)}}^\top b}{1 + \sigma^2 b^\top Q_{\text{DIP}}^{(m)} \left({Q_{\text{DIP}}^{(m)}}^\top \Sigma Q_{\text{DIP}}^{(m)} \right)^{-1} {Q_{\text{DIP}}^{(m)}}^\top b}$$

$$\beta_{\text{CIRM},0}^{(m)} = \left(1 - {\beta_{\text{CIRM}}^{(m)}}^\top H b\right) a_Y^{(m)} - {\beta_{\text{CIRM}}^{(m)}}^\top H a_X^{(m)}, \tag{86}$$

the corresponding CIRM target population risk is

$$\tilde{R}\left(f_{\text{CIRM}(m)}\right) = \frac{\sigma^2}{1 + \sigma^2 b^\top \Sigma^{-\frac{1}{2}} G_{\text{DIP}}^{(m)} \Sigma^{-\frac{1}{2}} b} + \frac{\left(a_Y^{(m)} - \widetilde{a}_Y\right)^2}{\left(1 + \sigma^2 b^\top \Sigma^{-\frac{1}{2}} G_{\text{DIP}}^{(m)} \Sigma^{-\frac{1}{2}} b\right)^2},$$

where $Q_{\text{DIP}}^{(m)}$ and $G_{\text{DIP}}^{(m)}$ are in the same way as $Q_{\text{DIP}}^{(1)}$ and $G_{\text{DIP}}^{(1)}$ in Theorem 1.

$G_{\text{DIP}}^{(m)} = \Sigma^{1/2} Q_{\text{DIP}}^{(m)} \left({Q_{\text{DIP}}^{(m)}}^\top \Sigma Q_{\text{DIP}} m \right)^{-1} {Q_{\text{DIP}}^{(m)}}^\top \Sigma^{1/2}$ is a projection matrix. $Q_{\text{DIP}}^{(m)} \in \mathbb{R}^{d \times d-1}$ is the matrix with columns formed by the vectors that complete the vector u to an orthonormal basis where

$$u^{(m)} = \begin{cases} \frac{a^{(m)} - \widetilde{a}}{\left\| a^{(m)} - \widetilde{a} \right\|_2}, & \text{if } a^{(m)} \neq \widetilde{a} \\ 0, & \text{otherwise.} \end{cases}$$

C.2 Proof of Corollary 6

Equation (31), (32) and (33) follow directly by plugging in $\Sigma = \frac{\sigma^2}{\rho} \mathbb{I}_d$ and intervention on Y equals to zero in the corresponding formula in Theorem 5. Recall that $P_{\text{CIP}} \in \mathbb{R}^{d \times p}$ is the matrix with columns formed with the orthonormal basis of span $\left(a_X^{(2)} - a_X^{(1)}, \ldots, a_X^{(M)} - a_X^{(1)}\right)$. Let $A \in \mathbb{R}^{d \times (M-1)}$ with $(m-1)$-th column $a_X^{(m)} - a_X^{(1)}$. According the assumption that $a_X^{(2)} - a_X^{(1)}, \ldots a_X^{(M)} - a_X^{(1)}$ are generated independently from the standard Gaussian distribution, $\frac{1}{d} A^\top A$ follows a multivariate Wishart distribution and its eigenvalue tail bound is well known (see e.g. Chapter 6 in Wainwright (2019)). We have

$$\mathbb{P}\left(\lambda_{\min}\left(A^\top A\right) \leq d \left((1-\delta) - \sqrt{\frac{M-1}{d}} \right)^2 \right) \leq \exp\left(-d\delta^2/2\right).$$

$$\mathbb{P}\left(\lambda_{\max}\left(A^\top A\right) \geq d \left((1+\delta) + \sqrt{\frac{M-1}{d}} \right)^2 \right) \leq \exp\left(-d\delta^2/2\right).$$

It implies with probability at least $1 - 2\exp(-d\delta^2/2)$, for $\delta < \frac{1}{2} - \sqrt{\frac{M-1}{d}}$, we have

$$\lambda_{\min}\left(A^\top A\right) > \frac{d}{2} \text{ and } \lambda_{\max}\left(A^\top A\right) < \frac{3d}{2}.$$

This also implies that with the same probability $A^\top A$ is full rank and $p = M - 1$.

For a fixed vector $v \in \mathbb{R}^d$, we have

$$\left\|P_{\text{CIP}}^\top v\right\|_2^2 = v^\top A\left(A^\top A\right)^{-1} A^\top v.$$

Note that

$$A^\top v = \sum_{m=2}^{M}\left(a_X^{(m)} - a_X^{(1)}\right)^\top v$$

is a sum of $(M-1)$ i.i.d. Gaussian random variables with mean 0 and variance $v^\top v$. Using the standard chi-square tail bound, we have

$$\mathbb{P}\left(\left\|A^\top v\right\|_2^2 \geq (M-1)v^\top v\left(1 + \frac{t}{\sqrt{M-1}}\right)\right) \leq \exp\left(-t^2/8\right).$$

$$\mathbb{P}\left(\left\|A^\top v\right\|_2^2 \leq (M-1)v^\top v\left(1 - \frac{t}{\sqrt{M-1}}\right)\right) \leq \exp\left(-t^2/8\right).$$

Combining the high probability bounds above, we have, with probability at least $1 - 2\exp\left(-d\delta^2/2\right) - 2\exp\left(-t^2/8\right)$,

$$\frac{(M-1)}{3d}v^\top v \leq \left\|P_{\text{CIP}}^\top v\right\|_2^2 \leq \frac{3(M-1)}{d}v^\top v,$$

given that $\delta < \frac{1}{2} - \sqrt{\frac{M-1}{d}}$ and $t \leq \frac{\sqrt{M-1}}{2}$. Note that it is possible to ensure $1 - 2\exp\left(-d\delta^2/2\right) - 2\exp\left(-t^2/8\right)$ to be close to 1 when d and M are both large and $M \ll d$. Taking $t = \frac{\sqrt{M-1}}{2}$ and $d = 6M$, this probability is larger than $1 - 2\exp(-d/32) - 2\exp(-M/32)$.

