Journal of Machine Learning Research 25 (2024) 1-57 Submitted 9/22; Revised 6/24; Published 9/24

On Causality in Domain Adaptation and Semi-Supervised Learning: an Information-Theoretic Analysis for Parametric Models

Xuetong Wu XUETONGW1@STUDENT.UNIMELB.EDU.AU
Department of Electrical and Electronic Engineering

Mingming Gong MINGMING.GONG@UNIMELB.EDU.AU
School of Mathematics and Statistics

Jonathan H. Manton JMANTON@UNIMELB.EDU.AU
Department of Electrical and Electronic Engineering

Uwe Aickelin UWE.AICKELIN@UNIMELB.EDU.AU
Department of Computing and Information Systems

Jingge Zhu JINGGE.ZHU@UNIMELB.EDU.AU
Department of Electrical and Electronic Engineering
University of Melbourne
Parkville, 3010, Australia

Editor: Ilya Shpitser

Abstract

Recent advancements in unsupervised domain adaptation (UDA) and semi-supervised learning (SSL), particularly incorporating causality, have led to significant methodological improvements in these learning problems. However, a formal theory that explains the role of causality in the generalization performance of UDA/SSL is still lacking. In this paper, we consider the UDA/SSL scenarios where we access m labelled source data and n unlabelled target data as training instances under different causal settings with a parametric probabilistic model. We study the learning performance (e.g., excess risk) of prediction in the target domain from an information-theoretic perspective. Specifically, we distinguish two scenarios: the learning problem is called causal learning if the feature is the cause and the label is the effect, and is called anti-causal learning otherwise. We show that in causal learning, the excess risk depends on the size of the source sample at a rate of $O(\frac{1}{m})$ only if the labelling distribution between the source and target domains remains unchanged. In anti-causal learning, we show that the unlabelled data dominate the performance at a rate of typically $O(\frac{1}{n})$. These results bring out the relationship between the data sample size and the hardness of the learning problem with different causal mechanisms.

Keywords: Causality, domain adaptation, semi-supervised learning, parametric models, generalization error

1. Introduction

A common obstacle in many real-world learning problems is that the training and testing data may originate from different distributions. Such a paradigm is known as the "domain adaptation" problem. Specifically, we consider the unsupervised domain adaptation (UDA) scenarios in which we have two datasets drawn from different distributions, namely the "source" and "target" distributions, respectively. The source dataset includes both features and labels, whereas the target dataset contains only features and no labels. The goal is to train a model that performs well on the **target** distribution. This assumption is particularly interesting because it reflects real-world scenarios where the target labels are often unavailable.

Schölkopf et al. (2012) began the pioneering work of developing a framework that links causal mechanisms with UDA, where the objective is to predict the label Y using feature X. They delve into two fundamental causal settings: the "causal learning" setting, where X is the cause of Y, and the "anti-causal learning" setting, where Y is the cause of X. An interesting empirical observation made in the paper is that semi-supervised learning (SSL) - a machine learning paradigm where the model is trained on a mix of labelled and unlabelled data - improves learning performance in the anti-causal direction but does not provide a similar boost in the causal direction. This finding suggests that, given known causal structures, we may be able to enhance the generalization capabilities of machine learning algorithms strategically. Even though numerous causality-driven machine learning algorithms have demonstrated their effectiveness empirically (Schölkopf et al., 2012; Zhang et al., 2013; Gong et al., 2016), the analytical part remains less investigated. Specifically, understanding how causality impacts learning performance and how the unlabelled target data and labelled source data contribute to the prediction under specific causal settings is yet to be deepened. This paper attempts to demystify how causal directions influence generalization ability and how the labelled source and unlabelled target data contribute to the prediction in the UDA/SSL settings under generative parametric models. Specifically, we examine the excess risk under various distribution shift conditions under the UDA setup, including the case of no distribution shifts as seen in SSL.

Our main results reveal that in the causal learning scenario, the unlabelled target data do not contribute to the prediction, and the source data only aids in reducing the excess risk when the conditional probability distribution $P(Y|X)$ remains consistent between source and target domains. Conversely, in anti-causal learning, unlabelled data are always useful. However, the usefulness of the source data, in terms of the convergence rate for excess risk, is contingent on the distribution shift conditions.In situations where the causal relationship between the feature and the label is unknown, improving generalization capability in domain adaptation requires careful consideration when making predictions from either a causal or anti-causal direction. This understanding enables us to design more efficient learning algorithms that are equipped to handle the challenges presented by complex real-world learning problems.

2. Related Work

Causal Inference and Machine Learning. Two important frameworks in causal inference are the potential outcome (counterfactual) framework and the structural causal model (SCM) (Holland, 1986; Hernán and Robins, 2010; Imbens and Rubin, 2015; Pearl and Mackenzie, 2018)[1], which allows reasoning about a system not only under observation but also under intervention, and they have become an influential tool in several machine learning problems. For example, Schölkopf et al. (2012) study the causal and anti-causal learning for domain adaptation with an additive noise SCM. Bottou et al. (2013) carry out the counterfactual analysis for the advertisement placement problem, allowing more flexibility in decision-making and thus improving the system performance. More recently, Schölkopf (2022) put forward significant issues such as i.i.d. assumptions and generalization ability of current machine learning algorithms and summarized the intrinsic connections between machine learning and the causality. Moraffah et al. (2020) reviewed several causal interpretable models and suggested that the causal interpretable model under these causal and anti-causal frameworks is a way to explain the black-box machine learning algorithms. Makhlouf et al. (2020) argue that causality-based machine learning algorithms are necessary to address the problem of fairness appropriately.

However, although the causal models are favourable for specific learning regimes, only a few works generally consider generalization ability. To name a few, Kilbertus et al. (2018) argue that the generalization capabilities for anti-causal learning problems are associated with the hypothesis space searching and validation, but no theoretical analysis is presented. Kuang et al. (2018) and Cui and Athey (2022) develop a stable learning algorithm that is robust across different underlying distributions and derives the generalization error bound with the "causal" features, which are stable across different environments. Arjovsky et al. (2019) propose the invariant risk minimization to generalize well across different domains. Chen and Bühlmann (2021) develop a theoretical framework via the linear structural causal models, allowing comparisons of the learning performance for existing domain adaptation methods.

Domain Adaptation Most techniques to conquer domain adaptation problems are purely statistics-based without referring to causal concepts. For example, the instance-based methods identify source samples that bear similarities to target samples based on the probability density ratio on the marginal distribution of features (Cortes et al., 2008; Gretton et al., 2009). The feature-based methods will seek a new latent space where the discrepancy of the empirical distribution embeddings between the source and target domains are small under some metric (Pan et al., 2010; Zhang et al., 2017). The popular deep learning-based methods will involve deep generative networks to align distributions between source and target domains (Tzeng et al., 2017; Shen et al., 2018). However, recent works have shown that introducing causal concepts leads to more robust and efficient algorithms for domain adaptation. The main idea is to identify and extract the transferable components that are invariant across different domains under certain causal models (Gong et al., 2016; Magliacane et al., 2018; Rojas-Carulla et al., 2018; Mahajan et al., 2021). Nevertheless, they mainly focus on the empirical verification of the effect of source samples instead of a theoretical analysis of their algorithms. To rigorously investigate the generalization ability

1. It is sometimes also called structural equation model (SEM).

and usefulness of the source and target data, Wu et al. (2021) give an attempt to interpret the transfer learning in terms of parametric probabilistic models. Kpotufe and Martinet (2018) study the covariate shift problem and derive the minimax rate with the notion of "transfer component". Cai and Wei (2021) investigate the concept drift problem and establish the optimal minimax convergence rate with weighted k-nearest neighbour classifier. Maity et al. (2022) consider the target shift condition and derive the optimal minimax rate in non-parametric classification.

Semi-Supervised Learning Semi-supervised learning aims to learn the predictor with scarce labelled and abundant unlabelled data. The crucial questions are when the unlabelled data are useful and how to avoid their negative impact. On the practical side, Schölkopf et al. (2012) find that the unlabelled data will be useful for prediction when these data are the effect of their corresponding (unknown) labels. Li and Zhou (2014) propose a robust SVM-based algorithm to prevent the unlabelled data from hurting the performance. Under generalized linear models, Yuval and Rosset (2022) analyze the effectiveness of the unlabelled data via risk minimization. On the theoretical side, Castelli and Cover (1996) and Zhang and Oles (2000) pose the parametric assumptions on data distributions and claim the value of the unlabelled data depends on the Fisher information matrices of the distribution parameters. A similar argument is made in Zhu (2020) that if the unlabelled data contain all information of the required parameters, they will be equally useful as the labelled data. Seeger (2000) and Liang et al. (2007) suggest that for certain data-generating processes, the unlabelled data is not useful from a Bayesian perspective. We refer to Mey and Loog (2019) for other plentiful theoretical results on semi-supervised learning. Our methods provide a pathway to probabilistically analyze the semi-supervised learning problem and definitude the conditions when the unlabelled data are useful from a causal point of view.

3. Preliminaries

In this paper, we use the convention that capital letters denote the random variables and small letters their realizations. We define $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. The notation $f(n) \asymp g(n)$ means that there exists some positive integer n_0 such that for all $n > n_0$, $c_1 g(n) \leq f(n) \leq c_2 g(n)$ always holds for some positive c_1 and c_2. We also use $f(n) = O(g(n))$ by meaning that there exists some integer n_0 such that for all $n > n_0$, $f(n) \leq c_3 g(n)$ always holds for some positive value c_3. We denote the KL divergence between two distributions P and Q by $\mathrm{KL}(P\|Q) = \mathbb{E}_P\left[\log \frac{dP}{dQ}\right]$. We use $P(X) \ll Q(X)$ to denote that the probability distribution $P(X)$ is absolutely continuous w.r.t. $Q(X)$. If not otherwise specified, the notation $\mathbb{E}_\theta[\cdot]$ denotes the expectation taken over all data examples involved that are drawn from P_θ.

3.1 Information Theory Basics

Before proceeding, we will define several common information theory quantities such as entropy, mutual information, and Kullback-Leibler divergence (KL divergence), and state several well-known results on these measures that will be referenced in the literature. For more information on the basics, the readers can refer to Cover and Thomas (2006). The

Shannon entropy of a discrete random variable X is defined as:

$$H(X) = -\sum_{x \in \mathcal{X}} \mathbb{P}(X = x) \log \mathbb{P}(X = x). \tag{1}$$

For continuous random variable X with the probability density function $p(x)$, the differential entropy is defined as:

$$h(X) = -\int p(x) \log p(x) dx. \tag{2}$$

Note that for discrete r.v., the Shannon entropy is always nonnegative and bounded by $\log |\mathcal{X}|$ while the differential entropy is considered as a measure of relative information and can be negative. Next, we define the Kullback-Leibler divergence: for two probability measures P and Q, if P is absolutely continuous with respect to Q, the Kullback-Leibler divergence between P and Q is:

$$D(P\|Q) = \int \log \left(\frac{dP}{dQ} \right) dP,$$

where $\frac{dP}{dQ}$ is the Radon-Nikodym derivative of P with respect to Q. The KL divergence roughly estimates how different the two distributions P and Q are. For any probability distributions P and Q over the space Ω such that P is absolutely continuous with respect to Q, we have the non-negativity property such that $D(P\|Q) \geq 0$ and the quantity is usually non-symmetric, e.g., $D(P\|Q) \neq D(Q\|P)$ if $P \neq Q$. We can then define the mutual information between the random variables X and Y as:

$$I(X;Y) = D(P(X,Y)\|P(X)P(Y)), \tag{3}$$

which is the Kullback-Leibler divergence between the joint distribution of X and Y and the product of the marginal distributions. From the definition, it is clear that $I(X;Y) = I(Y;X)$, and the first property of the KL divergence implies that $I(X;Y)$ is nonnegative and $I(X;Y) = 0$ when X and Y are independent. Furthermore, we also define conditional mutual information as

$$I(X,Y|Z) = \mathbb{E}_Z \left[D(P(X,Y|Z)\|P(X|Z)P(Y|Z)) \right],$$

where it represents the amount of information gained about X by observing Y given a third variable Z.

3.2 Prediction with Mixture Strategy

Considering the effectiveness and complexity of UDA and SSL problems, we use the parametric distribution models as a critical component of our approach. The reason for this choice is that the distribution shifts can be characterized concisely by the parameter changes. This approach allows for a rigorous statistical framework in which the complexities of the learning problem can be analyzed.

The mixture strategy is an important concept in the field of statistical inference that was leveraged from Clarke and Barron (1994, 1990); Merhav and Feder (1998) with the

application of universal prediction, which involves the construction of a mixture distribution over the model parameters for prediction when the true distribution (parameters) is unknown. Here, "universal" means that the predictor does not depend on the unknown underlying distribution and performs essentially as well as if the distribution was known in advance. Furthermore, given these complexities and the distributional shifts of data sources, a mixture strategy becomes a natural choice for tackling these challenges in different domain adaptation settings as it allows us to integrate source and target distribution information, enabling a comprehensive understanding of the learning performance.

The mixture strategy has been extensively studied in the literature, with several important works exploring its properties and applications in various fields. For example, Feder et al. (1992); Merhav and Feder (1998); Cover and Ordentlich (1996) mainly focused on situations where data is drawn independently and identically from a single parametric distribution, which is similar to traditional online learning problems. However, the bounds obtained through the conditional mutual information cannot provide more quantitative insights for analyzing the regret. To this end, the previous works such as Clarke (1999); Clarke and Barron (1990); Zhu (2020) provided an asymptotic analysis for the conditional mutual information under the conventional online learning or semi-supervised learning problems, where the regret approximation is associated with the sample size and the prior distribution over the distribution parameters.

Mathematically, let θ be the parameter of interest that is involved in the model distribution, and let $p(\theta)$ be the prior distribution over θ. Assume we have the training dataset $\mathcal{D}$ with each $Z_i \in \mathcal{D}$ i.i.d. drawn from a distribution $p^*_\theta(Z)$. If we consider the predictor ω to be a probability distribution over the data sample Z, the logarithmic loss is then defined as

$$\ell(\omega, Z) = -\log \omega(Z). \tag{4}$$

We can define the expected loss on test data Z' as

$$L := -\mathbb{E}_{\theta^*}\left[\log Q(Z'|\mathcal{D})\right]. \tag{5}$$

where the mixture strategy involves constructing a mixture distribution over Z' for the testing data given the training data as

$$Q(Z'|\mathcal{D}) = \frac{\int p(\mathcal{D}, Z'|\theta)p(\theta)d\theta}{\int p(\mathcal{D}|\theta)p(\theta)d\theta} = \int p_\theta(Z')Q(\theta|\mathcal{D})d\theta, \tag{6}$$

where $Q(\theta|\mathcal{D})$ is the conditional distribution of the parameter θ given the dataset $\mathcal{D}$ induced by $Q(\mathcal{D}) = \int p(\mathcal{D}|\theta)p(\theta)d\theta$ and the joint distribution $p(\mathcal{D}, \theta) = p(\mathcal{D}|\theta)p(\theta)$, and $p(\theta)$ is a prior distribution over θ. From a Bayesian perspective, we assign a probability distribution $p(\theta)$ over the parameter space to represent our prior knowledge, and we update the posterior with the training data to approximate the underlying distributions. With the mixture strategy, the excess risk w.r.t. the best estimator could be rewritten as:

$$\begin{aligned} R &:= -\mathbb{E}_{\theta^*}\left[\log Q(Z'|\mathcal{D})\right] - \mathbb{E}_{\theta^*}\left[\log p_{\theta^*}(Z')\right] && (7)\\ &= \mathbb{E}_{\theta^*}\left[\log \frac{P_{\theta^*}(Z')}{Q(Z'|\mathcal{D})}\right]. && (8)\\ &= I(Z'; \theta^*|\mathcal{D}). && (9) \end{aligned}$$

The above characterization implies that under logarithmic loss, with a specific prior $p(\theta)$, the excess risk induced by the mixture strategy is captured by the conditional mutual information between the sample Z' and distribution parameter that is evaluated at θ^* given the training data, which naturally gives an interpretation on the amount of information that the test data point Z' carries about the true parameter θ^*, given the whole training set $\mathcal{D}$. Such an information-theoretic framework has been established and studied in SSL and online learning problems (see Merhav and Feder (1998); Zhan and Taylor (2015); Zhu (2020) for references). One advantage of this framework is that information-theoretic tools are powerful in studying asymptotic behaviours as well as deriving learning performance bounds for various statistical problems. This characterization also ensures minimax optimality, which means that irrespective of the underlying parameters, the resultant learning rate is guaranteed to be optimal, even in the worst-case scenario. Additionally, information-theoretic quantities such as mutual information and KL divergence (relative entropy) give natural interpretations for the learning bounds. Furthermore, when it comes to distribution parameter estimation, the mixture model is particularly beneficial when the data is believed to be generated from a certain underlying process, as it can provide a probabilistic representation of the diverse sub-populations, and this is particularly valuable where only assuming a single distribution could lead to skewed or inaccurate results (such as the plug-in method). On the other hand, while estimating a single distribution offers simplicity, the model is sensitive to outliers and may fall short when the data complexity is high or the sample size is small. Taking advantage of the robustness of the mixture strategy, this paper expands on the findings of Merhav and Feder (1998) and Zhu (2020), which were initially applied to conventional learning scenarios where the source and target originate from the same distribution. In the following, we will examine both UDA and SSL learning bounds across various distribution shift conditions by leveraging a mixture strategy grounded in causal and anti-causal settings.

4. Problem Formulation

We consider the typical *unsupervised domain adaptation* problem for classification. Given the labelled source data $D_s^m = (X_s^{(1)}, Y_s^{(1)}, \cdots, X_s^{(m)}, Y_s^{(m)})$ and the unlabelled target data $D_t^{\mathrm{U},n} = (X_t^{(1)}, \cdots, X_t^{(n)})$, we assume each source sample is i.i.d. drawn from a probability distribution $P_S(X, Y)$ and takes value in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and each target sample is i.i.d. drawn from the marginal distribution of $P_T(X, Y)$ and takes value in $\mathcal{X}$. In general, $P_S(X, Y)$ is different from $P_T(X, Y)$, and both $\mathcal{X}$ and $\mathcal{Y}$ can be discrete or continuous. For simplicity, we consider the case where both X and Y are discrete in this paper. We point out that the analysis in the paper continues to hold for a continuous Y in the causal learning case and for a continuous X in the anti-causal learning case. We will predict the label Y_t' for the previously unseen sample X_t' in target domain, utilising the training sample D_s^m and $D_t^{\mathrm{U},n}$ with the learning algorithm $\mathcal{A} : \mathcal{Z}^m \times \mathcal{X}^n \times \mathcal{X} \to \mathcal{B}$, whose output b is the distribution-independent *predictor* for the outcome Y_t' in the predictor space $\mathcal{B}$. We define the loss function $\ell : \mathcal{B} \times \mathcal{Y} \to \mathbb{R}$ that evaluates the prediction performance. The learning task is to minimise the corresponding *excess risk* for its label Y_t' defined as

$$\mathcal{R}(b) := \mathbb{E}_{D_s^m, D_t^{\mathrm{U},n}, X_t', Y_t'} \left[\ell\left(b, Y_t'\right) - \ell(b^*, Y_t') \right], \tag{10}$$

where the expectation is taken with respect to all the source and target data, and b^* is the optimal predictor that can depend on the true distribution of the data. Particularly, we will also examine the excess risk under the condition $P_S(X,Y) = P_T(X,Y)$, commonly known as *semi-supervised learning*.