The high probability bound for $\left(u^\top b\right)^2$ can be obtained similarly as in the proof of Corollary 2. According to the proof of Corollary 2 in Appendix B.2, for $0 < t \leq d/2$, with probability at least $1 - \exp(-t/16) - 2\exp(-t/4)$, we have

$$\left(u^\top b\right)^2 \leq \frac{t}{d}\left\|b\right\|_2^2.$$

Putting the two high probability bound together, we conclude Corollary 6.

C.3 Proof of Corollary 8

This corollary follows from the proof of Theorem 5 in Appendix C.1. According to Equation (85), the m-th source population risk can be written as

$$\left(1-\beta^\top Hb\right)^2\sigma^2+\left(\left(1-\beta^\top Hb\right)a_Y^{(m)}-\beta^\top Ha_X^{(m)}-\beta_0\right)^2+\beta^\top H\Sigma H^\top\beta.$$

Similarly, the target population risk can be written as

$$\left(1-\beta^\top Hb\right)^2\sigma^2+\left(\left(1-\beta^\top Hb\right)\widetilde{a}_Y-\beta^\top H\widetilde{a}_X-\beta_0\right)^2+\beta^\top H\Sigma H^\top\beta.$$

The CIRM constraint ensures that $\beta^\top H\left(a_X^{(m)}-\widetilde{a}_X\right)=0$ according to Equation (85). Hence the only difference between the source population risk and target population risk lies in the $a_Y^{(m)}$ and $\widetilde{a}_Y$ dependent terms. Plugging the CIRM solution (86) into the source and target population risks, we obtain

$$\begin{aligned}R^{(m)}\left(f_{\text{CIRM}}^{(m)}\right)&=\tilde{R}\left(f_{\text{CIRM}}^{(m)}\right)+\left(1-{\beta_{\text{CIRM}}^{(m)}}^\top Hb\right)^2\left(a_Y^{(m)}-\widetilde{a}_Y\right)^2\\&=\tilde{R}\left(f_{\text{CIRM}}^{(m)}\right)+\frac{\left(a_Y^{(m)}-\widetilde{a}_Y\right)^2}{\left(1+\sigma^2b^\top\Sigma^{-\frac{1}{2}}G_{\text{DIP}}^{(m)}\Sigma^{-\frac{1}{2}}b\right)^2}.\end{aligned}$$

C.4 Proof of Corollary 7

Risk of DIP$\diamondsuit$: Since $X_{\text{P}}^{(m)}$ is uncorrelated with $\varepsilon_Y^{(m)}$, regressing $Y^{(m)}$ on $X_{\text{P}}^{(m)}$ gives back the coefficients in the data generation SCM. Solving Equation (39), we obtain

$$\gamma^{(m)}=\omega_{\text{P}}.$$

Similarly, since $X_{\text{P}}^{(m)}$ is uncorrelated with $X_{\text{D}}^{(m)}$, regressing of $X_{\text{D}}^{(m)}$ on $X_{\text{P}}^{(m)}$ gives back the coefficients in the data generation SCM. Solving Equation (40), we obtain

$${\Gamma^{(m)}}^\top=\left(\mathbb{I}_{d-r}-\mathbf{B}_{\text{D}}\right)^{-1}b_{\text{D}}\omega_{\text{P}}^\top+\left(\mathbb{I}_{d-r}-\mathbf{B}_{\text{D}}\right)^{-1}\mathbf{B}_{\text{d-p}}.$$

Use the definition of intermediate random variables in Equation (37), we observe that these variables satisfy the anticausal data generation Assumption 1

$$\begin{aligned}\begin{bmatrix}X_{\mathcal{I}}^{(m)}\\Y_{\mathcal{I}}^{(m)}\end{bmatrix}&=\begin{bmatrix}\mathbf{B}_{\text{D}}&b_{\text{D}}\\0&0\end{bmatrix}\begin{bmatrix}X_{\mathcal{I}}^{(m)}\\Y_{\mathcal{I}}^{(m)}\end{bmatrix}+\begin{bmatrix}a_{X,\text{D}}^{(m)}\\a_Y^{(m)}\end{bmatrix}+\begin{bmatrix}\varepsilon_{X,\text{D}}^{(m)}\\\varepsilon_Y^{(m)}\end{bmatrix}\\\begin{bmatrix}\widetilde{X}_{\mathcal{I}}\\\widetilde{Y}_{\mathcal{I}}\end{bmatrix}&=\begin{bmatrix}\mathbf{B}_{\text{D}}&b_{\text{D}}\\0&0\end{bmatrix}\begin{bmatrix}\widetilde{X}_{\mathcal{I}}\\\widetilde{Y}_{\mathcal{I}}\end{bmatrix}+\begin{bmatrix}\widetilde{a}_{X,\text{D}}\\\widetilde{a}_Y\end{bmatrix}+\begin{bmatrix}\widetilde{\varepsilon}_{X,\text{D}}\\\widetilde{\varepsilon}_Y\end{bmatrix}.\end{aligned}$$

Together with the assumption of no intervention on Y, we can apply Theorem 1 on the intermediate random variables to obtain the target risk of DIP$\diamondsuit$.

Risk of OLSTar: The data generation Assumption 3 gives the following equations

$$\begin{aligned}\widetilde{X}_{\mathrm{P}} &= H_{\mathrm{P}}\widetilde{a}_{X,\mathrm{P}} + H_{\mathrm{P}}\widetilde{\varepsilon}_{X,\mathrm{P}}\\ \widetilde{X}_{\mathrm{D}} &= H_{\mathrm{D}}b_{\mathrm{D}}Y + H_{\mathrm{D}}\mathbf{B}_{\text{d-p}}\widetilde{X}_{\mathrm{P}} + H_{\mathrm{D}}\widetilde{a}_{X,\mathrm{D}} + H_{\mathrm{D}}\widetilde{\varepsilon}_{X,\mathrm{D}}\\ \widetilde{Y} &= \omega_{\mathrm{P}}^{\top}\widetilde{X}_{\mathrm{P}} + \widetilde{\varepsilon}_{Y},\end{aligned}$$

where $H_{\mathrm{P}} = (\mathbb{I}_r - \mathbf{B}_{\mathrm{P}})^{-1}$, $H_{\mathrm{D}} = (\mathbb{I}_{d-r} - \mathbf{B}_{\mathrm{D}})^{-1}$. The target population risk for $(\beta_{\mathrm{P}}, \beta_{\mathrm{D}}, \beta_0)$ becomes

$$\begin{aligned}&\mathbb{E}\left(\widetilde{Y} - \beta_{\mathrm{P}}^{\top}\widetilde{X}_{\mathrm{P}} - \beta_{\mathrm{D}}^{\top}\widetilde{X}_{\mathrm{D}} - \beta_0\right)^2\\ &= \left(1 - \beta_{\mathrm{D}}^{\top}H_{\mathrm{D}}\omega_{\mathrm{D}}\right)^2\sigma^2 + \mathbb{E}\left(\omega_{\mathrm{P}}^{\top}\widetilde{X}_{\mathrm{P}} - \beta_{\mathrm{P}}^{\top}\widetilde{X}_{\mathrm{P}} - \beta_{\mathrm{D}}^{\top}H_{\mathrm{D}}b_{\mathrm{D}}\omega_{\mathrm{P}}^{\top}\widetilde{X}_{\mathrm{P}} - \beta_{\mathrm{D}}^{\top}H_{\mathrm{D}}\mathbf{B}_{\text{d-p}}\widetilde{X}_{\mathrm{P}}\right)^2\\ &+ \left(\beta_{\mathrm{D}}^{\top}H_{\mathrm{D}}\widetilde{a}_{X,\mathrm{D}} + \beta_0\right)^2 + \beta_{\mathrm{D}}^{\top}H_{\mathrm{D}}\Sigma_{\mathrm{D}}H_{\mathrm{D}}^{\top}\beta_{\mathrm{D}}.\end{aligned}$$

The minimization on β_{P} and β_0 can be solved easily and it remains a quadratic program on β_{D}. Since this quadratic program is similar to that in the proof of Theorem 1, we conclude by the referring to the part where we solve the quadratic program in Appendix B.1.

Appendix D. Additional CIRM variants: RII, RIIRMI, CIRMI

The target population risk of CIP and CIRM estimator discussed above still have dependence on $a_Y^{(1)} - \widetilde{a}_Y$. As we have explained after Theorem 5, this is because Assumption 2 does not rule out the unidentifiable scenario. When we add additional assumptions such as $b \notin \text{span}\left(a_X^{(2)} - a_X^{(1)}, \ldots, a_X^{(M)} - a_X^{(1)}\right)$, we show that there are new methods which take advantage of this assumption to get rid of $a_Y^{(1)} - \widetilde{a}_Y$ dependence in the risk. We introduce three new estimators.

- **RII-mean:** the population residual invariance and independent estimator where mean is matched across environments

$$\begin{aligned}f_{\mathrm{RII}}(x) &:= x^{\top}\beta_{\mathrm{RII}} + \beta_{\mathrm{RII},0}\\ \beta_{\mathrm{RII}}, \beta_{\mathrm{RII},0} &:= \underset{\beta,\beta_0}{\arg\min}\ \frac{1}{M}\sum_{k=1}^{M}\mathbb{E}_{(X,Y)\sim\mathcal{P}(k)}\left(Y - X^{\top}\beta - \beta_0\right)^2\\ &\qquad \text{s.t. } \mathbb{E}_{(X,Y)\sim\mathcal{P}(m)}\left[Y - X^{\top}\beta\right] = \mathbb{E}_{(X,Y)\sim\mathcal{P}(1)}\left[Y - X^{\top}\beta\right]\\ &\qquad \text{and } \mathbb{E}_{(X,Y)\sim\mathcal{P}(m)}\left[(Y - \mathbb{E}_{Y\sim\mathcal{P}_Y^{(m)}}[Y])\cdot(Y - X^{\top}\beta)\right] = 0, \forall m \in \{2,\ldots,M\}.\end{aligned} \tag{87}$$

  The idea of matching the residual has appeared in the invariant causal prediction paper Peters et al. (2016) and in the anchor regression paper Rothenhäusler et al. (2021). The residual independence penalty is also the core idea in the invariant causal prediction Peters et al. (2016) and Greenfeld and Shalit Greenfeld and Shalit (2020).

- **RIIRMI$^{(m)}$-mean:** the population residual invariant independent residual matching estimator with additional residual independence using m-th source environment

$$
\begin{aligned}
f^{(m)}_{\text{RIIRMI}}(x) &:= x^\top \beta^{(m)}_{\text{RIIRMI}} + \beta^{(m)}_{\text{RIIRMI},0} \\
\beta^{(m)}_{\text{RIIRMI}}, \beta^{(m)}_{\text{RIIRMI},0} &:= \arg\min_{\beta,\beta_0} \ \mathbb{E}_{(X,Y)\sim\mathcal{P}^{(m)}} \left(Y - X^\top\beta - \beta_0\right)^2 \\
&\text{s.t. } \mathbb{E}_{X\sim\mathcal{P}^{(m)}_X}\left[\beta^\top\left(X - \left(X^\top\beta_{\text{RII}}\right)\vartheta_{\text{RIIRMI}}\right)\right] = \mathbb{E}_{X\sim\widetilde{\mathcal{P}}_X}\left[\beta^\top\left(X - \left(X^\top\beta_{\text{RII}}\right)\vartheta_{\text{RIIRMI}}\right)\right] \\
&\qquad \text{and } \mathbb{E}_{(X,Y)\sim\mathcal{P}^{(m)}}\left[(Y - \mathbb{E}_{Y\sim\mathcal{P}^{(m)}_Y}[Y])\cdot(Y - X^\top\beta)\right] = 0, \qquad (88)
\end{aligned}
$$

  where, with $\mathcal{P}^{\text{source}}$ denote the uniform mixture of all source distributions,

$$
\vartheta_{\text{RIIRMI}} := \frac{\mathbb{E}_{(X,Y)\sim\mathcal{P}^{\text{source}}}\left[X\cdot(Y - \mathbb{E}[Y])\right]}{\mathbb{E}_{Y\sim\mathcal{P}^{\text{source}}_Y}\left[(Y - \mathbb{E}[Y])^2\right]}. \qquad (89)
$$