4.1 Causal Settings

In this section, we introduce the concept of causality within a supervised learning context involving feature variable X and label variable Y. Here we take an approach by establishing the learning model based on the parametric data distributions. We focus on scenarios where there are no other con-founders but only variables X and Y. Assume X is drawn from a finite set $\mathcal{X} = \{x_1, x_2, \ldots, x_k\}$ with k elements and the corresponding label Y is drawn from a finite set $\mathcal{Y} = \{y_1, y_2, \ldots, y_{k'}\}$ with k' elements. We then construct the parametric models under causal settings by specifying the joint distribution of X and Y as follows:

Definition 1 (Causal Settings) *We define two distinct learning settings based on the direction of causality for $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ using the following generation process:*

- ***Causal learning (Figure 1(a))*** *We say that "X causes Y" (denoted as $X \to Y$) if the pair (X,Y) is generated as follows: X is firstly generated according to the distribution P_{θ_X}. Given $X = x$, Y is generated from the distribution $P_{\theta_{Y_x}}$. This implies that the joint distribution of (X,Y) is given by*

$$P(x,y) = P_{\theta_X}(x) P_{\theta_{Y_x}}(y). \tag{11}$$

  We call a learning problem "causal learning" if the underlying causal mechanism satisfies $X \to Y$.

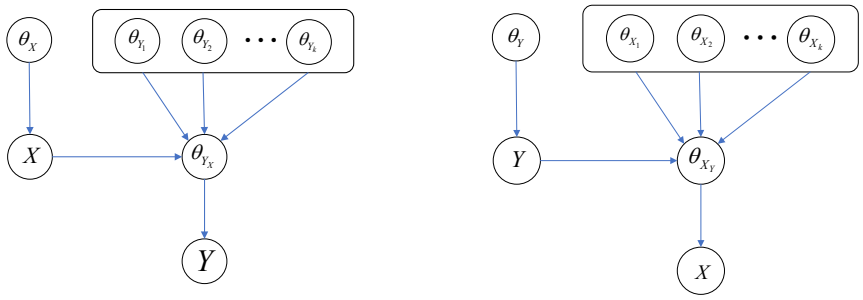
- ***Anti-causal learning (Figure 1(b))*** *We say that "Y causes X" (denoted as $Y \to X$) if the pair (X,Y) is generated as follows: Y is firstly generated according to the distribution P_{θ_Y}. Given $Y = y$, X is generated from the distribution $P_{\theta_{X_y}}$. This implies that the joint distribution of (X,Y) is given by*

$$P(x,y) = P_{\theta_Y}(y) P_{\theta_{X_y}}(x). \tag{12}$$

  We call a learning problem "anti-causal learning" if the underlying causal mechanism satisfies $Y \to X$.

These learning scenarios are conceptualized through parametric data generation mechanisms and sketched in Figure 1. When considering the causal setting $X \to Y$, we assume that X is drawn from the distribution P_{θ_X} and when we see a realization x_i of the random variable X, the distribution of the outcome variable Y is then characterized by a distinct parameter $\theta_{Y_{x_i}}$, and the observed outcome y is assumed to be drawn from the distribution $P_{\theta_{Y_{x_i}}}$. The double subscript notation is intentionally used to emphasize that the parameters $\theta_{Y_{x_i}}$ describe the distribution of Y, which is directly associated with the specific values of x_i. This framework inherently incorporates the concept of the "soft" intervention that alters the conditional probability distributions of the variables being intervened upon(Eberhardt and Scheines, 2007; Pearl, 2009, 1998; Imbens and Rubin, 2015), which is a fundamental concept in the

study of causality. By firstly setting $X = x_i$, we effectively intervene in the system, which allows for the direct examination of its impact on Y for different interventions. Hence, the model not only captures the association between X and Y but also provides a structured way to explore causal effects through interventions. For the anti-causal setting $Y \to X$, the procedure is analogous: the distribution of Y is defined by a parameter θ_Y^*, and upon intervening to set Y to y_i, the distribution of X is specified by the parameter $\theta_{X_{y_i}}$, from which we observe x through the distribution $P_{\theta_{X_{y_i}}}(X)$. In Definition 1, we assume that both X and Y are discrete variables for simplicity. However, it is important to note that our results also apply to cases with discrete causes and continuous effects.



(a) Causal Learning: $X \to Y$ (b) Anti-causal Learning: $Y \to X$

Figure 1: Causal settings for $X \to Y$ in (a) and $Y \to X$ in (b). We refer to the scenario in (a) as the "causal learning" setting because the direction of causation aligns with the direction of prediction, whereas the scenario in (b) is termed the "anti-causal learning" setting since the direction of causation is opposite to the direction of prediction.

We draw the diagram in Figure 1 to visualize the parametric models under these two different mechanisms. The models in Figures 1(a) and 1(b) are called "causal learning" and "anti-causal learning" respectively (Schölkopf et al., 2012), to mirror the causation direction in alignment with the prediction direction. In causal learning, the prediction direction coincides with the causation direction, whereas in anti-causal learning, the causation direction opposes the prediction direction.

Remark 2 *As we will show in the sequel, the causal structure of the data-generating process can be leveraged to enhance the prediction performance, which cannot be achieved by using the knowledge of the observational distribution of* (X, Y) *alone. Roughly speaking, under certain regularity conditions, we could learn the parameters* $(\theta_X, \theta_{Y_x}$*, etc.) directly from the unlabelled data, thus improving the prediction performance. As for the labelled source data, they can be partially profitable if the target domain shares some distribution parameters with the source domain. We will support these intuitions with our theoretical analysis in Section 5.*

Remark 3 *We make the following remarks regarding the definitions of the above settings.*

- *For simplicity, we will use the notation Y_x to denote a random variable if it is generated according to the distribution $P_{\theta_{Y_x}}(y)$ in the causal learning setting. Similarly, X_y denotes a random variable generated according to the distribution $P_{\theta_{X_y}}(x)$ in the anti-causal learning setting. More generally, we define a random variable Y_X if it is drawn from a random distribution $P_{\theta_{Y_X}}(y)$ induced by the random variable X. This notation also suggests an equivalent way of expressing the causality. Namely, we have $Y = \sum_{i=1}^{k} \mathbf{1}_{X=x_i} Y_{x_i}$ for the causal setting where X is generated according to P_{θ_X}, and $X = \sum_{i=1}^{k} \mathbf{1}_{Y=y_i} X_{y_i}$ for the anti-causal setting where Y is generated according to P_{θ_y}. This notation is consistent with the notations used in (Hernán and Robins, 2010; Imbens and Rubin, 2015; Cabreros and Storey, 2019), where the concept of potential outcome is used.*

- *Figure 1 suggests that the random variables X and $Y_{x_1}, Y_{x_2}, \ldots, Y_{x_k}$ are mutually* independent *in the causal settings. Similarly, $Y, X_{y_1}, \ldots, X_{y_k}$ are also mutually independent in the anti-causal learning setting.*

- *The causal setting outlined can also be specialized to parametric structural causal models as outlined by Hernán and Robins (2010); Pearl and Mackenzie (2018), which takes the form of the relationship $X' \to Y'$ by*

$$X' := N_X, \quad Y' := f(N_Y, X').$$

Here, f is a function that defines the parametric distributions of Y', with N_Y and N_X being independent random variables. This setup allows us to parameterize the distribution of X with N_X by identifying P_{θ_X} with P where P is the distribution of N_X. By setting $Y_{x_k} = f(x_k, N_Y)$, we could then model the distribution of the outcome by $P_{\theta_{Y_{x_k}}}(Y)$ where the parameters depend on the function f, N_Y and x_k. Then we could express Y as a sum over potential outcomes of X, represented as $Y = \sum_{i=1}^{k} \mathbf{1}_{X=x_i} f(x_i, N_Y)$, which simplifies to $Y = f(N_Y, X)$.

For the following discussion and main results, we assume that the causal relationship between X and Y is always unique, e.g., the causal direction is acyclic. Initially, we also assume the relationship is known for the theoretical analysis. In later parts of this discussion, we will also examine the case in which the causal direction of the underlying causal direction is unknown.

4.2 Parametric Models

When studying domain adaptation, we have two sets of random variables (X_s, Y_s) and (X_t, Y_t), where the former denotes the feature and label in the source domain and the latter for the target domain. We will consider two causal settings. The first one is given by $X_s \to Y_s$ and $X_t \to Y_t$ with the definition of causation given in Figure 1(a), namely the adaptation with the *causal learning* setting. We assume X_s, X_t take value in $\{x_1, x_2, \cdots, x_k\}$ and Y_s, Y_t take values in $\mathcal{Y}$, which could be either a continuous or discrete space. We will focus on parametric models in this work, and more precisely, the source distribution (similarly to target distribution) P_{X_s} is parameterized by a parameter θ_X^{s*} and the distributions of the outcome random variables $P_{Y_{x_i}}$ are also parameterized by the parameters $\theta_{Y_{x_i}}^{s*}$ for all

$i = 1, \ldots, k$. Then the joint distribution of the data pair (X_s, Y_s) and (X_t, Y_t) can be formulated as,

$$P_{\theta_s^*}(x_s, y_s) = P_{\theta_X^{s*}}(x_s) P_{\theta_{Y_{x_s}}^{s*}}(y_s), \tag{13}$$

$$P_{\theta_t^*}(x_t, y_t) = P_{\theta_X^{t*}}(x_t) P_{\theta_{Y_{x_t}}^{t*}}(y_t), \tag{14}$$

where we use θ_s^* and θ_t^* to encapsulate all the parameters:

$$\theta_s^* = (\theta_X^{s*}, \theta_{Y_{x_1}}^{s*}, \cdots, \theta_{Y_{x_k}}^{s*}) \in \Lambda, \tag{15}$$

$$\theta_t^* = (\theta_X^{t*}, \theta_{Y_{x_1}}^{t*}, \cdots, \theta_{Y_{x_k}}^{t*}) \in \Lambda. \tag{16}$$

For simplicity, we assume that every element in both θ_s^* and θ_t^* is a scalar in $\mathbb{R}$ and $\Lambda \subseteq \mathbb{R}^{k+1}$ is a closed set endowed with Lebesgue measure. In the sequel, we write $P_S(X) = P_T(X)$ (similarly for $P_S(Y|X)$) with the understanding that their underlying parameters are elementwise equal (e.g., $\theta_X^{s*} = \theta_X^{t*}$) and vice versa.

The second learning model we consider in this work is given by $Y_s \to X_s$ and $Y_t \to X_t$, where X_{s,y_i} and X_{t,y_i} denote the random outcomes given the treatment y_i in source and target domains, namely the adaptation with the *anti-causal learning* setting. The parameterization, in this case, is analogous to causal learning by regarding Y as a cause and X as an effect. Instead, we now assume Y_s, Y_t take value in $\{y_1, y_2, \cdots, y_{k'}\}$ and X_s, X_t take values in a continuous or discrete space $\mathcal{X}$ for the anti-causal learning. Similarly to the causal learning, we assume Y_s and Y_t are parameterized by θ_Y^{s*} and θ_Y^{t*}, and X_{s,y_i} and X_{t,y_i} are parameterized by $\theta_{X_{y_i}}^{s*}$ and $\theta_{X_{y_i}}^{t*}$ for all $i = 1, \cdots, k'$, and we use the same notation θ_s^* and θ_t^* to encapsulate all the parameters and every parameter in both θ_s^* and θ_t^* is a scalar in $\mathbb{R}$ and $\Lambda \subseteq \mathbb{R}^{k'+1}$ is a closed set endowed with Lebesgue measure.

Under causal learning ($X \to Y$), it can be seen that the unlabelled target data are generated only with θ_X^{t*} and thus do not contain knowledge about $\theta_{Y_{x_i}}^{t*}$ as they are statistically independent. Intuitively speaking, the parameters associated with $P(Y|X)$ in the target domain cannot be accurately estimated exclusively from the unlabelled data. However, under anti-causal learning ($Y \to X$), the unlabelled target data are associated with all parameters θ_Y^{t*} and $\theta_{X_{y_i}}^{t*}$ that induce the labelling distribution $P(Y|X)$ in the target domain. In addition, we make the following assumption for the data distributions in both causal settings.

Assumption 1 (Parametric IID data) *We assume the labelled source and unlabelled target samples are generated independently and identically under both causal learning and anti-causal learning. More precisely, the joint distribution of the data sequence pairs $P_{\theta_s^*, \theta_t^*}(D_t^{U,n}, D_s^m)$ can be written as*

$$P_{\theta_s^*, \theta_t^*}(D_t^{\mathrm{U},n}, D_s^m) = \prod_{i=1}^{n} P_{\theta_t^*}(X_t^{(i)}) \prod_{j=1}^{m} P_{\theta_s^*}(X_s^{(j)}, Y_s^{(j)}),$$

where $P_{\theta_t^*}(X_t^{(i)})$ is the marginal of $P_{\theta_t^*}(X_t^{(i)}, Y_t^{(i)})$. We also assume θ_t^* and θ_s^* are points in the interior of Λ. Furthermore, in both models, the parametric families for the cause and effect are assumed to be known in advance.

Based on the models defined above, the excess risk in Equation (10) can be written as

$$\begin{aligned}\mathcal{R}(b) &:= \mathbb{E}_{P_{\theta_s^*}(D_s^m)P_{\theta_t^*}(D_t^{\mathrm{U},n},X_t',Y_t')}\left[\ell\left(b,Y_t'\right)-\ell(b^*,Y_t')\right] \\ &= \mathbb{E}_{\theta_s,\theta_t}\left[\ell\left(b,Y_t'\right)-\ell(b^*,Y_t')\right] \end{aligned} \tag{17}$$

For simplicity, we use the notation $\mathbb{E}_{\theta_s,\theta_t}[\cdot]$ (similarly, $\mathbb{E}_{\theta_t}[\cdot]$ and $\mathbb{E}_{\theta_s}[\cdot]$) to denote the expectation taken over all source and target samples drawn from P_{θ_s} and P_{θ_t}.

5. Main Results

In this section, we will examine the excess risk for causal and anti-causal learning under various conditions of distribution shift, e.g., covariate shift (Gretton et al., 2009), target shift (Zhang et al., 2013), concept drift (Cai and Wei, 2021), etc.

Before diving into the details, we informally outline our main results in Table 1 under log-loss. Recall that in both causal and anti-causal learning, the goal is to learn the conditional distribution $P_T(Y|X)$ such that the label Y can be predicted from the feature X in the target domain. In causal learning, this corresponds to learning the outcome random variables Y_{t,x_i}. However, the unlabelled target data X_t ("cause" in this case) do not contain information about Y_{t,x_i} as they are independent under causal generating processes. Therefore, the unlabelled target data are not useful in the causal learning case, as indicated in the table. The usefulness of the source data depends on the causal settings. When the labelling distribution is invariant across two domains (e.g., $P_S(Y|X) = P_T(Y|X)$), the source data help reduce excess risk by providing information about Y_{t,x_i}, which is identical to Y_{s,x_i}. The learning rate is then shown to be $O(\frac{k}{m})$, where k is the number of parameters and m is the size of the source sample. On the other hand, if $P_S(Y|X) \neq P_T(Y|X)$, the source data generally do not provide information about Y_{t,x_i} and the excess risk does not converge to zero even with sufficient source and target data.

In anti-causal learning scenario ($Y \to X$, $P_S(X,Y) \neq P_T(X,Y)$), however, learning $P_T(Y|X)$ requires to estimate all the parameters of Y_t and X_{t,y_i}. Unlike causal learning, where $P_T(Y|X)$ is fully represented by the random outcome variables Y_{t,x_i}, in this case, we need to infer $P_T(Y|X)$ from the joint distribution $P_T(X,Y)$. We will show that the unlabelled target data is always useful in anti-causal learning under certain conditions. The source data can also contribute to learning, depending on the assumptions we have made about the distribution shift. For example, if $P_S(Y) \neq P_T(Y)$ and $P_S(X|Y) \neq P_T(X|Y)$ with the independence assumption, there is no reason for the source data to be useful for prediction in the target domain. Therefore, the rate, in this case, is $O(\frac{k'+1}{n})$, which solely depends on the number of unlabelled target data. Intuitively, this is the cost of learning $k'+1$ parameters with n unlabelled target samples. Under the target shift condition ($P_S(Y) \neq P_T(Y)$ and $P_S(X|Y) = P_T(X|Y)$), the source data helps in learning the outcome variables $X_{y_i}, i = 1,\ldots,k'$, which is evinced in the rate $O(\frac{1}{n} + \frac{k'}{m+n})$ that constitutes the learning of Y_t (with associated parameter θ_Y^{t*}) with a rate $O(\frac{1}{n})$ and $X_{t,y_i}, i = 1,\ldots,k'$ (with associated parameters $\theta_{X_{y_i}}^{t*}$) with a rate $O(\frac{k'}{n+m})$. Similarly, for the conditional shift ($P_S(Y) = P_T(Y)$ and $P_S(X|Y) \neq P_T(X|Y)$), the rate becomes $O(\frac{k'}{n} + \frac{1}{m+n})$ where sufficient source data boosts the learning of Y_t (associated with parameter θ_Y^{t*}) with a rate $O(\frac{1}{m+n})$, but are not helpful for learning outcomes variables X_{t,y_i}.

Causal Setting	Conditions	UT	LS	Rate
$X \to Y$	$P_S(X) \neq P_T(X)$, $P_S(Y\|X) \neq P_T(Y\|X)$	✗	✗	-
	$P_S(X) \neq P_T(X)$, $P_S(Y\|X) = P_T(Y\|X)$	✗	✓	$O(\frac{k}{m})$
	$P_S(X) = P_T(X)$, $P_S(Y\|X) \neq P_T(Y\|X)$	✗	✗	-
	$P_S(X) = P_T(X)$, $P_S(Y\|X) = P_T(Y\|X)$	✗	✓	$O(\frac{k}{m})$
$Y \to X$	$P_S(Y) \neq P_T(Y)$, $P_S(X\|Y) \neq P_T(X\|Y)$	✓	✗	$O(\frac{1+k'}{n})$
	$P_S(Y) \neq P_T(Y)$, $P_S(X\|Y) = P_T(X\|Y)$	✓	✓	$O(\frac{1}{n} + \frac{k'}{n+m})$
	$P_S(Y) = P_T(Y)$, $P_S(X\|Y) \neq P_T(X\|Y)$	✓	✓	$O(\frac{k'}{n} + \frac{1}{n+m})$
	$P_S(Y) = P_T(Y)$, $P_S(X\|Y) = P_T(X\|Y)$	✓	✓	$O(\frac{k'+1}{m+n})$

Table 1: (Informal) results on the effectiveness of source and unlabelled target data under causal and anti-causal learning problems."✓" and "✗" marks indicate whether the data are useful or not for the prediction under specific conditions and causal settings. "UT" and "LS" are abbreviated for "Unlabelled Target" and "labelled Source", respectively. The rate illustrates the convergence for the excess risk under log-loss in terms of the target sample size n and source sample size m. The "-" sign in the rate column means the risk will not converge to zero even if we have sufficient source and target data.

As a special case of domain adaptation, we also consider SSL where $P_S(X,Y) = P_T(X,Y)$. Using the same arguments in causal and anti-causal settings, we obtain a better rate of $O(\frac{k'+1}{m+n})$ in anti-causal learning, where the unlabelled target data take effect on prediction, compared to $O(\frac{k}{m})$ in causal learning, where the unlabelled target data are not helpful. For readers interested in empirical verification of our results, we substantiate the analysis with a toy example, which can be found in Section 6. More generally, our analysis also holds for the case when the cause is discrete and the effect can be either discrete or continuous. This is practically useful since the datasets in many real classification problems are usually anti-causal with a finite label space $\mathcal{Y}$ where the feature space is usually continuous (Schölkopf et al., 2012; Zhang et al., 2013; Gong et al., 2016). To summarize, different causation directions incentivize different learning complexity for generalization, which is reflected in the number of model parameters and the effectiveness of the data. It comes naturally when we could model both the source and target data from either $X \to Y$ or $Y \to X$ in some non-identifiable circumstances, we need to take the distribution shift conditions and sample sizes into account to achieve better learning performance. We will first show our main proof techniques in Section 5.1 and examples are followed in Section 6.