- **CIRMI$^{(m)}$-mean:** the population conditional invariant residual matching estimator with additional residual independence using m-th source environment

$$
\begin{aligned}
f^{(m)}_{\text{CIRMI}}(x) &:= x^\top \beta^{(m)}_{\text{CIRMI}} + \beta^{(m)}_{\text{CIRMI},0} \\
\beta^{(m)}_{\text{CIRMI}}, \beta^{(m)}_{\text{CIRMI},0} &:= \arg\min_{\beta,\beta_0} \ \mathbb{E}_{(X,Y)\sim\mathcal{P}^{(m)}} \left(Y - X^\top\beta - \beta_0\right)^2 \\
&\text{s.t. } \mathbb{E}_{X\sim\mathcal{P}^{(m)}_X}\left[\beta^\top\left(X - \left(X^\top\beta_{\text{CIP}}\right)\vartheta_{\text{CIRMI}}\right)\right] = \mathbb{E}_{X\sim\widetilde{\mathcal{P}}_X}\left[\beta^\top\left(X - \left(X^\top\beta_{\text{CIP}}\right)\vartheta_{\text{CIRMI}}\right)\right] \\
&\qquad \text{and } \mathbb{E}_{(X,Y)\sim\mathcal{P}^{(m)}}\left[(Y - \mathbb{E}_{Y\sim\mathcal{P}^{(m)}_Y}[Y])\cdot(Y - X^\top\beta)\right] = 0, \qquad (90)
\end{aligned}
$$

  where, with $\mathcal{P}^{\text{source}}$ denote the uniform mixture of all source distributions,

$$
\vartheta_{\text{CIRMI}} := \frac{\mathbb{E}_{(X,Y)\sim\mathcal{P}^{\text{source}}}\left[X\cdot(Y - \mathbb{E}[Y])\right]}{\mathbb{E}_{(X,Y)\sim\mathcal{P}^{\text{source}}}\left[(X^\top\beta_{\text{CIP}} - \mathbb{E}[X^\top\beta_{\text{CIP}}])\cdot(Y - \mathbb{E}[Y])\right]}. \qquad (91)
$$

Note that a common feature of the three estimators above is that they all have the residual independent constraint of the form

$$
\mathbb{E}_{(X,Y)\sim\mathcal{P}^{(m)}}\left[(Y - \mathbb{E}_{Y\sim\mathcal{P}^{(m)}_Y}[Y])\cdot(Y - X^\top\beta)\right] = 0.
$$

This constraint takes advantage of the anticausal prediction setting and restricts the estimator to ignorant of the intervention on Y.

Theorem 9 *Under data generation Assumption 2 and the additional assumption*

$$
b \notin \mathit{span}\left(a^{(2)}_X - a^{(1)}_X, \ldots, a^{(M)}_X - a^{(1)}_X\right), \qquad (92)
$$

the population target risk of ***RII-mean****,* ***RIIRMI$^{(1)}$-mean*** *and* ***CIRMI$^{(1)}$-mean*** *satisfies*

$$\tilde{R}\left(f_{RII}\right)=\frac{b^{\top}\left(\mathbb{I}_d-P_4P_4^{\top}\right)\Sigma^{1/2}\left(\mathbb{I}_d-G_4\right)\Sigma^{1/2}\left(\mathbb{I}_d-P_4P_4^{\top}\right)b}{\left\|\left(\mathbb{I}_d-P_4P_4^{\top}\right)b\right\|_2^4}, \tag{93}$$

where $P_4\in\mathbb{R}^{d\times p}$ *equals to* P_{CIP} *which is the matrix formed by an orthonormal basis of the* p*-dimensional subspace span* $\left(a_X^{(2)}-a_X^{(1)},\ldots,a_X^{(M)}-a_X^{(1)}\right)$, $G_4=\Sigma^{1/2}Q_4\left(Q_4^{\top}\Sigma Q_4\right)^{-1}Q_4^{\top}\Sigma^{1/2}$ *is a projection matrix of rank* $d-p-1$, $Q_4\in\mathbb{R}^{d\times(d-p-1)}$ *is the matrix with columns formed by completing the columns of* P_4 *and* $\frac{\left(\mathbb{I}_d-P_4P_4^{\top}\right)b}{\left\|\left(\mathbb{I}_d-P_4P_4^{\top}\right)b\right\|_2}$ *to an orthonormal basis of* $\mathbb{R}^d$ *via Gram-Schmidt orthogonalization,*

$$\tilde{R}(f_{RIIRMI}^{(m)})=\tilde{R}(f_{CIRMI}^{(m)})=\frac{b^{\top}\left(\mathbb{I}_d-uu^{\top}\right)\Sigma^{1/2}\left(\mathbb{I}_d-G_5\right)\Sigma^{1/2}\left(\mathbb{I}_d-uu^{\top}\right)b}{\left\|\left(\mathbb{I}_d-uu^{\top}\right)b\right\|_2^4}, \tag{94}$$

where $u=\frac{a_X^{(m)}-\widetilde{a}_X}{\left\|a_X^{(m)}-\widetilde{a}_X\right\|_2}$, $G_5=\Sigma^{1/2}Q_5\left(Q_5^{\top}\Sigma Q_5\right)^{-1}Q_5^{\top}\Sigma^{1/2}$ *is a projection matrix of rank* $d-2$, $Q_5\in\mathbb{R}^{d\times(d-2)}$ *is the matrix with columns formed by completing* u *and* v *to an orthonormal basis of* $\mathbb{R}^d$ *via Gram-Schmidt orthogonalization.*

We present a corollary that puts additional assumptions on how the interventions are positioned to make the results in Theorem 9 easier to understand.

Corollary 10 *In addition to Assumption 2 and the assumption in Equation* (92)*, assume* $\Sigma=\frac{\sigma^2}{\rho}\mathbb{I}_d$ *with* $\rho>0$*, then*

$$\tilde{R}\left(f_{RII}\right)=\frac{\sigma^2}{\rho\left\|b\right\|_2^2-\rho\left(P_4^{\top}b\right)^2} \tag{95}$$

$$\tilde{R}(f_{RIIRMI}^{(1)})=\tilde{R}(f_{CIRMI}^{(1)})=\frac{\sigma^2}{\rho\left\|b\right\|_2^2-\rho\left(u^{\top}b\right)^2}, \tag{96}$$

Additionally, if $a_X^{(2)}-a_X^{(1)},\ldots,a_X^{(M)}-a_X^{(1)},\xi$ *are generated independently from the standard Gaussian distribution and* $\widetilde{a}_X-a^{(1)}=P_{CIP}\xi$*, then with high probability, we have*

$$dim\left(span\left(a_X^{(2)}-a_X^{(1)},\ldots,a_X^{(M)}-a_X^{(1)}\right)\right)=M-1.$$

$$\frac{\sigma^2}{\rho\left(1-\frac{(M-1)}{3d}\right)\left\|b\right\|_2^2}\leq\tilde{R}\left(f_{RII}\right)\leq\frac{\sigma^2}{\rho\left(1-\frac{3(M-1)}{d}\right)\left\|b\right\|_2^2} \tag{97}$$

$$\tilde{R}(f_{RIIRMI}^{(m)})=\tilde{R}(f_{CIRMI}^{(m)})\leq\frac{\sigma^2}{\rho\left(1-\frac{c}{d}\right)\left\|b\right\|_2^2} \tag{98}$$

where c *is a constant.*

	Estimator	**Target population risk upper bound** interventions under general position
Intervention on Y (Corollary 10, M sources)	DIP$^{(m)}$	$\frac{\sigma^2}{1+\rho(1-\frac{c}{d})\left\|b\right\|_2^2} + \left(\widetilde{a}_Y - a_Y^{(m)}\right)^2$
	CIP	$\frac{\sigma^2}{1+\rho(1-\frac{c(M-1)}{d})\left\|b\right\|_2^2} + \frac{(\widetilde{a}_Y - \bar{a}_Y) - \Delta_Y}{\left[1+\rho(1-\frac{c(M-1)}{d})\left\|b\right\|_2^2\right]^2}$
	CIRM$^{(m)}$	$\frac{\sigma^2}{1+\rho(1-\frac{c}{d})\left\|b\right\|_2^2} + \frac{\left(\widetilde{a}_Y - a_Y^{(m)}\right)^2}{\left[1+\rho(1-\frac{c}{d})\left\|b\right\|_2^2\right]^2}$
	RII	$\frac{\sigma^2}{\rho(1-\frac{c(M-1)}{d})\left\|b\right\|_2^2}$
	RIIRMI$^{(m)}$	$\frac{\sigma^2}{\rho(1-\frac{c}{d})\left\|b\right\|_2^2}$
	CIRMI$^{(m)}$	$\frac{\sigma^2}{\rho(1-\frac{c}{d})\left\|b\right\|_2^2}$

Table 7: Summary of target population risk for different estimators for anticausal domain adaptation under the assumptions of Corollary 10. Here c is a constant. For simplicity, we only compare the target population risk upper bound under high probability when the interventions are generated i.i.d. Gaussian.

The target population risks of RII, RIIRMI, CIRMI are compared with CIRM under the assumptions of Corollary 10. Under the additional assumption (92), the new DA methods RIIRMI and CIRMI effectively remove the $\widetilde{a}_Y - a_Y^{(m)}$ dependency when compared to CIRM. However, RIIRMI and CIRMI have slightly worse target population risk when there is no intervention on the label $a_Y^{(1)} = a_Y^{(2)} = \ldots = \widetilde{a}_Y = 0$.

D.1 Proof of Theorem 9

Risk of RII: Using the SEM (74) and the residual expression (75), the RII objective (87) becomes

$$\begin{aligned}
&\min_{\beta,\beta_0} \left(1-\beta^\top Hb\right)^2 \sigma^2 + \frac{1}{M}\sum_{m=1}^{M}\left(\left(1-\beta^\top Hb\right)a_Y^{(m)} - \beta^\top H a_X^{(m)} - \beta_0\right)^2 + \beta^\top H\Sigma H^\top \beta \\
&\text{s.t. } \left(1-\beta^\top Hb\right)\left(a_Y^{(m)} - a_Y^{(1)}\right) + \beta^\top H\left(a_X^{(m)} - a_X^{(1)}\right) = 0,\ \forall m \in \{2,\ldots,M\} \\
&\text{and } \left(1-\beta^\top Hb\right)^2\sigma^2 = 0.
\end{aligned}$$

The second constraint above ensures that

$$1 - \beta^\top Hb = 0.$$

This observation allows us to simplify the RII minimization problem above and obtain

$$\begin{aligned}
&\min_{\beta,\beta_0} \frac{1}{M} \sum_{m=1}^{M} \left(0 - \beta^\top H a_X^{(m)} - \beta_0\right)^2 + \beta^\top H \Sigma H^\top \beta \\
&\text{s.t. } \beta^\top H \left(a_X^{(m)} - a_X^{(1)}\right) = 0, \ \forall m \in \{2, \ldots, M\} \\
&\text{and } \beta^\top H b = 1. \qquad (99)
\end{aligned}$$