Many theoretical results on generalization in domain adaptation depend on distributional conditions and algorithms. Notably, based on the covariate shift condition, Kpotufe and Martinet (2018) propose the "transfer component" that evaluates the support overlap between the source and target domains and derives the minimax rate for the generalization error. However, such a notion cannot be generally applied to other distribution shift conditions. Similarly, Cai and Wei (2021) determine the optimal minimax rate of convergence with the weighted k-nearest neighbour classifier using the notion of "relative signal exponent" based on the concept drift condition. Under the target shift condition, Maity et al. (2022) and Gong et al. (2016) derive the learning guarantees for the distribution reweighting strategies,

which are algorithm-dependent. While our analysis is restricted to parametric models, it applies to all possible distribution shift conditions. This applicability facilitates a unified framework for assessing learning performance from a causal viewpoint. It also offers an intuitive understanding of the values derived from source and target data. In particular, our result of the covariate shift condition offers the same insight when $P_T(X)$ is absolutely continuous w.r.t. $P_S(X)$ in Kpotufe and Martinet (2018), where the labelled source has the same value as the labelled target. The target shift result agrees with Maity et al. (2022) in the sense that the unlabelled target is equally useful as the labelled target data, achieving a rate of $O(\frac{1}{n})$. Under the concept drift condition, we argue that the excess risk does not converge, which is consistent with Theorem 3.1 in Cai and Wei (2021) for a large relative signal exponent and no labelled target data. Moreover, we prove in Lemma 13 that the excess risk is minimax optimal under log-loss.

5.1 Information-theoretic Characterization

In this section, we will outline our primary proof techniques for the findings presented in Table 1. Our proofs primarily build upon the work of Merhav and Feder (1998) and Zhu (2020), which originally focused on the sequential learning problem or semi-supervised learning problem. However, we extend their results by applying the mixture strategy to the UDA and SSL problems with the information-theoretic framework. To begin with, we first consider the *log-loss* (also known as the logarithmic loss), which is formally defined as follows.

Definition 4 (Log-loss) *Let the predictor b be a probability distribution over the target label Y_t'. The log-loss is then defined as,*

$$\ell(b, Y_t') = -\log b(Y_t'). \tag{18}$$

Given the testing feature X_t', training data D_s^m and $D_t^{U,n}$, we may view the predictor b as the conditional distribution $Q(Y_t'|D_t^{\mathrm{U},n}, D_s^m, X_t')$ over the unseen target label given the testing feature X_t' and the training data $D_s^m, D_t^{\mathrm{U},n}$. It could be proved that the true predictor b^* is given by the underlying target distribution as $b^*(Y_t') = P_{\theta_t^*}(Y_t'|X_t')$. Then the excess risk can be expressed as,

$$\mathcal{R}(b) = \mathbb{E}_{\theta_t^*, \theta_s^*}\left[\log \frac{P_{\theta_t^*}(Y_t'|X_t')}{Q(Y_t'|D_t^{\mathrm{U},n}, D_s^m, X_t')}\right] \tag{19}$$

Concerning the choice of the predictor $Q(Y_t'|D_t^{\mathrm{U},n}, D_s^m, X_t')$, we first define Θ_s and Θ_t as random vectors over Λ, which can be interpreted as a random guess of θ_s^* and θ_t^*. Note that Θ_s and Θ_t may share some common parameters, e.g., $\Theta_{s,i} = \Theta_{t,i}$ for ith entry. Then by *mixture strategy* (Merhav and Feder, 1998; Xie and Barron, 2000), we assign a probability distribution ω over Θ_s and Θ_t w.r.t. the Lebesgue measure to represent our prior knowledge and update the posterior with the incoming data to approximate the underlying distributions.

That is,

$$Q(Y_t'|D_t^{\mathrm{U},n}, D_s^m, X_t') = \frac{\int P_{\theta_t}(D_t^{\mathrm{U},n}, X_t', Y_t') P_{\theta_s}(D_s^m)\omega(\theta_t,\theta_s)d\theta_t d\theta_s}{\int P_{\theta_t}(X_t') P_{\theta_t}(D_t^{\mathrm{U},n}) P_{\theta_s}(D_s^m)\omega(\theta_t,\theta_s)d\theta_t d\theta_s}$$
$$= \int P_{\theta_t}(Y_t'|X_t') P(\theta_t,\theta_s|X_t', D_s^m, D_s^{\mathrm{U},n}) d\theta_s d\theta_t. \tag{20}$$

We can interpret (20) as estimating Y' in a two-step procedure. With a prior distribution ω, the first step is to learn the parameters θ_s, θ_t with the joint posterior $P(\theta_s, \theta_t|D_s^m, D_t^{\mathrm{U},n}, X_t')$. In the second step, the learned θ_t is applied for prediction in terms of the parametric distribution $P_{\theta_t}(Y_t'|X_t')$. One way to comprehend the mixture strategy is that we encode our prior knowledge over target and source domain distributions in terms of the prior distribution $\omega(\Theta_s, \Theta_t)$, and different distribution shift conditions correspond to different priors. By the mixture strategy, we give the excess risk under log-loss.

Theorem 5 (Excess Risk with Log-loss) *Under log-loss, let the predictor Q be the distribution in (20) with the prior distribution $\omega(\Theta_s, \Theta_t)$. Then the excess risk can be expressed as*

$$\mathcal{R}(b) = I(Y_t'; \theta_t^*, \theta_s^*|D_s^m, D_t^{\mathrm{U},n}, X_t'), \tag{21}$$

where the R.H.S. denotes the conditional mutual information $I(Y_t'; \Theta_t, \Theta_s|D_s^m, D_t^{\mathrm{U},n}, X_t')$ evaluated at $\Theta_t = \theta_t^*$ and $\Theta_s = \theta_s^*$.

All proofs in this paper can be found in the Appendix. A similar learning strategy can be used for more general loss functions. Given a general loss function ℓ, we define the predictor b as

$$b = \mathrm{argmin}_b \mathbb{E}_Q \left[\ell(b, Y_t')|X_t', D_t^{U,n}, D_s^m\right], \tag{22}$$

with the choice of the mixture strategy

$$Q(X_t', Y_t', D_t^n, D_s^m) = \int P_{\theta_t,\theta_s}(X_t', Y_t', D_t^{U,n}, D_s^m)\omega(\theta_t,\theta_s)d\theta_t d\theta_s$$

for some prior ω. The optimal predictor is then given by

$$b^* = \mathrm{argmin}_b \mathbb{E}_{\theta_t^*} \left[\ell(b, Y_t')|D_t^{U,n}, X_t'.\right] \tag{23}$$

We have the following theorem for β-exponential concave loss functions as follows.

Theorem 6 (Excess Risk with Exponential Concave Loss) *Assume the loss function is β-exponentially concave of b for any y. Then the excess risk induced by b and b^* in Equation (22) and (23) can be bounded as*

$$\mathcal{R}(b) \leq \frac{1}{\beta} I(Y_t'; \theta_t^*, \theta_s^*|D_s^m, D_t^{\mathrm{U},n}, X_t'). \tag{24}$$

The log-loss can be regarded as a special case with $\beta = 1$. One can refer to Lemma 1 (also the proof) in Zhu (2020) for more details and comments, which we will not repeat in our context. Likewise, if the loss function is bounded, we arrive at the following theorem.

Theorem 7 (Excess Risk with Bounded Loss) *Assume the loss function satisfies* $|\ell(b, y) - \ell(b^*, y)| \leq M$ *for any observation* y *and any two predictors* b, b^**. Then the excess risk can be bounded as*

$$\mathcal{R}(b) \leq M\sqrt{2I(Y_t'; \theta_t^*, \theta_s^* | D_s^m, D_t^{\mathrm{U},n}, X_t')}. \tag{25}$$

From the above theorems, we can see the analogy that the expected regrets induced by the mixture strategy are both characterized by CMI evaluated at θ_t^* and θ_s^*. Note that these results apply to both causal and anti-causal learning problems. Nevertheless, the characterization of learning performance in its present form is less informative because it does not show the effect of sample sizes and causal directions. To this end, we make some regularity assumptions on the parametric conditions (Clarke and Barron, 1990; Merhav and Feder, 1998; Zhu, 2020) and define the proper prior distribution to obtain an asymptotic approximation.

Assumption 2 (Parametric Distribution Conditions) *With the aforementioned parameterization, let* $\theta^* = (\theta_s^*, \theta_t^*)$ *denote the underlying parameters for labelled source and unlabelled target data. We assume:*

- ***Condition 1:*** *The source and target distributions* $P_{\theta_s}(X_s, Y_s)$ *and* $P_{\theta_t}(X_t)$ *is twice continuously differentiable at* θ_s^* *and* θ_t^* *for almost every* (X_s, Y_s) *and* X_t*.*

- ***Condition 2:*** *Define the Fisher information matrix*

$$\begin{aligned} I_s &= -\mathbb{E}_{\theta_s^*}[\nabla^2 \log P(X_s, Y_s | \theta_s^*)], \\ I_t &= -\mathbb{E}_{\theta_t^*}[\nabla^2 \log P(X_t | \theta_t^*)], \\ I_0 &= -\mathbb{E}_{\theta^*}[\nabla^2 \log P(X_t, X_s, Y_s | \theta^*)]. \end{aligned}$$

  We assume I_s *and* I_t *are positive definite and it holds that* I_0 *is also positive definite.*

- ***Condition 3*** *(Clarke and Barron, 1990): Assume that the convergence of a sequence of parameter values is equivalent to the weak convergence of the distributions they index. Particularly:*

$$\begin{aligned} \theta_s \to \theta_s^* &\Leftrightarrow P_{\theta_s}(X, Y) \to P_{\theta_s^*}(X, Y), \\ \theta_t \to \theta_t^* &\Leftrightarrow P_{\theta_t}(X) \to P_{\theta_t^*}(X), \end{aligned}$$

  for source and target domains, respectively.

- ***Condition 4****: Assume that for all* θ_s *in some neighbourhood of* θ_s^* *and* θ_t *in some neighbourhood of* θ_t^**, the normalized Rényi divergences of order* $1 + \lambda$*, the following holds*

$$\log \int P_{\theta_s^*}(x, y)^{1+\lambda} P_{\theta_s}(x, y)^{-\lambda} dx dy < \infty, \tag{26}$$

$$\log \int P_{\theta_t^*}(x)^{1+\lambda} P_{\theta_t}(x)^{-\lambda} dx < \infty \tag{27}$$

  for sufficiently small $\lambda > 0$*.*

- ***Condition 5***: *Assume that for all θ_s in some neighbourhood of θ_s^* and θ_t in some neighbourhood of θ_t^*, the moment generating function is bounded as*

$$\mathbb{E}_{\theta_s^*}\left[e^{\lambda \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(X_s, Y_s|\theta_s)}\right] < \infty, \tag{28}$$

$$\mathbb{E}_{\theta_t^*}\left[e^{\lambda \frac{\partial^2}{\partial\theta_j \partial\theta_k} \log p(X_t|\theta_t)}\right] < \infty, \tag{29}$$

for all $j, k = 1, \ldots, d$ with some small $\lambda > 0$, where d is determined based on the causal settings and conditional shifting conditions.

- ***Condition 6:*** *Let $l_s := [(\nabla \log p\,(X, Y \mid \theta_s^*)), \mathbf{0}_{d''}]^T$, $l_t := [\mathbf{0}_{d''}, \nabla \log p\,(X \mid \theta_t^*)]^T$, where $\mathbf{0}_{d''}$ denotes the zero vector with length d'', and d'' denotes the number of distribution parameters for both source and target domains. We also define l_s', l_t' as an independent copy of l_s and l_t, respectively. We assume the moment-generating functions*

$$\mathbb{E}\left[e^{\lambda l_s^T I_0 l_s}\right], \mathbb{E}\left[e^{\lambda l_s^T I_0 l_s'}\right], \mathbb{E}\left[e^{\lambda l_t^T I_0 l_t}\right],$$
$$\mathbb{E}\left[e^{\lambda l_t^T I_0 l_t'}\right], \mathbb{E}\left[e^{\lambda l_t^T I_0 l_s}\right]$$

exist for some small enough $\lambda > 0$.

Assumption 3 (Proper Prior) *We assume that the prior distribution $\omega(\Theta_s, \Theta_t)$ is continuous and positive over its whole support.*

Remark 8 *We impose the first three conditions on parametric distributions with the proper prior distribution to ensure that the posterior distribution of Θ_t and Θ_s asymptotically concentrates on neighbourhoods of θ_t^* and θ_s^* under both causal settings given sufficient source and target data. In particular, the positive definite Fisher information matrix and parameter uniqueness assumption imply that θ_t^* and θ_s^* are identifiable within Λ. We also impose some technical conditions to ensure that the posterior of the parameters converges to their true values at an appropriate rate. Additionally, for the anti-causal setting $Y \to X$, we exclude the case when outcome variable X has the same distribution for all y_i with Condition 2, that is, $P_{\theta_{X_{y_i}}}(X)$ is identical for all $y_i \in \mathcal{Y}$. Because in this case, X and Y are effectively independent, and the fisher information I_t is no longer positive definite as the distribution of X no longer depends on the parameter θ_Y^*.*

Remark 9 *The last three technical conditions are adopted and modified from Zhu (2020) to ensure that the posterior of the parameters converges to their true values at an appropriate rate for both source and target domains. We will mainly use these conditions for asymptotic estimation of KL divergence, e.g., see proof of Lemma 14.*

Remark 10 *Though asymptotically, the prior distribution does not affect the learning rate, its choice is crucial in practice, particularly with limited data. Priors should be selected based on parameter understanding, model complexity, and existing knowledge. For simple parametric models such as generalized linear models, we can adopt conjugate priors (Diaconis*

and Ylvisaker, 1979; Chen and Ibrahim, 2003) for updating parameters easily. For more complex models, we may require non-conjugate priors where the data are used to estimate the parameters of the prior distribution (Efron, 2012; Carlin and Louis, 2008). This is particularly useful when we have little prior knowledge about the distribution. In practice, the sensitivity analysis could also be conducted to assess the robustness of the posterior distribution to the choice of prior. This helps ensure that the posterior is not unduly influenced by the choice of prior.

5.2 Excess Risk in Causal Learning

In this section, we will characterize the excess risk asymptotically under causal learning. We first consider the learning scenario when $P_S(Y|X) = P_T(Y|X)$, which corresponds to SSL if $P_T(X) = P_S(X)$ and covariate shift regime otherwise. The random vector Θ_s and Θ_t can be explicitly written as

$$\Theta_s = (\Theta^s_X, \Theta^s_{Y_{x_1}}, \cdots, \Theta^s_{Y_{x_k}}) = (\Theta^s_X, \Theta^s_{Y_X}), \tag{30}$$

$$\Theta_t = (\Theta^t_X, \Theta^t_{Y_{x_1}}, \cdots, \Theta^t_{Y_{x_k}}) = (\Theta^t_X, \Theta^t_{Y_X}), \tag{31}$$

where $\Theta_{Y_X} = (\Theta_{Y_{x_1}}, \cdots, \Theta_{Y_{x_k}})$ for succinctness. We assume Θ^t_X and Θ^s_X are independent of $\Theta^s_{Y_X}$ and $\Theta^t_{Y_X}$, but we will keep $\Theta^s_{Y_X}$ and $\Theta^t_{Y_X}$ identical according to the assumption $P_S(Y|X) = P_T(Y|X)$, written as $\Theta^{st}_{Y_X}$. With the proper prior distribution, we simplify the mixture distribution Q as follows by omitting the unlabelled target data as follows:

$$Q(Y'_t|D_t^{\mathrm{U},n}, D_s^m, X'_t) = \int P(Y'_t|\theta^{st}_{Y_X}, X'_t) P(\theta^{st}_{Y_X}|D_s^m) d\theta^{st}_{Y_X},$$

where the knowledge transfer depends on the conditional posterior $P(\theta^{st}_{Y_X}|D_s^m)$. Since $P_S(Y|X) = P_T(Y|X)$, without any labels from the target domain, we can only learn the parameters of the random outcomes Y_X from the source data. On the other hand, if the assumption $P_S(Y|X) = P_T(Y|X)$ does not hold, namely, the concept drift if $P_S(X) = P_T(X)$ and general shift condition otherwise, the mixture strategy in (20) becomes

$$Q(Y'_t|D_t^{\mathrm{U},n}, D_s^m, X'_t) = \int P_{\theta^t_{Y_{X'_t}}}(Y'_t) \omega(\theta^t_{Y_{X'_t}}) d\theta^t_{Y_{X'_t}}$$

due to the mutual independence properties of the distribution parameters. In this case, neither the unlabelled target data nor the source data are useful for the estimation, the prediction is only piloted by the prior distribution $\omega(\theta^t_{Y_X})$ as the initial estimate for $\theta^{t*}_{Y_X}$. As a result, the excess risk, in this case, does not go to zero even if we have enough source and target data. To formally state the idea, we give the asymptotic estimation in the following main theorem.

Theorem 11 (Excess Risk with Causal Learning) *In addition to Assumption 1,2 and 3, we also assume that X causes Y in both source and target domains. Let Θ_s and Θ_t be parameterized in (30) and (31). As $m \to \infty$, the mixture strategy under log-loss yields:*

- *(General shift and Concept drift) For any $P_{\theta^{t*}_X}(X) \ll P_{\theta^{s*}_X}(X)$, if $P_S(Y|X) \neq P_T(Y|X)$:*

$$\mathcal{R}(b) = \mathbb{E}_{\theta^{t*}_X}[\mathrm{KL}(P_{\theta^{t*}_{Y_{X'_t}}}(Y'_t) \| Q(Y'_t|X'_t)], \tag{32}$$

where $Q(Y'_t|X'_t) = \int P_{\theta^t_{Y_{X'_t}}}(Y'_t)\omega(\theta^t_{Y_{X'_t}})d\theta^t_{Y_{X'_t}}$ *for a certain prior* ω *over* $\Theta^t_{Y_{X'_t}}$.

- *(covariate shift and SSL) For any* $P_{\theta^{t*}_X}(X) \ll P_{\theta^{s*}_X}(X)$*, if* $P_S(Y|X) = P_T(Y|X)$*:*

$$\mathcal{R}(b) \asymp \frac{k}{m}. \tag{33}$$

From the above theorem, it is clear that the target data are not useful without labels and n does not occur in the rate. This is understandable because such data do not contain information about $P(Y|X)$ due to the independence assumptions between X and Y_{x_i}. If the conditional distribution remains unchanged between source and target domains, the excess risk converges with the rate of $O(\frac{k}{m})$.

5.3 Excess Risk in Anti-Causal Learning

We now turn to the opposite causal direction where $Y \to X$. Similarly, we define the random variable Θ_s and Θ_t with the same form as (30) and (31) by

$$\Theta_s = (\Theta^s_Y, \Theta^s_{X_{y_1}}, \cdots, \Theta^s_{X_{y_{k'}}}) = (\Theta^s_Y, \Theta^s_{X_Y}), \tag{34}$$

$$\Theta_t = (\Theta^t_Y, \Theta^t_{X_{y_1}}, \cdots, \Theta^t_{X_{y_{k'}}}) = (\Theta^t_Y, \Theta^t_{X_Y}). \tag{35}$$

At this stage, we do not particularize any conditions on the parameters. From the Bayes rule, we rewrite the mixture distribution Q in terms of the above parameterization as

$$\begin{aligned} Q(Y'_t|D_t^{\mathrm{U},n}, D_s^m, X'_t) &= \frac{\int P_{\theta_t}(D_t^{\mathrm{U},n}, X'_t, Y'_t)P_{\theta_s}(D_s^m)\omega(\theta_t,\theta_s)d\theta_t d\theta_s}{\int P_{\theta_t}(X'_t)P_{\theta_t}(D_t^{\mathrm{U},n})P_{\theta_s}(D_s^m)\omega(\theta_t,\theta_s)d\theta_t d\theta_s} \\ &= \frac{\int P(Y'_t|\theta_t, X'_t)P(\theta_t|D_t^{\mathrm{U},n}, X'_t, \theta_s)d\theta_t P(\theta_s|D_s^m)d\theta_s}{\int P(\theta_t|D_t^{\mathrm{U},n}, X'_t, \theta_s)d\theta_t P(\theta_s|D_s^m)d\theta_s} \\ &= \int P(Y'_t|\theta_t, X'_t)P(\theta_t|D_t^{\mathrm{U},n}, X'_t, \theta_s)d\theta_t P(\theta_s|D_s^m)d\theta_s. \end{aligned}$$

To interpret, the mixture strategy first provides an estimate of θ_s from the source data, then knowledge is transferred from θ_s to θ_t with the prior distribution $\omega(\theta_t|\theta_s)$, which induces the posterior $P(\theta_t|X'_t, D_t^{U,n}, \theta_s)$ along with the features $X'_t, D_t^{U,n}$ in the target domain, since the unlabelled data may contain all the information of θ^*_t under the anti-causal parameterization. Eventually, the prediction of Y'_t will be based on the estimated θ_t and X'_t.

With condition 3 under Assumption 2, we require that the true parameters θ^*_t are identifiable given sufficient unlabelled target data, where its distribution is a mixture distribution, i.e., $\sum_{y\in\mathcal{Y}} P_{\theta_Y}(y)P_{\theta_{X_y}}(X)$. In general, this is a strong condition where the mixture distributions, such as the Bernoulli mixture, do not satisfy the assumption (Gyllenberg et al., 1994) and the parameters within their support are **not** identifiable. But for certain types of families, the parameters are identifiable up to *label swapping*, such as Gaussian (Teicher, 1963), exponential families (Barndorff-Nielsen, 1965), and many other finite continuous mixture distributions (McLachlan et al., 2019). Under label swapping, the posterior of the parameters approaches one of all permutations (Marin et al., 2005) and our result holds only

up to the permutation where we simply set θ^* to be the parameters for that permutation. To solve the label swapping problem, the methods proposed include the specification of parameterization constraints (Marin et al., 2005; McLachlan et al., 2019), a relabelling algorithm (Stephens, 2000), and constraint clustering (Grün and Leisch, 2009). Once the label swapping is addressed, the mixed distributions are identifiable (Titterington et al., 1985; McLachlan et al., 2019) and our results hold for estimating the corresponding θ^* as well. For illustration, we give a simple example of a categorical mixture distribution identifiable by adding structural constraints to the parameterization in Section 6. We will now consider different distribution shift scenarios under anti-causal learning and derive the corresponding asymptotic estimation for the excess risk.

Theorem 12 (Excess Risk with Anti-causal Learning) *In addition to Assumptions 1, 2 and 3, we also assume $Y \to X$ in both source and target domains. Let Θ_s and Θ_t be parameterized in (34) and (35). As $m \asymp n^p$ for some $p > 0$ and $n \to \infty$, the mixture strategy under log-loss yields:*

- *(General shift) If $P_S(Y) \neq P_T(Y)$, $P_S(X|Y) \neq P_T(X|Y)$,*

$$\mathcal{R}(b) \asymp \frac{1 + k'}{n}. \tag{36}$$

- *(Conditional shift) If $P_S(Y) = P_T(Y)$, $P_S(X|Y) \neq P_T(X|Y)$,*

$$\mathcal{R}(b) \asymp \frac{k'}{n} + \frac{1}{n \vee n^p}. \tag{37}$$

- *(Target shift) If $P_S(Y) \neq P_T(Y)$, $P_S(X|Y) = P_T(X|Y)$,*

$$\mathcal{R}(b) \asymp \frac{1}{n} + \frac{k'}{n \vee n^p}. \tag{38}$$

- *(SSL) If $P_S(Y) = P_T(Y)$, $P_S(X|Y) = P_T(X|Y)$,*

$$\mathcal{R}(b) \asymp \frac{k' + 1}{n \vee n^p}. \tag{39}$$

In contrast to causal learning, in the general shift case, we can achieve good generalization ability only with the unlabelled target data, while the source data do not help at all. This result confirms the value of unlabelled data, which is consistent with the intuition from Figure 1. In the conditional shift and target shift cases, we can further show that the source data can only help improve the excess risk from $O(\frac{k'+1}{n})$ to $O(\frac{k'+1-j}{n} + \frac{j}{n \vee n^p})$ depending on how many j common parameters θ_s^* and θ_t^* share. Intuitively, $O(\frac{k'+1-j}{n})$ can be viewed as the learning cost for $k' + 1 - j$ domain-specific parameters and $O(\frac{j}{n \vee n^p})$ as the learning cost for domain-sharing parameters. Therefore, the source data are incapable of changing the overall rate since the unlabelled target data always dominates the rate. In SSL, the rate $O(\frac{k'+1}{n \vee n^p})$ indicates that unlabelled target data are as useful as the labelled source data and that sufficient source data (e.g., $p > 1$) can indeed change the convergence rate. The results show that the *learning complexity* under different causal directions will vary. This

crucial distinction discloses how the causal relationships affect the model complexity and its generalization ability.

Our results in Theorem 5, 11, 12 establish the convergence rate for the mixture strategy. Here we show that this strategy is in fact optimal for log-loss.

Lemma 13 (Worst-Case Excess Risk) *For log-loss,*

$$\min_{b} \max_{\theta_s^*, \theta_t^*} \mathcal{R}(b) = \max_{\omega(\theta_s, \theta_t)} I(Y_t'; \Theta_t, \Theta_s | D_s^m, D_t^{\mathrm{U},n}, X_t'),$$

where (Θ_t, Θ_s) *is endowed with some prior distribution* ω*.*

This lemma exactly characterizes the excess risk for log-loss in the worst case. It shows that the worst-case regret is captured by the same CMI term as in Theorem 5, although maximized w.r.t. the prior distribution over the source and target parameters. However, it can be shown that the maximization does not change the convergence rate of the mutual information term (Clarke and Barron, 1994; Merhav and Feder, 1998). In other words, the convergence rate in Theorem 11, 12 is indeed optimal and cannot be improved using a different learning algorithm. Even though we only consider the log-loss in the previous analysis, the results can be extended straightforwardly in the case of other general loss functions, such as exponentially concave or bounded losses, where the excess risk is captured by the same CMI term in Theorem 5 (see Theorem 7 for bounded losses as an example).

6. Experiments

In this section, we begin by confirming our main results with a toy example, for which we elaborate on the case when the data can be modeled both as causal learning and anti-causal learning. Subsequently, we extend the idea to tackle real-world challenges like the classification of handwritten digits. For these scenarios, we parametrize the data distribution using the Gaussian mixture model as an approximation, and the insights drawn from our experimental results reflect a similarity to those deduced from our theoretical analysis, confirming the effectiveness of the source and target data in more complicated learning problems.

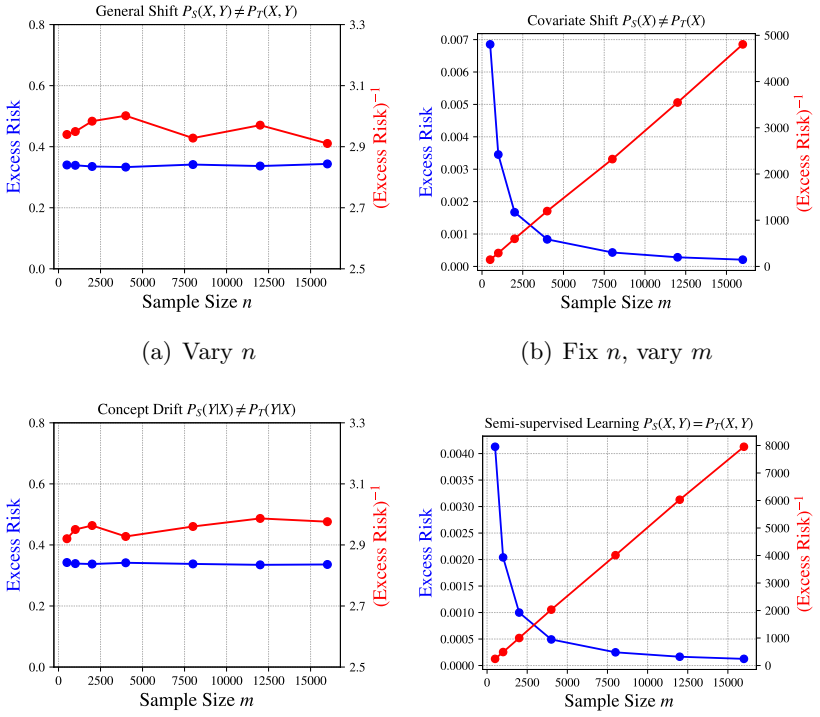
6.1 A toy example

We will numerically confirm our main results using a toy example. We consider a simple example where $\mathcal{Y} = \{0, 1\}$ and $\mathcal{X} = \{1, 2, 3, 4\}$. In causal learning, we model the data distributions as

$$\begin{aligned} X &\sim \mathrm{Cat}(\theta_{x_1}, \theta_{x_2}, \theta_{x_3}, \theta_{x_4}) \\ Y_{x_i} &\sim \mathrm{Ber}(\theta_{Y_{x_i}}) \text{ for } i = 1, 2, 3, 4. \end{aligned}$$

We set $\theta_X^{t*} = (0.25, 0.25, 0.25, 0.25)$ and $\theta_{Y_X}^{t*} = (0.3, 0.4, 0.5, 0.6)$ for synthetic experiments, and we will vary $\theta_X^{s*} = (0.6, 0.1, 0.1, 0.2)$ and $\theta_{Y_X}^{s*} = (0.5, 0.5, 0.3, 0.5)$ for the covariate shift and concept drift conditions, respectively. The parameters are estimated using the maximum likelihood algorithm and used in the prediction. We run experiments 3000 repeatedly and the results are shown in Figure 2. For the general shift case in (a), we fix $m = 2000$ and vary

n from 500 to 16000 and it can be seen that with the unlabelled target sample increasing, the risk will remain around 0.34 and hence does not converge in this case. We sketch the regret for covariate shift and semi-supervised learning in figures (b) and (d), here we fix $n = 2000$ and vary m from 500 to 16000. It can be seen in that $\mathcal{R}(b)$ in blue converges to zero with m increasing in these two cases, then we also plot the $\mathcal{R}(b)^{-1}$ in red to show the rate. The reciprocal of the excess risk is linear in the source sample size, which coincides with our theoretical analysis. It is worth pointing out that the slopes are different in these two cases because the quantity will depend on the Fisher information matrix of $P_{\theta^{s*}_{Y_X}}(Y)$ and the distribution of the covariate X varies across two domains. For concept drift learning in (c), we fix $n = 2000$ and vary m from 500 to 16000. Similar to the general shift case, the excess risk is maintained around 0.34 as well, which is independent of the source sample size m.



(a) Vary n (b) Fix n, vary m

(c) Fix n, vary m (d) Fix n, vary m

Figure 2: Excess risk comparisons under causal learning. (a) and (c) represents the results of $\mathcal{R}(b)$ for general shift case and concept drift learning, where we vary n from 500 to 16000 in (a), and fix $n = 2000$ but vary m from 500 to 16000 in (c). We sketch the results $\mathcal{R}(b)$ for covariate shift and semi-supervised learning in (b) and (d), here we fix $n = 2000$ and vary m from 500 to 16000. We also plot $\mathcal{R}(b)^{-1}$ to show the rate w.r.t. m. We plot all excess risks in blue and their reciprocals in red. All results are derived by 3000 experimental repeats.

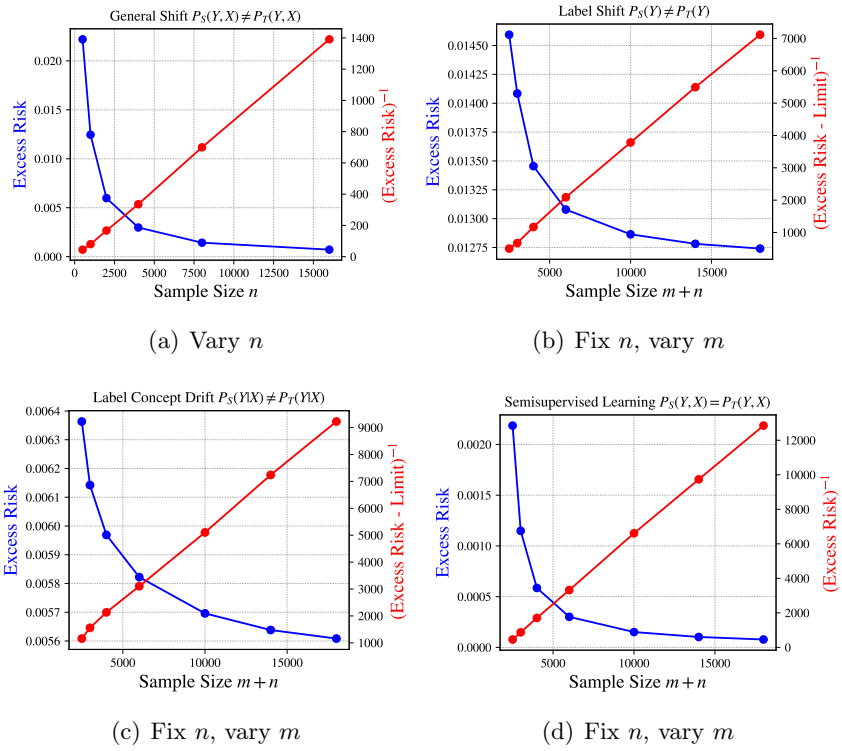
In anti-causal learning, we will model the distributions of the outcome random variables as

$$
\begin{aligned}
Y &\sim \mathrm{Ber}(\theta_Y), \\
X_0 &\sim \mathrm{Cat}(\theta_0, \theta_0 + 0.55, \theta_0 + 0.2, 0.25 - 3\theta_0), \\
X_1 &\sim \mathrm{Cat}(\theta_1, \theta_1 + 0.25, 0.4 - 3\theta_1, \theta_1 + 0.35).
\end{aligned}
$$

For experiments, we set $\theta_Y^{t*} = 0.5$ and $\theta_{X_0}^{t*} = \theta_{X_1}^{t*} = 0.05$ as an example, and we will vary $\theta_Y^{s*} = 0.7$ and $\theta_{X_0}^{s*} = \theta_{X_1}^{t*} = 0.01$ for the target shift and conditional shift conditions, respectively. Using the maximum likelihood algorithm, we sketch the results in Figure 3. For the general shift case in (a), the excess risk converges as n becomes larger, and more explicitly $\mathcal{R}(b)^{-1}$ is linear in n, which confirms our theoretical result. For target shift and conditional shift in (b) and (c), it can be seen that $\mathcal{R}(b)$ converges to a non-zero value λ with m increasing in these two cases, then we also plot the $(\mathcal{R}(b) - \lambda)^{-1}$ to show the rate w.r.t. the sample size $m + n$. These two curves indicate that the source data can only help reduce the excess risk up to a constant. For semi-supervised learning in (d), as expected, the excess risk will converge to zero as m increases. It is also observed that the slope of the reciprocal is higher compared to the general shift condition, implying the source data contain more information than the unlabelled target data and lead to higher scaling factor c (e.g., $O(\frac{c}{m})$) in the rate. We empirically depict the rate of learning performance under different causal mechanisms and domain shift conditions, from which the usefulness of the source and target data is manifested.

6.2 Experiments with Real Datasets

In this section, we shift our focus to real-world datasets (e.g., the MNIST dataset) for anti-causal learning to further reinforce our idea in practical scenarios. Although the core of our analysis lies in the assumption that the data distribution is parametric, this is often not the case when dealing with real-world data. As such, we need to find a parametric model to approximate the true underlying distribution with finite samples. In the following, we use Gaussian mixture models (GMM) to approximate the data, where we assume each class label y_i corresponds to a specific cluster of features and these features are modeled by a Gaussian distribution denoted as $P_{X_{y_i}}(x)$, with parameters including a mean vector μ_i and a covariance matrix Σ_i. Our implementation of this model is based on the expectation-maximization (EM) algorithm (Dempster et al., 1977) by efficiently estimating the initial GMM parameters from the labelled data, and the parameters will be updated with the additional unlabelled data or data with a distributional shift. This framework has been applied to semi-supervised learning and unlabelled domain adaptation problems where the details are outlined in Algorithm 1. While there would exist a potential mismatch between the parametric model and the true underlying distribution and some estimation errors, the empirical results nevertheless demonstrate that anti-causal learning can enhance prediction performance when we efficiently use unlabelled target data and source data.



(a) Vary n

(b) Fix n, vary m

(c) Fix n, vary m

(d) Fix n, vary m

Figure 3: Excess risk comparisons under anti-causal learning. (a) represents the results of $\mathcal{R}(b)$ and $\mathcal{R}(b)^{-1}$ for general shift case, and we vary n from 500 to 16000. We sketch the results $\mathcal{R}(b)$ for label shift, label concept drift and semi-supervised learning in (b), (c) and (d). Here we fix $n = 2000$ and vary m from 500 to 16000. It can be seen in that $\mathcal{R}(b)$ converges to a non-zero value λ with m increasing in (b) and (c), then we also plot $(\mathcal{R}(b) - \lambda)^{-1}$ to show the rate w.r.t. $m+n$. We plot all the excess risks in blue and their reciprocals in red. All results are derived by 3000 experimental repeats.

Semi-supervised Learning

The MNIST dataset[2] (LeCun et al. (1998)) serves as a well-recognized standard for benchmarking, comprising 70,000 grayscale, handwritten digit images (ranging from 0 to 9), each of pixel size 28×28. It is a frequent choice for testing various machine learning algorithms, particularly in image classification scenarios. Our analysis will primarily focus on exploring the usefulness of unlabelled data under the anti-causal learning setting using the Gaussian mixture model, specifically with the MNIST dataset. To achieve this, we select two digits at random (for instance, 2 and 5) and construct a dataset comprising 100 labelled samples, while varying the unlabelled sample size from 0 to 1,000. By introducing a small set of labelled target data, we can accurately determine the correct labels, addressing the potential label-swapping issue that may arise with the unlabelled data only. Our goal is to demon-

2. http://yann.lecun.com/exdb/mnist/

Algorithm 1: Anti-Causal Learning with GMMs

Data: A small set of labelled target training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ with N samples, where x_i are features and y_i are labels, unlabelled target training dataset $\mathcal{D}_U = \{x_i\}_{i=1}^{M}$ with M samples, labelled source training dataset $\mathcal{D}' = \{(x_i, y_i)\}_{i=1}^{L}$ with L samples and test dataset $\mathcal{D}_T = \{(x_i, y_i)\}_{i=1}^{T}$ with T samples

Result: Improved prediction performance using GMM on $\mathcal{D}_T$.

Initialize K, the number of Gaussian components, corresponding to the number of class labels.

Initialize parameters $\Theta = \{\mu_k, \Sigma_k\}_{k=1}^{K}$ for each Gaussian component.

Step 1: Feature Engineering

Conduct feature engineering using methods such as PCA or other feature selection with $\mathcal{D}$, and $\mathcal{D}_U$ or D' depending on the SSL/UDA tasks

Step 2: Parameter Estimation

for *each class label* $k = 1$ *to* K **do**

    Estimate μ_k and Σ_k using EM algorithm on $\mathcal{D}$ with corresponding instances with label k.