The objective (99) is a quadratic program with linear constraints. First, the one-dimensional quadratic program on β_0 can be solved easily by setting the derivative to zero.

$$\begin{aligned}
\beta_0^* &= \frac{1}{M} \sum_{m=1}^{M} \beta^{*\top} H a_X^{(m)} \\
&= \beta^{*\top} H a_X^{(1)},
\end{aligned}$$

where the last equality follows from the first constraint in Equation (99). Second, since H is invertible, we can re-parametrize the optimization on β as follows

$$\begin{aligned}
&\min_{\gamma \in \mathbb{R}^d} \gamma^\top \Sigma \gamma \qquad (100) \\
&\text{s.t. } \gamma^\top \left(a_X^{(m)} - a_X^{(1)}\right) = 0, \ \forall m \in \{2, \ldots, M\} \\
&\text{and } \gamma b = 1. \qquad (101)
\end{aligned}$$

Let P_4 be the matrix with columns formed by an orthonormal basis of the p-dimensional subspace span $\left(a^{(2)} - a^{(1)}, \ldots, a^{(M)} - a^{(1)}\right)$. Define

$$v := \frac{\left(\mathbb{I}_d - P_4 P_4^\top\right) b}{\left\| \left(\mathbb{I}_d - P_4 P_4^\top\right) b \right\|_2}.$$

v is well defined because $\left\| \left(\mathbb{I}_d - P_4 P_4^\top\right) b \right\|_2 \neq 0$ by assumption 92. By construction, we have $P_4^\top v = 0$ and also

$$\left(\frac{v}{\left\| \left(\mathbb{I}_d - P_4 P_4^\top\right) b \right\|_2} \right)^\top b = 1 \qquad (102)$$

.

Let $Q_4 \in \mathbb{R}^{d \times (d-p-1)}$ be the matrix with columns formed by completing the columns of P_4 and v to an orthonormal basis of $\mathbb{R}^d$ via Gram-Schmidt orthogonalization. Because of Equation (102), the following map

$$\begin{aligned}
\mathbb{R}^{d-p-1} &\to \mathbb{R}^d \\
\zeta &\mapsto \frac{v}{\left\| \left(\mathbb{I}_d - P_4 P_4^\top\right) b \right\|_2} + Q_4 \zeta
\end{aligned}$$

constitutes a bijection between $\mathbb{R}^{d-p-1}$ and the set $\left\{\gamma \in \mathbb{R}^d \mid P^\top \gamma = 0, \gamma^\top b = 1\right\}$. With the change of variable, the constrained quadratic program in Equation (100) is equivalent to the following unconstrained one

$$\min_{\zeta \in \mathbb{R}^{d-p-1}} \left(\frac{v}{\left\| \left(\mathbb{I}_d - P_4 P_4^\top \right) b \right\|_2} + Q_4 \zeta \right)^\top \Sigma \left(\frac{v}{\left\| \left(\mathbb{I}_d - P_4 P_4^\top \right) b \right\|_2} + Q_4 \zeta \right) \tag{103}$$

Solving the unconstrained quadratic program by setting gradient to zero, we obtain the minimizer

$$\zeta^* = -\frac{\left({Q_4}^\top \Sigma Q_4 \right)^{-1} {Q_4}^\top \Sigma v}{\left\| \left(\mathbb{I}_d - P_4 P_4^\top \right) b \right\|_2}.$$

Transforming the variables back, the RII estimator is

$$\begin{aligned} \beta_{\mathrm{RII}} &= \left(\mathbb{I}_d - \mathbf{B} \right)^\top \frac{\Sigma^{-1/2} \left(\mathbb{I}_d - G_4 \right) \Sigma^{1/2} v}{\left\| \left(\mathbb{I}_d - P_4 P_4^\top \right) b \right\|_2} \\ \beta_{\mathrm{RII},0} &= -\beta_{\mathrm{RII}}^\top H a_X^{(1)}, \end{aligned} \tag{104}$$

where $G_4 = \Sigma^{1/2} Q_4 \left(Q_4^\top \Sigma Q_4 \right)^{-1} Q_4^\top \Sigma^{1/2}$ is a projection matrix of rank $d-p-1$. According to the form of the target residual (76), the target population risk for (β, β_0) takes the following form

$$\left(1 - \beta^\top H b \right)^2 \sigma^2 + \left(\left(1 - \beta^\top H b \right) \widetilde{a}_Y - \beta^\top H \widetilde{a}_X - \beta_0 \right)^2 + \beta^\top H \Sigma H^\top \beta.$$

Plugging in the RII estimator into the equation above, because $\beta_{\mathrm{RII}}^\top H b = 1$ and $\widetilde{a}_X - a_X^{(1)} \in \operatorname{span}(a_X^{(2)} - a_X^{(1)}, \ldots, a_X^{(m)} - a_X^{(1)})$, we obtain the target population risk of RII

$$\tilde{R}(f_{\mathrm{RII}}) = \frac{v^\top \Sigma^{1/2} \left(\mathbb{I}_d - G_4 \right) \Sigma^{1/2} v}{\left\| \left(\mathbb{I}_d - P_4 P_4^\top \right) b \right\|_2^2}.$$

Risk of RIIRMI: We start by deriving $\vartheta_{\mathrm{RIIRMI}}$ defined in Equation (89). It calculates the correlation between X and Y. Using the SEM (74), we obtain

$$\vartheta_{\mathrm{RIIRMI}} = H b. \tag{105}$$

Note that just like in CIRM, the vector $\vartheta_{\mathrm{RIIRMI}}$ is also co-linear with the Y component in the covariate expression in Equation (81). So the idea of RIIRMI is very similar to that of CIRM: it uses $\widetilde{X}^\top \beta_{\mathrm{RII}}$ as a proxy for the unobserved $\widetilde{Y}$ to remove the $\widetilde{Y}$ part in the covariates so that the DIP matching based ideas can still be applied.