Step 3: SSL/UDA with GMM

while *not converged* **do**

    For SSL: Use unlabelled data $\mathcal{D}_U$ to update Θ by the EM algorithm

    For DA: Use labelled source data $\mathcal{D}'$ to update Θ by the EM algorithm

Step 4: Prediction

for *each new instance* x *in* $\mathcal{D}_T$ **do**

    Predict label y by selecting the Gaussian component k that maximizes $P_{X_k}(x)$ with parameters (μ_k, Σ_k).

strate that incorporating unlabelled data can still improve the performance of the model effectively. The initial step in our approach involves data preprocessing, which includes applying principal component analysis (PCA) to both the labelled and unlabelled datasets to reduce the input feature dimensionality from 784 down to a manageable number - 20 in our experiment. This reduction aids in addressing the curse of dimensionality, enhancing the computational speed and potentially boosting the Gaussian mixture model's performance. Following this, we establish an initial Gaussian mixture model using the labelled data only. Then we follow the procedures in Algorithm 1 to update the parameters of the initial GMM. We will finally compare the performance of the updated GMM with its initial model using a test set from the same digit pair with the size of 3,000.

Figure 4 illustrates the test set accuracy for different sizes of unlabelled data for the digit pair (2, 5). Our observations indicate that integrating unlabelled data significantly improves the model performance. Correspondingly, as the size of unlabelled data increases, the model accuracy also sees an increase, achieving approximately 99% accuracy when the data size exceeds 500. This improvement indicates that unlabelled data indeed helps estimate the distribution parameters in the context of anti-causal learning, and this also empirically validates the results we presented in Table 1. To visualize the model performance on these

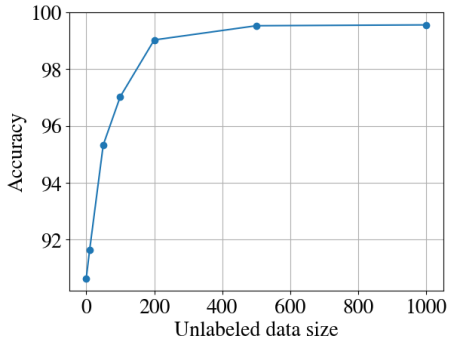
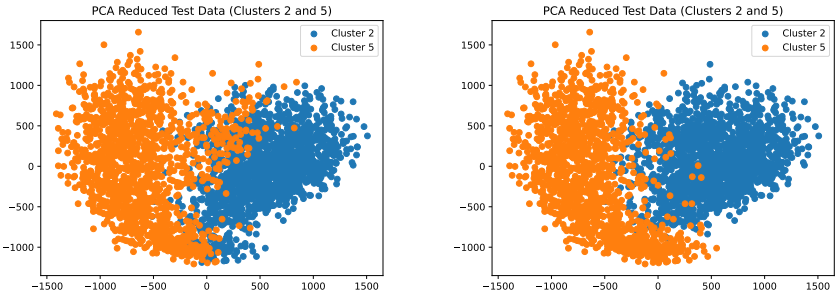


Figure 4: Accuracy v.s. unlabelled sample size for digit pair (2, 5)



(a) Initial GMM

(b) Update GMM

Figure 5: Visualization of clusters for various source and target combinations for digit pair (2, 5) for initial and updated GMM with 100 labelled data and 500 unlabelled data

two clusters, we further illustrate the clusters by plotting the two most significant principle components in Figure 5. It demonstrates that updated GMM learning can indeed make two clusters more distinct and separable than the initial GMM, which leads to higher accuracy. To provide a more comprehensive demonstration of the usefulness of unlabelled data, we have randomly selected several additional digit pairs and conducted experiments with varying amounts of unlabelled data. We summarize the result in Table 2. From the table, we can see that, in all cases, the unlabelled data help improve the accuracy in predictions, and as the sample size of unlabelled data increases, the accuracy also improves correspondingly. However, due to the variability between different digit pairs, and randomness from train and test sampling and estimation errors, the extent to which unlabelled data improves accuracy varies across different experiments. Through the experimental validation conducted on the MNIST dataset, our results confirm the substantial impact of unlabelled data on enhancing the performance of the anti-causal learning setting, particularly under conditions where labelled samples are limited. This establishes the crucial role of anti-causal learning settings in practical applications when it comes to semi-supervised learning problems.

Unlabelled Size	(2,5)	(5, 9)	(3, 8)	(4, 7)	(0, 6)	(2, 3)
0	0.896	0.531	0.575	0.854	0.893	0.855
50	0.941	0.623	0.817	0.858	0.942	0.865
200	0.990	0.636	0.852	0.905	0.983	0.884
500	0.991	0.774	0.891	0.917	0.985	0.937

Table 2: Performance comparison of different sizes of datasets on various digit pairs

Unlabelled Domain adaptation

We further assess the effectiveness of anti-causal learning in the realm of unlabelled domain adaptation. Here, we include three different source data domains for comparisons: the United States Postal Service (USPS) dataset (Hull, 1994), an adapted MNIST dataset with added Gaussian noise, and a colour-infused MNIST dataset with colored backgrounds added to the digits. The USPS dataset, frequently used for digit recognition and domain adaptation tasks, consists of 9,298 grayscale images of handwritten digits (0-9) with a pixel resolution of 16×16. For the target domain, we randomly select two digits from the MNIST dataset to create a dataset containing 100 labelled samples. Subsequently, we will introduce the aforementioned three source data, each with 500 labelled samples, to help update the distribution parameters learned from the initial GMM. We aim to examine whether introducing an additional labelled dataset can significantly improve model performance, particularly when the causal mechanisms and generating distributions are closely similar. We apply a similar algorithm used in semi-supervised learning where we first apply PCA to both source and target data, and then we construct an initial GMM with the target data and then update the GMM using the EM algorithm on the source data. Here we pick various digit pairs to evaluate the effectiveness of the source data, and the results are summarized in Table 3.

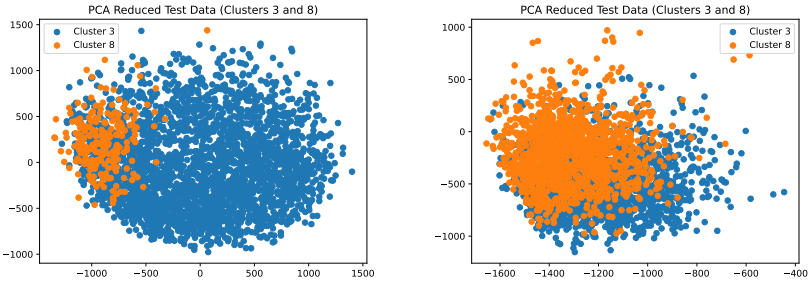
Source	(2,5)	(5,9)	(3, 8)	(4,7)	(0,6)	(2,3)
-	0.896	0.531	0.575	0.854	0.900	0.850
Colored MNIST	0.989	0.636	0.860	0.857	0.985	0.933
Noisy MNIST	0.993	0.926	0.882	0.889	0.979	0.946
USPS	0.971	0.835	0.840	0.525	0.550	0.510

Table 3: Performance comparison of different source datasets on various digit pairs. Here the sign '-' represents the accuracy with only 100 labelled MNIST data without any source data, while the remaining three rows are the performance with 500 additional colored MNIST, noisy MNIST and USPS data, respectively.

As can be observed from the above table, we compared the model performance by accuracy between not using source data and using three different types of source data. In most cases, the introduction of source data showed an improvement over not using source data, validating the beneficial impact of source data on target performance enhancement. Moreover, when comparing different source data, the colored MNIST and noisy MNIST are

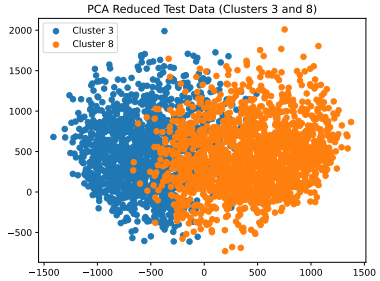
closer to the original MNIST in terms of the conditional generating distribution $P(X|Y)$, and they do perform better than the USPS in almost all cases.

We also plot the two main components for clusters 3 and 8 in Figure 6 to visualize the constructed GMM model. We can infer from the figure that the GMM model trained without using source data yields the poorest performance, as it fails to distinguish between digits 3 and 8 accurately, and moreover, the prediction of digit 8 is noticeably biased, contradicting the testing label distributions. Upon the introduction of source data, the GMM model trained with the additional USPS dataset still exhibits a substantial overlap between 3 and 8 in the test set, implying a less optimal performance. On the other hand, with the colored MNIST dataset, the two clusters are more separated, representing the best prediction performance.

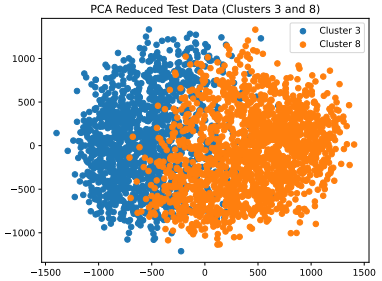


(a) MNIST

(b) MNIST + USPS



(c) MNIST + Colored MNIST



(d) MNIST + Noisy MNIST

Figure 6: Visualization of clusters for various source and target combinations for digit pair (3,8)

Referring to the table, we also noticed some cases where the use of USPS actually undermined the model accuracy for the digit pair $(4, 7)$, $(0, 6)$ and $(2, 3)$. We point out that this does not contradict our earlier analytical results (source data should never degrade the performance). The reason is that the GMM models used to train the classifier are only approximations of the "true" model, and importantly, the testing data is *not* from these approximating parametric models but from the real dataset, whereas our analytical results hold under the assumption that both training and testing data are from parametric models.

Nevertheless, we see that satisfactory results can still be achieved in many scenarios with this empirical setup, even when approximations are used, showing effective guidance of our theoretical results.

Multi-classification with SSL and UDA

In the previous section, we provided a simplified comparison of SSL and UDA by focusing on results involving just two numerical categories. These experiments helped clearly demonstrate the data's practical value through 2D visual representations. In this section, we aim to assess the comprehensive performance across the dataset by applying our algorithm to data that includes all labels, e.g., the multi-classification of handwritten digits ranging from 0 to 9. For experiments, we randomly select 200 samples from the MNIST dataset for our initial labelled target dataset. Then, to explore the impact of additional training data, we gradually increase the number of these extra training samples from 400 to 5000. These additional samples are sourced from various datasets, including unlabelled MNIST samples or labelled samples from variants of the MNIST dataset (such as coloured MNIST and noisy MNIST) and the USPS dataset. Furthermore, we investigate how the number of PCA dimensions and the number of clusters in our model affect its performance. The results are organized across three tables. Table 4 details how varying the size of additional data samples impacts the model performance. Table 5 explores the influence of changing the dimensions within PCA. Lastly, Table 6 examines the effects of altering the number of clusters in GMM. From the results, we identify some key insights as follows.

Sample sizes	400	800	1600	3200	5000
-	0.397				
Unlabelled MNIST	0.364	0.545	0.606	0.623	0.636
Colored MNIST	0.399	0.400	0.481	0.455	0.531
Noisy MNIST	0.483	0.420	0.567	0.468	0.562
USPS	0.354	0.367	0.271	0.259	0.335

Table 4: Effect of the sample size for additional training instances, where we set $N = 200$, $K = 10$ and PCA dimension to be 15. Here the sign '-' represents the accuracy with only 200 labelled MNIST data without any source data, while the rest four rows are the results with additional unlabelled MNIST, colored MNIST, noisy MNIST and USPS data, respectively (the same applies to tables below).

- **Additional Training Samples:** Including extra unlabelled MNIST samples steadily improves the model's performance, showing the value of unlabelled data in SSL. Nonetheless, the effect of augmenting the dataset with colored or noisy MNIST samples varies, indicating that while adding more training data from similar distributions can be advantageous, the presence of distribution shifts or noise might occasionally degrade the performance. The decrease in performance with USPS samples highlights the difficulty in adapting the model to different data distributions, also previously observed in Table 3 where the testing data distribution deviates from these approximating

PCA dimension	5	15	25	35	45
-	0.470	0.397	0.195	0.485	0.372
Unlabelled MNIST	0.531	0.606	0.287	0.506	0.445
Colored MNIST	0.512	0.481	0.113	0.456	0.353
Noisy MNIST	0.538	0.567	0.137	0.441	0.461
USPS	0.264	0.271	0.139	0.132	0.094

Table 5: Effect of the cluster number where we set $N = 200$, $M = L = 1600$ and the cluster number to be 10

Cluster number	10	15	20	25	30
-	0.590	0.616	0.609	0.563	0.491
Unlabelled MNIST	0.680	0.669	0.696	0.614	0.468
Colored MNIST	0.570	0.578	0.458	0.423	0.335
Noisy MNIST	0.693	0.695	0.727	0.649	0.592
USPS	0.458	0.487	0.479	0.338	0.223

Table 6: Effect of the cluster number where we set $N = 400$, $M = L = 1600$ and PCA dimension to be 15

parametric models, emphasizing that in practice, the data might be instead useless if the generating distribution varies too much in the anti-causal direction.

- **PCA Dimensions:** The link between the number of dimensions in PCA and how well a model performs is complex, showing that there is not a clear connection between adding more dimensions and achieving better performance. The best number of PCA dimensions changes depending on the dataset, suggesting the importance of a customized strategy for reducing dimensions that focuses on preserving key features while eliminating the effect of other factors, such as noise. This concept is especially clear when looking at the decline in performance across all dimension levels with USPS data, demonstrating the difficulties in applying a one-size-fits-all approach to different datasets.

- **Cluster Number:** The effectiveness of the model changes as the number of clusters changes. There is performance improvement up to a certain cluster number for particular datasets, and then it starts to decrease as the cluster increases. This indicates that there is an ideal number of clusters that can enhance the model's performance, a trend that is particularly noticeable with unlabelled and noisy MNIST datasets. On the other hand, for colored MNIST and USPS datasets, the performance tends to worsen as the number of clusters increases. This could be caused by over-segmentation or the loss of important features due to too many clusters.

These experiments examine the impact of different factors, such as additional data sample size, the number of PCA dimensions, and the number of clusters on the performance of models across various datasets for anti-causal learning. In the anti-causal learning setup,

more unlabelled data without the distribution shift generally boosts the model performance, but adding labelled source data (such as the refactored MNIST datasets and USPS in our example) does not always lead to better results, pointing to the importance of causal direction and data generating mechanisms. The optimal number of PCA dimensions and clusters is not one-size-fits-all but needs customization for each dataset to ensure key parameters are retained while minimizing noises from the redundant features. For some datasets like unlabelled and noisy MNIST, a specific cluster number can improve performance, whereas for others, like colored MNIST and USPS, it may cause problems due that the testing data may not be drawn from these approximating GMM distributions and possibly over-segmentation with large cluster numbers or the loss of important features with small PCA dimensions.

7. Extensions to Unknown Causal Settings

Even though in this work we primarily focus on the setup where the setting is known to be either causal learning or anti-causal learning, it is also interesting to consider the scenario where the underlying relationship between X and Y is acyclic but **unknown**. We ask the question, which causal direction should we use for prediction? Our strategy is that given the statistics from the observed data (X, Y), we try to fit the data with both causal-learning and anti-causal learning settings and decide which setting will enable us to make predictions more efficiently. Notice that it could be the case that the chosen setting is not the true underlying mechanism (and perhaps not physically possible). However, this is irrelevant as far as the prediction is concerned, as we only work with observed data and will not intervene in the system. By the same argument, we could choose either setting for the prediction *even if the true causal setting is known.* So it is tempting to carry out this comparison even if we know the true direction. However, it does not seem to be fruitful in general. Indeed, as pointed out by Kocaoglu et al. (2017) and Compton et al. (2020), if we want to use an anti-causal learning setting to fit the data generated from a causal learning setting (or vice versa), this "artificial" fitting is in general much more complicated than fitting from the true underlying setting, which would make the prediction more difficult.

If the causal relationship between X and Y for a certain learning problem is unknown and we can model the data from both directions, our results imply that we should use whichever model achieves a better learning performance. This can be viewed as a causal model selection problem. Referring to Table 1, for semi-supervised learning, the rate from the causal direction will be $O(\frac{k}{m})$ while $O(\frac{k'+1}{m+n})$ for anti-causal learning if we have abundant source data ($n \ll m$) and $k < k' + 1$, fitting from the causal direction will be easier. In contrast, if we have abundant target data ($m \ll n$), then fitting from the anti-causal direction will be more favourable. Using similar arguments in the domain adaptation scenarios, if the covariate shift assumption does not hold, the source data will be unhelpful from the causal direction, and we should always fit from the anti-causal direction. Otherwise, the model selection is, again, determined by the sample sizes m and n.

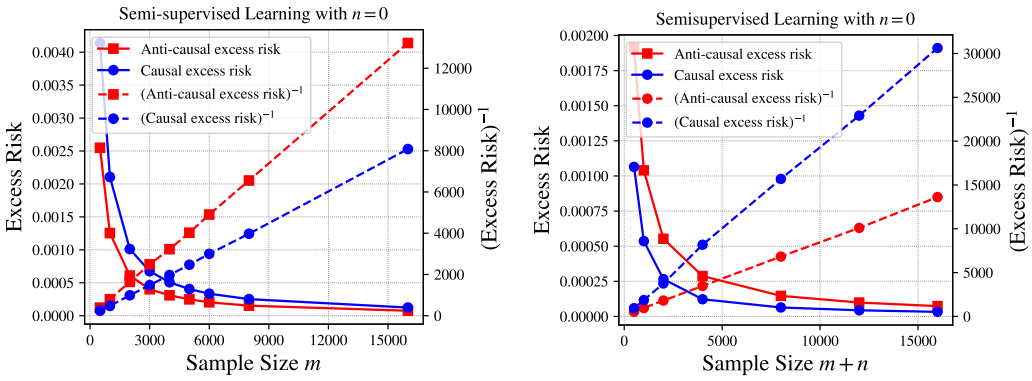
In an attempt to investigate the model selection issue, we examine the excess risk from numerical analysis for the aforementioned parametric models under the semi-supervised learning condition for the sake of simplicity. We will consider the distribution $P_S(X, Y) =$

$P_T(X, Y)$ from the anti-causal direction as:

$$\begin{aligned} Y &\sim \text{Ber}(0.5), \\ X_0 &\sim \text{Cat}(0.05, 0.6, 0.25, 0.1), \\ X_1 &\sim \text{Cat}(0.05, 0.3, 0.25, 0.4). \end{aligned}$$

by setting $\theta_Y = 0.5$, $\theta_0 = 0.05$ and $\theta_1 = 0.05$. We can also model the same joint distribution from the causal directions by choosing the parameters as follows:

$$\begin{aligned} X &\sim \text{Cat}(0.05, 0.45, 0.25, 0.25), \\ Y_{x_1} &\sim \text{Ber}(0.5), \quad Y_{x_2} \sim \text{Ber}(\frac{1}{3}), \\ Y_{x_3} &\sim \text{Ber}(0.5), \quad Y_{x_3} \sim \text{Ber}(0.8). \end{aligned}$$



(a) $k = 4, k' = 2$ (b) $k = 1, k' = 2$

Figure 7: Excess risk comparisons fitting from causal (red) and anti-causal (blue) under semi-supervised learning with labelled data only. Results are derived by the same parameterization from anti-causal learning with $k' = 2$ but different parameterization from causal learning with $k = 4$ in (a) and $k = 1$ in (b).

By varying the sample size m from 500 to 16000, we plot the excess risk $\mathcal{R}(b)$ under causal and anti-causal learning settings in Figure 7(a). It is observed that both directions produce the same rate of $O(\frac{1}{m})$. Compared to the causal case, fitting from the anti-causal direction enjoys a lower regret, and the slope of its reciprocal is higher, which implies that it is "easier" to learn the distribution $P(Y|X)$ from the anti-causal direction. Roughly speaking, the reason is that learning $\theta^*_{Y_X}$ requires $k = 4$ parameters, but the inference from the anti-causal direction only requires $k' + 1 = 3$ parameters, which decreases the model uncertainty and hence the better performance. Rigorously speaking, the slope (or scaling factor in the rate) depends on the information dimension (see Haussler and Opper (1995) for reference). For example, under causal learning, the convergence rate is proved to be $\frac{k}{2m}$ and the slope will be $\frac{2}{k}$ where $k = 4$ is the number of parameters for $\theta^*_{Y_X}$ in this case. It is also confirmed from the figure that the slope is roughly $\frac{1}{2}$. The same argument applies

in the anti-causal learning, and the convergence rate is $\frac{k'+1}{2m}$ when m is sufficiently large, leading to a lower regret since $k > k' + 1$. With such parameterization, it is always better to fit from the anti-causal direction.

However, if we model the distribution from the causal directions by setting $\theta_Y = 0.5$ in the following way:

$$\begin{aligned} X &\sim \text{Cat}(0.05, 0.45, 0.25, 0.25), \\ Y_{x_1} &\sim \text{Ber}(\theta_Y), \quad Y_{x_2} \sim \text{Ber}(\theta_Y - \frac{1}{6}), \\ Y_{x_3} &\sim \text{Ber}(\theta_Y), \quad Y_{x_4} \sim \text{Ber}(\theta_Y + 0.3), \end{aligned} \tag{40}$$

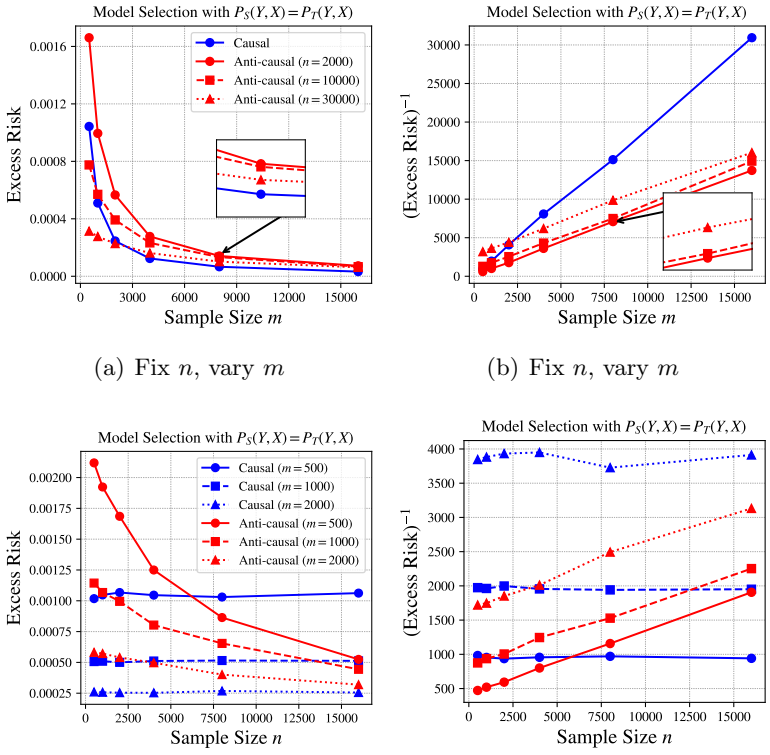
With such a restriction, the number of parameters k is reduced to 1. We successively repeat the experiment and plot the result in Figure 7(b). The excess risk, in this case, becomes lower than fitting from the anti-causal direction and the rate is improved to approximately $\frac{1}{2m}$. The results indicate that the model selection depends on how we parameterize the data distributions, particularly the number of parameters from each causal direction.

In the above example, we only consider the labelled data. The unlabelled samples, however, are not useful for the causal direction but will take effect from the anti-causal direction from Figure 2(a) and 3(a). Both causal learning and anti-causal learning can be more favorable than the other option, depending on the sample sizes. For instance, if we have abundant unlabelled data and limited labelled data, referring to Table 1, fitting from anti-causal direction yields the rate $O(\frac{k'+1}{m+n})$, which is better than the rate $O(\frac{k}{m})$ under causal direction if $n \gg m$.

To numerically illustrate, we conduct the experiments with the parameterization in (34) from anti-causal direction and (40) from causal direction under semi-supervised learning with both labelled and unlabelled data. We then plot the results in Figure 8.

We firstly vary m from 500 to 16000 by fixing $n = 2000, 10000$ and 30000 to show the effectiveness of labelled data. We plot the corresponding results of $\mathcal{R}(b)$ in subfigure 8(a) and $\mathcal{R}(b)^{-1}$ in 8(b). In 8(a), we only plot one curve in blue since n does not affect the excess risk from the causal direction. The remaining three red curves are derived by fitting from the anti-causal direction with an increasing n, from top to bottom. One can observe that a larger n will incur a smaller initial excess risk when $m = 500$. However, the convergence rates are identical for all three cases. Since the slope of $\mathcal{R}(b)^{-1}$ is higher from the causal direction, when m is large enough ($m > 2000$), even with large unlabelled data ($n = 30000$), the excess risk is still higher fitting from the anti-causal direction.

The subfigure 8(c) shows the results of $\mathcal{R}(b)$ and 8(d) of $\mathcal{R}(b)^{-1}$ by varying n from 500 to 16000 and fixing $m = 500, 1000$ and 2000. In 8(c), from top to bottom, three blue curves correspond to $m = 500, 1000$ and 2000 by the causal direction and the three red curves by the anti-causal direction. In this case, the excess risk from the causal direction is almost a constant depending on m, regardless of the unlabelled sample size n. Furthermore, a higher m incurs a lower regret. On the contrary, from the anti-causal learning direction, the excess risk will converge as n goes sufficiently large. Selecting an appropriate model strongly hinges on the unlabelled target sample size n. For example, in our formulation, when $m = 500$, we may need more than 6500 extra unlabelled samples to achieve a lower regret, and if m doubles, we will need to double the required unlabelled samples to achieve a comparable expected risk.



(a) Fix n, vary m

(b) Fix n, vary m

(c) Fix m, vary n

(d) Fix m, vary n

Figure 8: The figure shows the excess risk comparisons by varying m and n under causal and anti-causal learning for semi-supervised learning. The subfigure 8(a) shows the results of $\mathcal{R}(b)$ and 8(b) of $\mathcal{R}(b)^{-1}$ (sharing the same legend) by varying m from 500 to 16000 and fixing $n = 2000, 10000$ and 30000, respectively. In 8(a), the blue curve shows the result by fitting from the causal direction and from top to bottom, the other three red curves are derived by fitting from the anti-causal direction with $n = 500, 1000$ and 2000. The subfigure 8(c) shows the results of $\mathcal{R}(b)$ and 8(d) of $\mathcal{R}(b)^{-1}$ (sharing the same legend) by varying n from 500 to 16000 and fixing $m = 1000, 2000$ and 3000, respectively. In 8(c), from top to bottom, three blue curves correspond to $m = 500, 1000$ and 2000 by the causal direction and the three red curves by the anti-causal direction. All results are derived by 3000 experimental repeats.

Overall, for a general domain adaptation task without knowing the underlying causal mechanism, if we can model the data with parameterised distributions for both causal and anti-causal directions without some physical constraints, both models can be more favourable than the other option depending on how we do the parameterization, how many data samples we have and how different the source and target domains are.

8. Conclusions

This paper proposes a probabilistic framework articulating the connection between SSL/UDA and causal mechanisms. We explicitly characterize the rate of learning performance under different causal mechanisms and domain shift conditions, from which the usefulness of the source and target data is manifested. However, in our analysis, the parametric characterization of both source and target data is crucial. A possible future direction is to relax the assumptions on parametric conditions to general probability distributions and find the excess risk in terms of the sample sizes. Our analysis also heavily relies on the generating processes we skectch in Figure 1 (e.g., X and Y are unconfounded), and the possible future work could be performing a similar analysis for the case with more than two variables (e.g., causal setting with con-founders), which improves the generality and applicability in real-world problems. We have also observed that incorporating unlabelled data and labelled source data could significantly enhance the model performance for the target domain on both synthetic data and real benchmarks. Due to the discrepancy between the approximated parametric distribution and the underlying data distribution for real-world scenarios, our theoretical analysis cannot directly carry over. In addition, developing a method that can effectively handle non-parametric distributions is also a potential direction worth exploring.

Appendix A. Appendix: Proofs

A.1 Mixture Asymptotics Lemma

Lemma 14 (Mixture Asymptotics) *Under Assumption 1,2,3 and assume $m \asymp n^p$ for some $p > 0$ and let $n \to \infty$, then the mixture strategy yields*

$$D(P_{\theta^*}(D_s^m, D_t^{U,n}) \| Q(D_s^m, D_t^{U,n})) = \frac{d}{2} \log \frac{1}{2\pi e} + \log \frac{1}{\omega(\theta^*)} + \frac{1}{2} \log \det(I_{st}) + o(\frac{1}{n \vee m}), \tag{41}$$

where $\theta^* \in \mathbb{R}^d$ denotes the total parameters that characterize the source and target distributions and d denotes the total dimension, depending on the causal directions and distribution shifting conditions. The Fisher information matrix associated with D_s^m and $D_t^{U,n}$ is defined as $I_{st} = -\mathbb{E}_{\theta^*}[\nabla^2 \log P(D_s^m, D_t^{U,n} | \theta^*)]$.

Proof The proof and result is a generalization of Clarke and Barron (1990); Zhu (2020) with some modifications to fit our purpose. Without the loss of generality, we first assume that the source parameter and target parameter will have $\tilde{k} + 1 - c$ domain-specific parameters and c domain-sharing parameters, where $\tilde{k} = k$ for causal learning and $\tilde{k} = k'$ for anti-causal learning. c will vary under different shift conditions. For example, under the target shift condition in anti-causal learning, c will be k' for identical parameters $\theta_{X_{y_i}}$ in both domains; In conditional shift condition, $c = 1$ since $\theta_Y^{s*} = \theta_Y^{t*}$. With a little abuse of notation in this section, we denote the true source-specific parameters by $\theta_s^* \in \mathbb{R}^{\tilde{k}+1-c}$, the target-specific parameters by $\theta_t^* \in \mathbb{R}^{\tilde{k}+1-c}$ and the domain-sharing parameters as $\theta_c^* \in \mathbb{R}^c$. Then the source data (X, Y) is drawn from the distribution $P_{\theta_c^*, \theta_s^*}$ and the target data X is drawn from the distribution $P_{\theta_c^*, \theta_t^*}$ under such parameterization. For simplicity, we can write the joint domain parameters $\theta^* = (\theta_c^*, \theta_s^*, \theta_t^*)$ and the joint distribution for the source domain data and target domain data is expressed by

$$P_{\theta^*}(D_s^m, D_t^{U,n}) = P_{\theta_c^*, \theta_s^*}(D_s^m) P_{\theta_c^*, \theta_t^*}(D_t^{U,n}) = \prod_{i=1}^{m} P_{\theta_c^*, \theta_s^*}(D_s^m) \prod_{j=1}^{n} P_{\theta_c^*, \theta_t^*}(D_t^{U,n}). \tag{42}$$

Based on the notations above, we define the score functions by

$$l_s(\theta_s, \theta_c) = \nabla \log P(D_s^m | \theta_s, \theta_c), \tag{43}$$

$$l_t(\theta_t, \theta_c) = \nabla \log P(D_t^{U,n} | \theta_t, \theta_c), \tag{44}$$

$$l_{st}(\theta) = \nabla \log P(D_t^{U,n}, D_s^m | \theta). \tag{45}$$

Note that

$$l_{st}(\theta^*) = \begin{bmatrix} l_s(\theta_s^*, \theta_c^*) \\ \mathbf{0}_{\tilde{k}+1-c} \end{bmatrix} + \begin{bmatrix} \mathbf{0}_{\tilde{k}+1-c} \\ l_t(\theta_t^*, \theta_c^*) \end{bmatrix}, \tag{46}$$

where $\mathbf{0}_{\tilde{k}+1-c}$ denotes the zero vector with length $\tilde{k}+1-c$. We next restate the corresponding Fisher information matrix,

$$I_s = -\mathbb{E}_{\theta_s^*,\theta_c^*}[\nabla^2 \log P(X_s, Y_s|\theta_s^*, \theta_c^*)] \in \mathbb{R}^{(\tilde{k}+1)\times(\tilde{k}+1)}, \tag{47}$$

$$I_t = -\mathbb{E}_{\theta_t^*,\theta_c^*}[\nabla^2 \log P(X_t|\theta_t, \theta_c)] \in \mathbb{R}^{(\tilde{k}+1)\times(\tilde{k}+1)} \tag{48}$$

$$I_0 = -\mathbb{E}_{\theta^*}[\nabla^2 \log P(X_s, Y_s, X_t|\theta^*)] \in \mathbb{R}^{(2\tilde{k}+2-c)\times(2\tilde{k}+2-c)}, \tag{49}$$

$$I_{st} = -\mathbb{E}_{\theta^*}[\nabla^2 \log P(D_t^{U,n}, D_s^m|\theta^*)] \in \mathbb{R}^{(2\tilde{k}+2-c)\times(2\tilde{k}+2-c)}. \tag{50}$$

Their corresponding *empirical* versions are denoted by,

$$\tilde{I}_s(\theta_s, \theta_c) = -[\nabla^2 \log P(X_s, Y_s|\theta_s, \theta_c)] \in \mathbb{R}^{(\tilde{k}+1)\times(\tilde{k}+1)}, \tag{51}$$

$$\tilde{I}_t(\theta_t, \theta_c) = -[\nabla^2 \log P(X_t|\theta_t, \theta_c)] \in \mathbb{R}^{(\tilde{k}+1)\times(\tilde{k}+1)}, \tag{52}$$

$$\tilde{I}_0(\theta) = -[\nabla^2 \log P(X_s, Y_s, X_t|\theta^*)] \in \mathbb{R}^{(2\tilde{k}+2-c)\times(2\tilde{k}+2-c)}, \tag{53}$$

$$\tilde{I}_{st}(\theta) = -[\nabla^2 \log P(D_t^{U,n}, D_s^m|\theta)] \in \mathbb{R}^{(2\tilde{k}+2-c)\times(2\tilde{k}+2-c)}. \tag{54}$$

For convenience, if not otherwise stated we will simply omit brackets for θ^* in the sequel, e.g., we write $\tilde{I}_{st}(\theta^*)$ as $\tilde{I}_{st}$. Define the neighbourhood of θ^* by $N_\delta = \{\theta : \|\theta - \theta^*\| \leq \delta\}$ where the norm in $\mathbb{R}^{2\tilde{k}+2-c}$ is defined as

$$\|\xi\|^2 = \xi^T I_0 \xi. \tag{55}$$

Define

$$L(\theta^*) = l_{st}^T(\theta^*) I_{st}^{-1} l_{st}(\theta^*). \tag{56}$$

Note that,

$$\begin{aligned}
\mathbb{E}[L(\theta^*)] &= \mathbb{E}[\mathrm{Tr}(I_{st}^{-1} l_{st}(\theta^*)^T l_{st}(\theta^*))] \\
&= \mathrm{Tr}(I_{st}^{-1} \mathbb{E}[l_{st}(\theta^*)^T l_{st}(\theta^*)]) \\
&= \mathrm{Tr}(I_{st}^{-1} I_{st}) \\
&= 2\tilde{k} + 2 - c.
\end{aligned} \tag{57}$$

For $0 < \epsilon < 1$ and $\delta > 0$, we define three events $A(\delta, \epsilon)$, $B(\delta, \epsilon)$ and $C(\delta)$ as

$$A(\delta, \epsilon) = \left\{ \int_{N_\delta^c} P\left(D_m^s, D_t^{U,n} \mid \theta\right) \omega(\theta) d\theta \leq \epsilon \int_{N_\delta} P\left(D_m^s, D_t^{U,n} \mid \theta\right) \omega(\theta) d\theta \right\}, \tag{58}$$

$$B(\delta, \epsilon) := \{(1-\epsilon)(\theta - \theta^*)^T I_{st} (\theta - \theta^*) \leq (\theta - \theta^*)^T \left(\tilde{I}_{st}(\theta')\right) (\theta - \theta^*), \tag{59}$$

$$\leq (1+\epsilon)(\theta - \theta^*)^T I_{st} (\theta - \theta^*) \quad \text{for all } \theta, \theta' \in N_\delta\} \tag{60}$$

$$C(\delta) := \left\{L(\theta^*) \leq \min\{n, m\}\delta^2\right\}, \tag{61}$$

and

$$\rho(\delta, \theta^*) = \sup_{\theta \in N_\delta} \left| \frac{\omega(\theta)}{\omega(\theta^*)} \right|. \tag{62}$$

Following the similar procedures in Clarke and Barron (1990), we have the following upper and lower bounds on the density ratio.

Lemma 15 *We assume condition 3 in Assumption 2 holds that P_θ^* is twice differentiable around θ^* and I_{st} is positive definite. With proper prior $\omega(\theta)$, then on the set of $A \cap B$, we have,*

$$\frac{Q(D_m^s, D_n^{t,U})}{P_{\theta^*}(D_m^s, D_n^{t,U})} \leq (1+\epsilon)\omega(\theta^*)e^{\rho(\delta,\theta^*)}(2\pi)^{\frac{2\tilde{k}+2-c}{2}} e^{\frac{1}{2(1-\epsilon)}L(\theta^*)} \det((1-\epsilon)I_{st})^{-\frac{1}{2}}. \tag{63}$$

Further, on the set of $B \cap C$, we have the lower bound,

$$\frac{Q(D_m^s, D_n^{t,U})}{P_{\theta^*}(D_m^s, D_n^{t,U})} \geq \omega(\theta^*)e^{-\rho(\delta,\theta^*)}(2\pi)^{\frac{2\tilde{k}+2-c}{2}} e^{\frac{1}{2(1+\epsilon)}L(\theta^*)}(1-2^{\frac{2\tilde{k}+2-c}{2}}e^{-\epsilon^2(n\wedge m)\delta^2/8}) \det((1+\epsilon)I_{st})^{-\frac{1}{2}}. \tag{64}$$

Proof In both cases, we will use the Laplace method to give an upper and lower bound on the density ratio, for the upper bound, if we restrict on A and B, then,

$$\begin{aligned}
\frac{Q(D_m^s, D_n^{t,U})}{P_{\theta^*}(D_m^s, D_n^{t,U})} &\leq (1+\epsilon)\int_{N_\delta} \frac{P_\theta(D_m^s, D_n^{t,U})}{P_{\theta^*}(D_m^s, D_n^{t,U})}\omega(\theta)d\theta \\
&= (1+\epsilon)\int_{N_\delta} e^{\log\frac{P_\theta(D_m^s, D_n^{t,U})}{P_{\theta^*}(D_m^s, D_n^{t,U})}}\omega(\theta)d\theta \\
\text{(Taylor Expansion)} \quad &= (1+\epsilon)\int_{N_\delta} e^{(\theta-\theta^*)^T l_{st}(\theta^*) - \frac{1}{2}(\theta-\theta^*)^T \tilde{I}_{st}(\theta')(\theta-\theta^*)}\omega(\theta)d\theta \\
\text{(Definition of } \rho) \quad &\leq (1+\epsilon)\omega(\theta^*)e^{\rho(\delta,\theta^*)}\int_{N_\delta} e^{(\theta-\theta^*)^T l_{st}(\theta^*) - \frac{1}{2}(\theta-\theta^*)^T \tilde{I}_{st}(\theta')(\theta-\theta^*)}d\theta \\
\text{(Event } B) \quad &\leq (1+\epsilon)\omega(\theta^*)e^{\rho(\delta,\theta^*)}\int_{N_\delta} e^{(\theta-\theta^*)^T l_{st}(\theta^*) - \frac{1}{2}(1-\epsilon)(\theta-\theta^*)^T I_{st}(\theta-\theta^*)}d\theta \\
&\overset{(*)}{=} (1+\epsilon)\omega(\theta^*)e^{\rho(\delta,\theta^*)}e^{\frac{1}{2(1-\epsilon)}l_{st}^T(\theta^*)I_{st}^{-1}l_{st}(\theta^*)}\int_{N_\delta} e^{-\frac{1}{2}(1-\epsilon)(\theta-u)^T I_{st}(\theta-u)}d\theta \\
&\leq (1+\epsilon)\omega(\theta^*)e^{\rho(\delta,\theta^*)}e^{\frac{1}{2(1-\epsilon)}l_{st}^T(\theta^*)I_{st}^{-1}l_{st}(\theta^*)}\int_{N_\delta\cup N_\delta^c} e^{-\frac{1}{2}(1-\epsilon)(\theta-u)^T I_{st}(\theta-u)}d\theta \\
\text{(Gaussian integral)} \quad &= (1+\epsilon)\omega(\theta^*)e^{\rho(\delta,\theta^*)}e^{\frac{1}{2(1-\epsilon)}L(\theta^*)}(2\pi)^{\frac{2\tilde{k}+2-c}{2}} \det((1-\epsilon)I_{st})^{-\frac{1}{2}},
\end{aligned}$$

where we define $u = \theta^* + \frac{1}{1-\epsilon}(\hat{\theta}-\theta^*)$ and $\hat{\theta} = \theta^* + I_{st}^{-1}l_{st}(\theta^*)$ provided that I_{st} positive definite. We also use the identity in (*) by completing the square,

$$(\theta-\theta^*)^T l_{st}(\theta^*) - \frac{1}{2}(1-\epsilon)(\theta-\theta^*)^T I_{st}(\theta-\theta^*) = -\frac{1-\epsilon}{2}(\theta-u)^T I_{st}(\theta-u) + \frac{1}{2(1-\epsilon)}L(\theta^*). \tag{65}$$