With the $\vartheta_{\mathrm{RIIRMI}}$ expression (105) and the β_{RII} expression (104), the LHS of the first RIIRMI constraint (88) becomes

$$\begin{aligned} & X^{(m)} - \left({X^{(m)}}^\top \beta_{\mathrm{RII}} \right) \vartheta_{\mathrm{RIIRMI}} \\ &= \left(Y^{(m)} - {X^{(m)}}^\top \beta_{\mathrm{RII}} \right) H b + H a_X^{(m)} + H \varepsilon_X^{(m)} \\ &= \left(\left(1 - \beta_{\mathrm{RII}}^\top H b \right) Y^{(m)} - \beta_{\mathrm{RII}}^\top H a^{(m)} - \beta_{\mathrm{RII}}^\top H \varepsilon^{(m)} \right) H b + H a_X^{(m)} + H \varepsilon_X^{(m)} \\ &\overset{(i)}{=} \left(0 - \beta_{\mathrm{RII}}^\top H a^{(1)} - \beta_{\mathrm{RII}}^\top H \varepsilon^{(m)} \right) H b + H a_X^{(m)} + H \varepsilon_X^{(m)}, \end{aligned} \tag{106}$$

where the last inequality follows from the two constraints in the RII estimator (87). Similar expression can be obtained for the target environments by taking into account that $\widetilde{a}_X \in \text{span}\left(a_X^{(1)}, \ldots, a_X^{(M)}\right)$,

$$\widetilde{X} - \left(\widetilde{X}^\top \beta_{\text{RII}}\right) \vartheta_{\text{RIIRMI}} = \left(\beta_{\text{RII}}^\top H a^{(1)} + \beta_{\text{RII}}^\top H \widetilde{\varepsilon}\right) Hb + H\widetilde{a}_X + H\widetilde{\varepsilon}_X. \tag{107}$$

Multiply Equation (106) and (107) by β and take expectation, we obtain a simplified form of the first RIIRMI constraint (12)

$$\beta^\top H a_X^{(m)} = \beta^\top H \widetilde{a}_X.$$

Note that the first RIIRMI constraint is the same as the one in DIP or CIRM. With the above observation, RIIRMI$^{(1)}$ objective (88) becomes

$$\begin{aligned} &\min_{\beta, \beta_0} \left(0 - \beta^\top H a_X^{(m)} - \beta_0\right)^2 + \beta^\top H \Sigma H^\top \beta \\ &\text{s.t. } \beta^\top H \left(a_X^{(m)} - \widetilde{a}\right) = 0 \\ &\text{and } \beta^\top H b = 1. \end{aligned} \tag{108}$$

This objective is similar to the one (85) in the proof of CIRM, with the only difference that there is one additional constraint $\beta^\top H b = 1$. This objective is also similar to the one (99) in the proof of RII, with the difference that there is only one constraint of the type $\beta^\top H(a_X^{(m)} - \widetilde{a}_X) = 0$. Following the proof of RII to solve the quadratic program with linear constraints, we define

$$v := \frac{\left(\mathbb{I}_d - uu^\top\right) b}{\left\|\left(\mathbb{I}_d - uu^\top\right) b\right\|_2},$$

where $u = \frac{a_X^{(m)} - \widetilde{a}_X}{\left\|a_X^{(m)} - \widetilde{a}_X\right\|_2}$ and define $Q_5 \in \mathbb{R}^{d \times (d-2)}$ be the matrix with columns formed by completing u and v to an orthonormal basis of $\mathbb{R}^d$ via Gram-Schmidt orthogonalization, then the RIIRMI estimator is

$$\begin{aligned} \beta_{\text{RIIRMI}}^{(m)} &= \left(\mathbb{I}_d - \mathbf{B}\right)^\top \frac{\Sigma^{-1/2}\left(\mathbb{I}_d - G_5\right)\Sigma^{1/2}\left(\mathbb{I}_d - uu^\top\right) b}{\left\|\left(\mathbb{I}_d - uu^\top\right) b\right\|_2^2} \\ \beta_{\text{RIIRMI},0}^{(m)} &= -{\beta_{\text{RIIRMI}}^{(m)}}^\top H a_X^{(m)}. \end{aligned} \tag{109}$$

where $G_5 = \Sigma^{1/2} Q_5 \left(Q_5^\top \Sigma Q_5\right)^{-1} Q_5^\top \Sigma^{1/2}$ is a projection matrix of rank $d-2$. According to the form of the target residual (76), the target population risk for (β, β_0) takes the following form

$$\left(1 - \beta^\top H b\right)^2 \sigma^2 + \left(\left(1 - \beta^\top H b\right) \widetilde{a}_Y - \beta^\top H \widetilde{a}_X - \beta_0\right)^2 + \beta^\top H \Sigma H^\top \beta.$$

Plugging in the RIIRMI[1] estimator into the equation above, because $\beta_{\text{RII}}^\top H b = 1$ and $\beta^\top H a_X^{(m)} = \beta^\top H \widetilde{a}_X$, we obtain the target population risk of RIIRMI[1]

$$\tilde{R}(f_{\text{RIIRMI}}^{(m)}) = \frac{b^\top \left(\mathbb{I}_d - uu^\top\right) \Sigma^{1/2} \left(\mathbb{I}_d - G_5\right) \Sigma^{1/2} \left(\mathbb{I}_d - uu^\top\right) b}{\left\|\left(\mathbb{I}_d - uu^\top\right) b\right\|_2^4}.$$

Risk of CIRMI: The proof for CIRMI follows easily by combining parts of proofs in CIRM and RIIRMI. The CIRMI$^{(1)}$ estimator has one additional constraint compared to the CIRM$^{(1)}$ estimator in Equation (85)

$$\begin{aligned}
&\min_{\beta,\beta_0} \left(1-\beta^\top Hb\right)^2 \sigma^2 + \left(\left(1-\beta^\top Hb\right) a_Y^{(1)} - \beta^\top H a_X^{(1)} - \beta_0\right)^2 + \beta^\top H\Sigma H^\top \beta \\
&\text{s.t. } \beta^\top H\left(a_X^{(1)} - \widetilde{a}\right) = 0 \\
&\text{and } 1-\beta^\top Hb = 0.
\end{aligned} \tag{110}$$