For the lower bound, we have,

$$\begin{aligned}
\frac{Q(D_m^s, D_n^{t,U})}{P_{\theta^*}(D_m^s, D_n^{t,U})} &\geq \int_{N_\delta} \frac{P_\theta(D_m^s, D_n^{t,U})}{P_{\theta^*}(D_m^s, D_n^{t,U})}\omega(\theta)d\theta \\
&= \int_{N_\delta} e^{\log \frac{P_\theta(D_m^s, D_n^{t,U})}{P_{\theta^*}(D_m^s, D_n^{t,U})}}\omega(\theta)d\theta \\
\text{(Taylor Expansion)} \ &= \int_{N_\delta} e^{(\theta-\theta^*)^T l_{st}(\theta^*) - \frac{1}{2}(\theta-\theta^*)^T \tilde{I}_{st}(\theta')(\theta-\theta^*)}\omega(\theta)d\theta \\
\text{(Event } B) &\geq \omega(\theta^*)e^{-\rho(\delta,\theta^*)} \int_{N_\delta} e^{(\theta-\theta^*)^T l_{st}(\theta^*) - \frac{1}{2}(1+\epsilon)(\theta-\theta^*)^T I_{st}(\theta-\theta^*)}d\theta \\
&= \omega(\theta^*)^{-\rho(\delta,\theta^*)} e^{\frac{1}{2(1+\epsilon)}L(\theta^*)} \int_{N_\delta} e^{-\frac{(1+\epsilon)}{2}(\theta-u)^T I_{st}(\theta-u)}d\theta \\
&= \omega(\theta^*)^{-\rho(\delta,\theta^*)} e^{\frac{1}{2(1+\epsilon)}L(\theta^*)} \Bigg[\int_{\mathbb{R}^{2\tilde{k}+2-c}} e^{-\frac{(1+\epsilon)}{2}(\theta-u)^T I_{st}(\theta-u)}d\theta \\
&\quad - \int_{N_\delta^c} e^{-\frac{(1+\epsilon)}{2}(\theta-u)^T I_{st}(\theta-u)}d\theta \Bigg].
\end{aligned}$$

Here we define $u = \theta^* + \frac{1}{1+\epsilon}(\hat{\theta} - \theta^*)$ and $\hat{\theta} = \theta^* + I_{st}^{-1} l_{st}(\theta^*)$. Since we restrict to the event C and the norm is w.r.t. I_0, given Condition 2 such that $I_{st} \succcurlyeq (n \wedge m) I_0$, we have that for any $\theta \in N_\delta^c$,

$$\begin{aligned}
(\theta-u)^T I_{st}(\theta-u) &\geq (n\wedge m)(\theta-u)^T I_0(\theta-u) && (66)\\
\text{(Definition of } \|\cdot\|) \ &= (n\wedge m)\|\theta-u\|^2 && (67)\\
&= (n\wedge m)\|\theta-\theta^* - \frac{1}{1+\epsilon}(\hat{\theta}-\theta^*)\|^2 && (68)\\
&= (n\wedge m)\|\theta-\theta^* - \frac{1}{1+\epsilon}(I_{st}^{-1} l_{st}(\theta^*))\|^2 && (69)\\
&\geq (n\wedge m)(\|\theta-\theta^*\| - \frac{1}{1+\epsilon}\|I_{st}^{-1} l_{st}(\theta^*)\|)^2 && (70)\\
&\geq (n\wedge m)(\|\theta-\theta^*\| - \frac{1}{1+\epsilon}\sqrt{l_{st}^T(\theta^*) I_{st}^{-1} I_0 I_{st}^{-1} l_{st}(\theta^*)})^2 && (71)\\
&\geq (n\wedge m)(\|\theta-\theta^*\| - \frac{1}{1+\epsilon}\sqrt{\frac{1}{n\wedge m} l_{st}^T(\theta^*) I_{st}^{-1} l_{st}(\theta^*)})^2 && (72)\\
\text{(Event C)} \ &\geq (n\wedge m)(\delta - \frac{1}{1+\epsilon}\sqrt{\delta^2})^2 && (73)\\
&\geq \frac{\epsilon^2}{(1+\epsilon)^2}(n\wedge m)\delta^2. && (74)
\end{aligned}$$

Hence in the second integral in the lower bound, for any $\theta \in N_\delta^c$, the integrand is not greater than

$$e^{-\frac{(1+\epsilon)}{2}(\theta-u)^T I_{st}(\theta-u)} \leq e^{-\frac{(n\wedge m)\epsilon^2\delta^2}{4(1+\epsilon)}} e^{-\frac{(1+\epsilon)(n\wedge m)\|\theta-u\|^2}{4}}. \tag{75}$$

By expanding the terms, using the Gaussian integration and rearranging the integration, we have the lower bound and this completes the proof of this lemma. ■

With substantially small δ and ϵ, the integrand of the KL divergence term will approach $\frac{2\tilde{k}+2-c}{2}\log\frac{1}{2\pi}+\log\frac{1}{\omega(\theta^*)}+\frac{1}{2}\log\det(I_{st})-\frac{1}{2}L(\theta^*)$, hence we define the remaining term R_{st} by

$$R_{st}=\frac{P_{\theta^*}(D_m^s,D_n^{t,U})}{Q(D_m^s,D_n^{t,U})}-\frac{2\tilde{k}+2-c}{2}\log\frac{1}{2\pi}-\log\frac{1}{\omega(\theta^*)}-\frac{1}{2}\log\det(I_{st})+\frac{1}{2}L(\theta^*). \tag{76}$$

Using the similar argument in Zhu (2020) and Clarke and Barron (1990), we can show that the expected remaining term is upper-bounded and lower-bounded by

$$\mathbb{E}[R_{st}]\geq-\log(1+\epsilon)-\rho(\delta,\theta^*)-\frac{\epsilon}{2(1-\epsilon)}(2\tilde{k}+2-c)+\frac{2\tilde{k}+2-c}{2}\log\frac{1}{1-\epsilon} \tag{77}$$

$$+\mathbb{P}((A\cap B)^c)\left(\log\mathbb{P}((A\cap B)^c)+\frac{2\tilde{k}+2-c}{2}\log\frac{1}{2\pi}\right)-\mathbb{P}((A\cap B)^c)\log\frac{\det(I_{st})^{\frac{1}{2}}}{\omega(\theta^*)}, \tag{78}$$

and

$$\mathbb{E}[R_{st}]\leq\rho(\delta,\theta^*)+\frac{\epsilon}{2(1+\epsilon)}(2\tilde{k}+2-c)+\frac{2\tilde{k}+2-c}{2}\log\frac{1}{1+\epsilon}-\log\left(1-2^{\frac{2\tilde{k}+2-c}{2}}e^{-\epsilon^2(m\wedge n)\delta^2/8}\right)$$

$$+\mathbb{E}[L(\theta^*)\mathbf{1}_{(B\cap C)^c}]+\mathbb{P}((B\cap C)^c)\left(\frac{2\tilde{k}+2-c}{2}\log\frac{1}{2\pi}+|\log\int_{N_\delta}\omega(\theta)d\theta|+\log\frac{\det(I_{st})^{\frac{1}{2}}}{\omega(\theta^*)}\right)$$

$$+\mathbb{P}((B\cap C)^c)\mathbb{E}\left[\sup_{\theta,\theta'}(\theta-\theta^*)\nabla\log P_{\theta'}(D_s^m,D_t^{U,n})\right]$$

$$+\mathbb{P}((B\cap C)^c)^{\frac{1}{2}}\mathbb{E}\left[\sup_{\theta,\theta'}(\theta-\theta^*)\nabla^2\log P_{\theta'}(D_s^m,D_t^{U,n})\right]^{\frac{1}{2}}.$$

By application of Condition 2 in Assumption 2, with sufficiently small δ, the upper bound will go to zero if the probability of the data pair $D_t^{U,n}$ and D_s^m belong to the set $P(A^c)$, $P(B^c)$ and $P(C^c)$ is $o(\frac{1}{n\vee m})$. In the following, we will show that the probability of A^c, B^c and C^c will decay exponentially fast with $m\wedge n$ so that the expected remaining term will converge as $o(\frac{1}{n\vee m})$ under the regime that $m=cn^p$ for some $c>0$ and finite $p>0$.

Lemma 16 *Assume condition 4 holds so that for all $\theta\in N_\delta$, let $v=n\vee m$, then for sufficiently small δ, there is an $r>0$ and $\rho>0$ so that,*

$$\mathbb{P}((D_t^{U,n},D_s^m)\in A^c(\delta,e^{-vr}))=O(e^{-(m\wedge n)\rho}). \tag{79}$$

Proof For any given $r'>0$, we define the event

$$U=\left\{\int_{N_\delta}\omega(\theta)P_\theta(D_t^{U,n},D_s^m)d\theta>e^{-vr'}P_{\theta^*}(D_t^{U,n},D_s^m)\right\}. \tag{80}$$

We can bound the probability of A^c by

$$\begin{aligned}
\mathbb{P}\left(A^c\left(\delta, e^{-vr}\right)\right) &= \mathbb{P}\left(\int_{N_\delta} P(D_t^{U,n}, D_s^m \mid \theta)\omega(\theta)d\theta < e^{vr}\int_{N_\delta^c} P(D_t^{U,n}, D_s^m \mid \theta)\omega(\theta)d\theta\right) \\
&\leq \mathbb{P}\left(U \cap \left(\int_{N_\delta} P(D_t^{U,n}, D_s^m \mid \theta)\omega(\theta)d\theta \ < e^{vr}\int_{N_\delta^c} P(D_t^{U,n}, D_s^m \mid \theta)\omega(\theta)d\theta\right)\right) + \mathbb{P}\left(U^c\right) \\
&\leq \mathbb{P}\left(P\left(D_t^{U,n}, D_s^m \mid \theta^*\right) < e^{v(r+r')}\int_{N^c} \omega(\theta)P(D_t^{U,n}, D_s^m \mid \theta)d\theta\right) \\
&\quad + \mathbb{P}\left(e^{vr'}\int_{N_\delta} P(D_t^{U,n}, D_s^m \mid \theta)\omega(\theta)d\theta < P\left(D_t^{U,n}, D_s^m \mid \theta^*\right)\right).
\end{aligned}$$

For the first term, we use the argument in Clarke and Barron (1990) (Eq. (6.6)) and Zhu (2020) (Lemma 7) and it can be concluded that it is of the order of $O(e^{-(n\wedge m)r''})$ for some r'' under the Condition 3 for soundness of the parametric families. For the second term, define $Q(D_t^{U,n}, D_s^m \mid N_\delta) = \int_{N_\delta} P(X|\theta)\omega(\theta|N_\delta)d\theta$ and $\omega(\theta|N_\delta) = \frac{\omega(\theta)}{\int_{N_\delta}\omega(\theta)d\theta}$ and $\tilde{r} = r' - \frac{1}{v}\log\int_{N_\delta}\omega(\theta)d\theta$, we can write the probability as,

$$\mathbb{P}\left(e^{vr'}\int_{N_\delta} P(D_t^{U,n}, D_s^m \mid \theta)\omega(\theta)d\theta < P\left(D_t^{U,n}, D_s^m \mid \theta^*\right)\right) = \mathbb{P}\left(\log\frac{P(D_t^{U,n}, D_s^m \mid \theta^*)}{Q(D_t^{U,n}, D_s^m \mid N_\delta)} > v\tilde{r}\right) \tag{81}$$

$$\leq \mathbb{P}\left(\log P\left(D_t^{U,n}, D_s^m \mid \theta^*\right) - \int_{N_\delta}\log P(D_t^{U,n}, D_s^m \mid \theta)\omega\left(\theta \mid N_\delta\right)d\theta > v\tilde{r}\right) \tag{82}$$

$$\leq \mathbb{P}\left(\int_{N_{s,\delta}}\log\frac{P\left(D_s^m \mid \theta_c^*, \theta_s^*\right)}{P\left(D_s^m \mid \theta_c, \theta_s\right)}\omega\left(\theta \mid N_{s,\delta}\right)d\theta + \int_{N_{t,\delta}}\log\frac{P\left(D_t^{U,n} \mid \theta_c^*, \theta_t^*\right)}{P\left(D_t^{U,n} \mid \theta_c, \theta_t\right)}\omega\left(\theta \mid N_{t,\delta}\right)d\theta > v\tilde{r}\right) \tag{83}$$

$$= \mathbb{P}\left(\sum_{i=1}^m g_s\left(Z_s^{(i)}\right) + \sum_{j=1}^n g_t\left(X_t^{(j)}\right) > v\tilde{r}\right) \tag{84}$$

$$\leq \mathbb{P}\left(\frac{1}{v}\sum_{i=1}^m g_s\left(Z_s^{(i)}\right) > \tilde{r}/2\right) + \mathbb{P}\left(\frac{1}{v}\sum_{j=1}^n g_t\left(X_t^{(j)}\right) > \tilde{r}/2\right) \tag{85}$$

$$\leq \mathbb{P}\left(\frac{1}{m}\sum_{i=1}^m g_s\left(Z_s^{(i)}\right) > \tilde{r}/2\right) + \mathbb{P}\left(\frac{1}{n}\sum_{j=1}^n g_t\left(X_t^{(j)}\right) > \tilde{r}/2\right), \tag{86}$$

where we define,

$$g_s(Z_s^{(i)}) := \int_{N_{s,\delta}} \log \frac{P\left(Z_s^{(i)} \mid \theta_s^*, \theta_c^*\right)}{P(Z_s^{(i)} \mid \theta_s, \theta_c)} \omega\left(\theta_s, \theta_c \mid N_{s,\delta}\right) d\theta_s d\theta_c, \tag{87}$$

$$g_t(X_t^{(j)}) := \int_{N_{t,\delta}} \log \frac{P\left(X_t^{(j)} \mid \theta_t^*, \theta_c^*\right)}{P(X_t^{(j)} \mid \theta_t, \theta_c)} \omega\left(\theta_t, \theta_c \mid N_{t,\delta}\right) d\theta_s d\theta_c, \tag{88}$$

$$\omega(\theta_c, \theta_s \mid N_{s,\delta}) := \frac{\omega(\theta_c, \theta_s)}{\int_{N_{s,\delta}} \omega(\theta_c, \theta_s) d\theta_c d\theta_s}, \tag{89}$$

$$\omega(\theta_c, \theta_t \mid N_{t,\delta}) := \frac{\omega(\theta_c, \theta_t)}{\int_{N_{t,\delta}} \omega(\theta_c, \theta_t) d\theta_c d\theta_t}. \tag{90}$$

In this case, we use a slightly different notation that $N_{s,\delta} = \{\theta_{sc} : \|\theta_{sc} - \theta_{sc}^*\| \leq \delta\}$, where $\theta_{sc} = (\theta_s, \theta_c)$ denotes the source parameters and the norm is w.r.t. the Fisher information matrix I_s, e.g., $\|\theta_{sc}\|^2 = \theta_{sc}^T I_s \theta_{sc}$. Similarly, $N_{t,\delta} = \{\theta_{tc} : \|\theta_{tc} - \theta_{tc}^*\| \leq \delta\}$ where $\theta_{tc} = (\theta_t, \theta_c)$ denotes the target parameters and norm is w.r.t. the Fisher information matrix I_t as defined previously. The second inequality holds due to that $I_t \prec I_t + I_{sc}$ and $I_s \prec I_s + I_{tc}$ for $I_{sc} = -\mathbb{E}_{\theta_s^*, \theta_c^*}[\nabla^2 \log P(X_s, Y_s | \theta_t^*, \theta_c^*)]$ and $I_{tc} = -\mathbb{E}_{\theta_t^*, \theta_c^*}[\nabla^2 \log P(X_t | \theta_t^*, \theta_c^*)]$ the fisher information matrix w.r.t. θ_c^* in both source and target domains, with the fact that $N_\delta \subset N_{s,\delta}$ and $N_\delta \subset N_{t,\delta}$. If the source and target domain share the same parameters (e.g., $c = \tilde{k} + 1$), then our case generalizes to Lemma 7 in Zhu (2020).