In fact, the above optimization problem is exactly the same as RIIRMI in Equation (108). Consequently, the CIRMI solution is the same as that of RIIRMI.

$$\begin{aligned}
\beta^{(1)}_{\text{CIRMI}} &= \beta^{(1)}_{\text{RIIRMI}} \\
\beta^{(1)}_{\text{CIRMI},0} &= \beta^{(1)}_{\text{RIIRMI},0},
\end{aligned} \tag{111}$$

$$\tilde{R}(f^{(1)}_{\text{CIRMI}}) = \tilde{R}(f^{(1)}_{\text{RIIRMI}}).$$

D.2 Proof of Corollary 10

The proof of Corollary 10 follows similarly as that of Corollary 6 in Appendix C.2.

References

M Amini and Patrick Gallinari. Semi-supervised learning with an explicit label-error model for misclassified data. In *Proceedings of the 18th International Joint Conferences on Artificial Intelligence*, pages 555–560, 2003.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations (ICLR)*, 2019.

Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 769–776, 2013.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 137–144, 2007.

Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, pages 129–136, 2010.

John Blitzer, Sham Kakade, and Dean Foster. Domain adaptation with coupled subspaces. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 173–181. JMLR Workshop and Conference Proceedings, 2011.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

Peter Bühlmann. Invariance, causality and robustness. *Statistical Science*, 35(3):404–426, 2020.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, pages 11190–11201, 2019.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. Semi-supervised learning. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

Corinna Cortes and Mehryar Mohri. Domain adaptation in regression. In *International Conference on Algorithmic Learning Theory*, pages 308–323. Springer, 2011.

Corinna Cortes and Mehryar Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, 2014.

Corinna Cortes, Mehryar Mohri, and Andrés Muñoz Medina. Adaptation algorithm and theory based on generalized discrepancy. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 169–178, 2015.

John Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*, 2018.

Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of Science*, 74(5):981–995, 2007.

Li Fei-Fei. Imagenet: crowdsourcing, benchmarking & other cool things. In *CMU VASC Seminar*, volume 16, pages 18–25, 2010.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*, 2017.

Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. A unified view of label shift estimation. In *Advances in Neural Information Processing Systems*, volume 33, pages 3290–3300. Curran Associates, Inc., 2020.

Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7):1414–1430, 2016.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524, 2019.

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

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

Ian Goodfellow, Patrick McDaniel, and Nicolas Papernot. Making machine learning robust against adversarial inputs. *Communications of the ACM*, 61(7):56–66, 2018.

Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *2011 International Conference on Computer Vision*, pages 999–1006. IEEE, 2011.

Daniel Greenfeld and Uri Shalit. Robust learning with the Hilbert-Schmidt independence criterionilbert-schmidt independence criterion. In *International Conference on Machine Learning*, pages 3759–3768. PMLR, 2020.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *Machine Learning*, 110(2):303–348, 2021.

Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Algorithms and theory for multiple-source adaptation. In *Advances in Neural Information Processing Systems*, pages 8246–8256, 2018.

Peter J Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, pages 73–101, 1964.

Fredrik D Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 527–536. PMLR, 2019.

Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning*, pages 5468–5479. PMLR, 2020.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Yann LeCun. The MNIST database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*, 1998.

Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Yitong Li, Michael Murias, Samantha Major, Geraldine Dawson, and David Carlson. On target shift in adversarial domain adaptation. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 616–625. PMLR, 2019.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International conference on machine learning*, pages 3122–3130. PMLR, 2018.

Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 10869–10879, 2018.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *22nd Conference on Learning Theory, COLT*, 2009.

Geoffrey J McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

Nicolai Meinshausen. Causality from a distributional robustness point of view. In *2018 IEEE Data Science Workshop (DSW)*, pages 6–10. IEEE, 2018.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18, 2013.

Jerzy Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, 2019.

Kamal Nigam, Andrew McCallum, and Tom Mitchell. Semi-supervised text classification using EM. *Semi-Supervised Learning*, pages 33–56, 2006.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2009.

Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2010.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8026–8037, 2019.

Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Springer, 2000.

Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, pages 579–595, 2014.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1406–1415, 2019.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. MIT Press, 2009.

Aditi Raghunathan, Jacob Steinhardt, and Percy S Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems*, pages 10877–10887, 2018.

Adwait Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In *Conference on empirical methods in natural language processing*, 1996.

Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younès Bennani. A survey on domain adaptation theory. *arXiv preprint arXiv:2004.11829*, 2020.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

Dominik Rothenhäusler, Peter Bühlmann, Nicolai Meinshausen, et al. Causal dantzig: fast inference in linear structural equation models with hidden variables under additive interventions. *Annals of Statistics*, 47(3):1688–1722, 2019.

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

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3): 211–252, 2015.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Coference on Machine Learning*, pages 459–466, 2012.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Amos Storkey. When training and test sets are different: characterizing learning transfer. *Dataset shift in machine learning*, pages 3–28, 2009.

Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J Gordon. Domain adaptation with conditional distribution matching and generalized label shift. *Advances in Neural Information Processing Systems*, 33, 2020.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre Alvise-Rebuffi, Ira Ktena, and Taylan Cemgil. A fine-grained analysis on distribution shift. *arXiv preprint arXiv:2110.11328*, 2021.

Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.

Zikang Yuan, Dongfu Zhu, Cheng Chi, Jinhui Tang, Chunyuan Liao, and Xin Yang. Visual-inertial state estimation with pre-integration correction for robust mobile augmented reality. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1410–1418, 2019.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International Conference on Machine Learning*, pages 7523–7532. PMLR, 2019.

Xiaojin Jerry Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2005.