■

Lemma 17 *Assume condition 5 holds so that for sufficiently small δ, there is some $\rho > 0$ such that,*

$$\mathbb{P}((D_t^{U,n}, D_s^m) \in B^c(\delta, \epsilon)) = O(e^{-\rho(m \wedge n)}). \tag{91}$$

Proof The proof exactly follow Zhu (2020) with similar assumptions, which is omitted here. ■

Lemma 18 *Assume condition 6 holds, then for sufficiently small δ, there is a $\rho > 0$ so that,*

$$\mathbb{P}((D_t^{U,n}, D_s^m) \in C^c(\delta)) = O(e^{-(m \wedge n)\rho}). \tag{92}$$

Proof We firstly expand the term $L(\theta^*)$ by:

$$L(\theta^*) = l_{st}^T I_{st}^{-1} l_{st}^T \tag{93}$$

$$= \sum_{i=1}^m l_{s,i}^T I_{st}^{-1} l_{s,i} + \sum_{i \neq k}^m l_{s,i}^T I_{st}^{-1} l_{s,k} + \sum_{i=1}^n l_{t,i}^T I_{st}^{-1} l_{t,i} + \sum_{i \neq k}^n l_{t,i}^T I_{st}^{-1} l_{t,k} \tag{94}$$

$$+ 2 \sum_{i=1}^n \sum_{k=1}^m l_{t,i}^T I_{st}^{-1} l_{s,k} \tag{95}$$

Then we have that,

$$\begin{aligned}
&\mathbb{P}((D_t^{U,n}, D_s^m) \in C^c(\delta)) = \mathbb{P}(L(\theta^*) > (n \wedge m)\delta^2) \\
&\leq \mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m} l_{s,i}^T I_{st}^{-1} l_{s,i} \geq \frac{(n \wedge m)\delta^2}{6m}\right) + \mathbb{P}\left(\frac{1}{m(m-1)}\sum_{i\neq k}^{m} l_{s,i}^T I_{st}^{-1} l_{s,k} \geq \frac{(n \wedge m)\delta^2}{6m(m-1)}\right) \\
&+ \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} l_{t,i}^T I_{st}^{-1} l_{t,i} \geq \frac{(n \wedge m)\delta^2}{6n}\right) + \mathbb{P}\left(\frac{1}{n(n-1)}\sum_{i\neq k}^{n} l_{t,i}^T I_{st}^{-1} l_{t,k} \geq \frac{(n \wedge m)\delta^2}{6n(n-1)}\right) \\
&+ \mathbb{P}\left(\frac{2}{nm}\sum_{i=1}^{n}\sum_{k=1}^{m} l_{t,i}^T I_{st}^{-1} l_{s,k} \geq \frac{(n \wedge m)\delta^2}{3nm}\right)
\end{aligned}$$

We first consider the case where $m = cn^p$ for $p \geq 1$, then we can show that these five terms will decay exponentially fast. We first bound the expected value by

$$\begin{aligned}
\mathbb{E}[l_{s,i}^T I_{st} l_{s,i}] &= \mathrm{Tr}(I_{st}^{-1}\mathbb{E}[l_{s,i}^T l_{s,i}]) && (96)\\
&\leq \frac{1}{n \wedge m}\mathrm{Tr}(I_0^{-1}) I_s && (97)\\
&\leq \frac{1}{n \wedge m}\mathrm{Tr}(I_0^{-1}) I_0) && (98)\\
&= \frac{2\tilde{k} + 2 - c}{n \wedge m} && (99)
\end{aligned}$$

since $I_s \prec I_0$ due to the Condition 2. Also we have for large m,

$$\mathbb{E}[l_{t,i}^T I_{st} l_{t,i}] = \frac{2\tilde{k} + 2 - c}{m}, \tag{100}$$

and

$$\begin{aligned}
\mathbb{E}[l_{t,i}^T I_{st} l_{t,k}] &= 0, && (101)\\
\mathbb{E}[l_{s,i}^T I_{st} l_{s,k}] &= 0, && (102)\\
\mathbb{E}[l_{t,i}^T I_{st} l_{s,k}] &= 0 && (103)
\end{aligned}$$

due to that $l_{s,i}$ and $l_{s,k}$ are mutually independent. Since the Condition 6 holds, we will use the Chernoff bound again so that the inequality is bounded by $O(e^{-\rho(n\wedge m)})$ for some $\rho > 0$ under the case that $m = cn^p$ for $p \geq 1$ as Zhu (2020) (Lemma 9) suggested, where the details are omitted here. For the case where $m = cn^p$ for some $0 < p < 1$, since for large $n \gg m$,

$$\mathbb{E}[l_{t,i}^T I_{st} l_{t,i}] = \frac{2\tilde{k} + 2 - c}{n}. \tag{104}$$

We can upper bound the term on the source score function by,

$$\begin{aligned}
\mathbb{E}[l_{s,i}^T I_{st} l_{s,i}] &= \mathrm{Tr}(I_{st}^{-1}\mathbb{E}[l_{s,i}^T l_{s,i}]) && (105)\\
&\leq \frac{2\tilde{k} + 2 - c}{n \wedge m}. && (106)
\end{aligned}$$

Then similar argument can be made that the probability is bounded by $O(e^{-\rho'(n\wedge m)})$ for some $\rho' > 0$, and this completes the proof for all $p > 0$.

■

Overall, putting everything together we complete the proof.

■

A.2 Proof of Theorem 5

Proof We firstly show that given any prior over Θ_s and Θ_t,

$$
\begin{aligned}
&I(Y_t'; \Theta_t, \Theta_s | D_t^{U,n}, D_s^m, X_t') \\
&= I(\Theta_t, \Theta_s; Y_t', X_t', D_t^{U,n}, D_s^m) - I(\Theta_t, \Theta_s; X_t', D_t^{U,n}, D_s^m) \\
&= D(P_{\Theta_t,\Theta_s}(D_t^{U,n}, D_s^m, Y_t', X_t') \| Q(D_t^{U,n}, D_s^m, Y_t', X_t')) \\
&\quad - D(P_{\Theta_s,\Theta_t}(D_s^m, D_t^{U,n}, X_t') \| Q(D_s^m, D_t^{U,n}, X_t')) \\
&= \int \left(\mathbb{E}_{\theta_s,\theta_t} \left[\log \frac{P_{\theta_t,\theta_s}(D_t^{U,n}, D_s^m, Y_t', X_t')}{Q(D_t^{U,n}, D_s^m, Y_t', X_t')} \right] - \mathbb{E}_{\theta_s,\theta_t} \left[\log \frac{P_{\theta_t,\theta_s}(D_s^m, D_t^{U,n}, X_t')}{Q(D_s^m, D_t^{U,n}, X_t')} \right] \right) \omega(\theta_s, \theta_t) d\theta_s d\theta_t \\
&= \int \left(\mathbb{E}_{\theta_s,\theta_t} \left[\log \frac{P_{\theta_t}(Y_t' | X_t')}{Q(Y_t' | D_t^{U,n}, D_s^m, X_t')} \right] \right) \omega(\theta_s, \theta_t) d\theta_s d\theta_t,
\end{aligned}
$$

where in the last equality we use the chain rule and the assumption that both source and target data are drawn in an i.i.d. way under Assumption 1. The mutual information density at $\Theta_s = \theta_s^*$ and $\Theta_t = \theta_t^*$ is then given by

$$
\begin{aligned}
\mathcal{R}(b) &= I(Y_t'; \theta_t^*, \theta_s^* | D_t^{U,n}, D_s^m, X_t') \\
&= \mathbb{E}_{\theta_s^*,\theta_t^*} \left[\log \frac{P_{\theta_t^*}(Y_t' | X_t')}{Q(Y_t' | D_t^{U,n}, D_s^m, X_t')} \right],
\end{aligned}
$$

which completes the proof. ■

A.3 Proof of Theorem 7

Proof We can show that the expected excess risk can be bounded by

$$
\begin{aligned}
\mathcal{R}(b) &= \mathbb{E}_{\theta_t^*,\theta_s^*}\left[\ell(b,Y_t') - \ell(b^*,Y_t')\right] \\
&= \mathbb{E}_{D_s^m,D_t^{U,n},X_t',Y_t'}\mathbb{E}_{Y_t'}\left[\ell(b,Y_t') - \ell(b^*,Y_t')|D_s^m,D_t^{U,n},X_t'\right] \\
&= \mathbb{E}_{D_s^m,D_t^{U,n},X_t'}\sum_{y_t'}\left(\ell(b,y_t') - \ell(b^*,y_t')\right)P_{\theta_s^*,\theta_t^*}(y_t'|D_s^m,D_t^{U,n},X_t') \\
&= \mathbb{E}_{D_s^m,D_t^{U,n},X_t'}\sum_{y_t'}\left(\ell(b,y_t') - \ell(b^*,y_t')\right)(P_{\theta_s^*,\theta_t^*}(y_t'|D_s^m,D_t^{U,n},X_t') \\
&\quad - Q(y_t'|D_s^m,D_t^{U,n},X_t') + Q(y_t'|D_s^m,D_t^{U,n},X_t')) \\
&\overset{(a)}{\leq} \mathbb{E}_{D_s^m,D_t^{U,n},X_t'}\sum_{y_t'}\left(\ell(b,y_t') - \ell(b^*,y_t')\right)(P_{\theta_s^*,\theta_t^*}(y_t'|D_s^m,D_t^n,X_t') - Q(y_t'|D_s^m,D_t^{U,n},X_t')) \\
&\overset{(b)}{\leq} M\mathbb{E}_{D_s^m,D_t^{U,n},X_t'}\sum_{y_t'}(P_{\theta_s^*,\theta_t^*}(y_t'|D_s^m,D_t^{U,n},X_t') - Q(y_t'|D_s^m,D_t^{U,n},X_t')) \\
&\overset{(c)}{\leq} M\mathbb{E}_{D_s^m,D_t^{U,n},X_t'}\sqrt{2D\left(P_{\theta_s^*,\theta_t^*}(Y_t'|D_s^m,D_t^{U,n},X_t')\|Q(y_t'|D_s^m,D_t^{U,n},X_t')\right)} \\
&\overset{(d)}{\leq} M\sqrt{2\mathbb{E}_{D_s^m,D_t^{U,n},X_t'}D\left(P_{\theta_s^*,\theta_t^*}(Y_t'|D_s^m,D_t^{U,n},X_t')\|Q(Y_t'|D_s^m,D_t^{U,n},X_t')\right)} \\
&= M\sqrt{2D\left(P_{\theta_t^*}\|Q|D_s^m,D_t^{U,n},X_t'\right)} \\
&= M\sqrt{2D(P_{\theta_t^*}(Y_t'|X_t')\|Q(Y_t'|D_t^{U,n},D_s^m,X_t'))} \\
&= M\sqrt{2I(Y_t';\Theta_t=\theta_t^*,\Theta_s=\theta_s^*|D_s^m,D_t^{U,n},X_t')},
\end{aligned}
$$

where in (a) we use the definition of Q, then (b) holds since we assume the loss function is bounded, (c) follows from the Pinsker's inequality, (d) holds from the Jensen's inequality. ■

A.4 Proof of Theorem 11

We firstly consider the scenario for covariate shift condition where $P_S(X) \neq P_T(X)$ and $P_S(Y|X) = P_T(Y|X)$.

Proof Knowing the conditions $\theta_{Y_{x_i}}^{s*} = \theta_{Y_{x_i}}^{t*}$ for every $i = 1, 2, \cdots, k$, we choose the prior distribution $\omega(\Theta_s, \Theta_t)$ as

$$
\omega(\Theta_s,\Theta_t) = \omega(\Theta_X^t)\omega(\Theta_X^s)\omega(\Theta_{Y_X}^{st}). \tag{107}
$$

In the causal setting, Θ_X^s is usually considered as independent of Θ_X^t and $\Theta_{Y_X}^s$. We also set the parameter $\Theta_{Y_X}^t = \Theta_{Y_X}^s$ from the assumption $P_S(Y|X) = P_T(Y|X)$ and denote it by $\Theta_{Y_X}^{st}$. With a proper prior distribution, we will arrive at the asymptotic estimation of the

expected excess risk as

$$
\begin{aligned}
&D(P_{\theta_t^*,\theta_s^*}(D_t^{\mathrm{U},n}, X_t', Y_t', D_s^m)\|Q(D_t^{\mathrm{U},n}, X_t', Y_t', D_s^m)) - D(P_{\theta_t^*,\theta_s^*}(D_t^{\mathrm{U},n}, X_t', D_s^m)\|Q(D_t^{\mathrm{U},n}, X_t', D_s^m)) \\
&= D(P_{\theta_X^{t*}}(D_t^{\mathrm{U},n}, X_t')\|Q(D_t^{\mathrm{U},n}, X_t')) + D(P_{\theta_X^{s*}}(X_s^m)\|Q(X_s^m)) + D(P_{\theta_{Y_X}^*}(Y_{X',t}', Y_{X,s}^m)\|Q(Y_{X',t}', Y_{X,s}^m)) \\
&\quad - D(P_{\theta_X^{t*}}(D_t^{\mathrm{U},n}, X_t')\|Q(D_t^{\mathrm{U},n}, X_t')) + D(P_{\theta_X^{s*}}(X_s^m)\|Q(X_s^m)) - D(P_{\theta_{Y_X}^*}(Y_{X,s}^m)\|Q(Y_{X,s}^m)) \\
&= D(P_{\theta_{Y_X}^*}(Y_{X',t}', Y_{X,s}^m)\|Q(Y_{X',t}', Y_{X,s}^m)) - D(P_{\theta_{Y_X}^*}(Y_{X,s}^m)\|Q(Y_{X,s}^m)) \qquad (108) \\
&= \frac{1}{2}\log\det\mathbf{I}_1 + \frac{k}{2}\log\frac{1}{2\pi e} + \log\frac{1}{\omega(\theta_{Y_X}^*)} - \frac{1}{2}\log\det\mathbf{I}_0 - \frac{k}{2}\log\frac{1}{2\pi e} - \log\frac{1}{\omega(\theta_{Y_X}^*)} + o(\frac{1}{m}) \qquad (109) \\
&= \frac{1}{2}\frac{\log\det(\mathbf{I}_1)}{\log\det(\mathbf{I}_0)} + o(\frac{1}{m}), \qquad (110)
\end{aligned}
$$

where we use the i.i.d. property of the data distribution and the independence property of the prior distribution among Θ_X^t, Θ_X^s and $\Theta_{Y_X}^{st}$. Since $Y_{X',t}'$ and $Y_{X,s}^m$ are parameterized by the same set of parameters $\theta_{Y_X}^*$, we denote the Fisher information matrix of $P_{\theta_{Y_X}^*}(Y)$ for source and target domains by

$$
\begin{aligned}
&I(\theta_{Y_{x_i}}^*) = \mathbb{E}_{Y_{x_i}}[\partial^2 \log P_{\theta_{Y_{x_i}}^*}(Y)/(\partial\theta_{Y_{x_i}})^2], \text{ for } i = 1, 2, \cdots, k, \qquad (111) \\
&I_s(\theta_{Y_X}^*) = -\mathbb{E}_{Y_{X_s}}\left[\partial^2 \log P_{\theta_{Y_X}^*}(Y_X)/\partial\theta_j\partial\theta_k\right]_{j,k=1,2,\cdots,k} = \mathrm{diag}[P_{\theta_X^{s*}}(X = x_i) * I(\theta_{Y_{x_i}}^*)]_{i=1,\cdots,k}, \qquad (112) \\
&I_t(\theta_{Y_X}^*) = -\mathbb{E}_{Y_{X_t}}\left[\partial^2 \log P_{\theta_{Y_X}^*}(Y_X)/\partial\theta_j\partial\theta_k\right]_{j,k=1,2,\cdots,k} = \mathrm{diag}[P_{\theta_X^{t*}}(X = x_i) * I(\theta_{Y_{x_i}}^*)]_{i=1,\cdots,k}, \qquad (113)
\end{aligned}
$$

due to the mutually independence property of Y_{X_i}. Then $\mathbf{I}_1$ and $\mathbf{I}_0$ are expressed as follows.

$$
\begin{aligned}
\mathbf{I}_0 &= mI_s(\theta_{Y_X}^*), \qquad (114) \\
\mathbf{I}_1 &= mI_s(\theta_{Y_X}^*) + I_t(\theta_{Y_X}^*). \qquad (115)
\end{aligned}
$$

With the assumptions that the Fisher information matrix around true $\theta_{Y_X}^*$ are bounded and positive definite, we can calculate the excess risk by

$$
\begin{aligned}
\mathcal{R}(b) &= \frac{1}{2}\frac{\log\det(\mathbf{I}_1)}{\log\det(\mathbf{I}_0)} + o(\frac{1}{m}) \qquad (116) \\
&= \frac{1}{2}\log\det\left(\mathbf{I}_k + \frac{1}{m}I_t(\theta_{Y_X}^{t*})I_s^{-1}(\theta_{Y_X}^{s*})\right) + o(\frac{1}{m}). \qquad (117)
\end{aligned}
$$

We then use the expansion of determinant:

$$
\det(\mathbf{I} + \frac{1}{m}A) = 1 + \frac{1}{m}\mathrm{Tr}(A) + o(1/m). \qquad (118)
$$

As a consequence,

$$\mathcal{R}(b) = \frac{1}{2}\log\left(1 + \frac{1}{m}\operatorname{Tr}(I_t(\theta^*_{Y_X})I_s^{-1}(\theta^*_{Y_X})) + o(1/m)\right) + o(\frac{1}{m}) \tag{119}$$

$$= \frac{1}{2}\log\left(1 + \frac{1}{m}\sum_{i=1}^{k}\frac{P_{\theta^{t*}_X}(X=x_i)}{P_{\theta^{s*}_X}(X=x_i)} + o(1/m)\right) + o(\frac{1}{m}) \tag{120}$$

$$\asymp \left(\frac{\sum_{i=1}^{k}\frac{P_{\theta^{t*}_X}(X=x_i)}{P_{\theta^{s*}_X}(X=x_i)}}{m}\right) \tag{121}$$

$$\asymp \frac{k}{m}. \tag{122}$$

given that $P_{\theta^{t*}_X}(X=x_i)$ and $P_{\theta^{t*}_X}(X=x_i)$ are positive and bounded for any i. In other word, the convergence is guaranteed only when the source and target domains share the same support of the input X. For the case $\theta^{s*}_X = \theta^{t*}_X$, using the same procedure, by choosing

$$\omega(\Theta_s, \Theta_t) = \omega(\Theta^{st}_X)\omega(\Theta^{st}_{Y_X}). \tag{123}$$

we will also arrive at

$$\mathcal{R}(b) = \frac{1}{2}\log\left(1 + \frac{1}{m}\operatorname{Tr}(I_t(\theta^{t*}_{Y_X})I_s^{-1}(\theta^{s*}_{Y_X})) + o(1/m)\right) + o(\frac{1}{m}) \tag{124}$$

$$\asymp \frac{k}{m}. \tag{125}$$

which leads to the same rate and completes the proof. ■

Next we will look at the concept drift scenario where $P_S(Y|X) \neq P_T(Y|X)$ and $P_S(X) = P_T(X)$.

Proof Knowing the conditions $\theta^{s*}_{Y_{x_i}} \neq \theta^{t*}_{Y_{x_i}}$ for every $i = 1, 2, \cdots, k$, if $P_S(X) = P_T(X)$, we choose the prior distribution $\omega(\Theta_s, \Theta_t)$ as

$$\omega(\Theta_s, \Theta_t) = \omega(\Theta^{st}_X)\omega(\Theta^{s}_{Y_X})\omega(\Theta^{t}_{Y_X}). \tag{126}$$

following the similar machinery in the covariate shift conditions. Then the mixture distribution Q becomes

$$Q(Y'_t|D_t^{\mathrm{U},n}, D_s^m, X'_t) \tag{127}$$

$$= \frac{\int P_{\theta_t}(D_t^{\mathrm{U},n}, X'_t, Y'_t)P_{\theta_s}(D_s^m)\omega(\theta_t,\theta_s)d\theta_t d\theta_s}{\int P_{\theta_t}(X'_t)P_{\theta_t}(D_t^{\mathrm{U},n})P_{\theta^*_s}(D_s^m)\omega(\theta_t,\theta_s)d\theta_t d\theta_s} \tag{128}$$

$$= \int P_{\theta_t}(Y'_t|X'_t)P(\theta_t,\theta_s|X'_t, D_s^m, D_s^{\mathrm{U},n})d\theta_s d\theta_t \tag{129}$$

$$= \int P_{\theta^t_{Y_X}}(Y'_t|X'_t)P(\theta^{st}_X, \theta^s_{Y_X}, \theta^t_{Y_X}|X'_t, D_s^m, D_s^{\mathrm{U},n})d\theta^s_{Y_X}d\theta^t_{Y_X}\theta^{st}_X \tag{130}$$

$$\overset{(a)}{=} \int P(Y'_t|X'_t, \theta_{Y_X})\omega(\theta_{Y_X})d\theta_{Y_X} \tag{131}$$

$$= \int P_{\theta_{Y_{X'_t}}}(Y'_t)\omega(\theta_{Y_{X'_t}})d\theta_{Y_{X'_t}}, \tag{132}$$

where (a) holds because X'_t, D_s^m and $D_s^{\mathrm{U},n}$ are all independent of $\Theta^t_{Y_X}$. Therefore, the excess risk becomes,

$$\begin{aligned}\mathcal{R}(b) =&\mathbb{E}_{\theta_s^*,\theta_t^*,X'_t,Y'_t}\left[\log\frac{P(Y'_t|\theta^{t*}_{Y|X},X'_t)}{Q(Y'_t|D_t^{\mathrm{U},n},D_s^m,X'_t)}\right]\\ =&\mathbb{E}_{\theta^{t*}_X}[\mathrm{KL}(P_{\theta^{t*}_{Y_{X'_t}}}(Y'_t)\|Q(Y'_t|X'_t)].\end{aligned}\tag{133}$$

If $P_S(X)\neq P_T(X)$, we choose the prior distribution $\omega(\Theta_s,\Theta_t)$ as,

$$\omega(\Theta_s,\Theta_t)=\omega(\Theta^t_X)\omega(\Theta^s_X)\omega(\Theta^s_{Y_X})\omega(\Theta^t_{Y_X}),\tag{134}$$

where we will end up with the same results as (133). ■

A.5 Proof of Theorem 12

Before proving Theorem 12, we first restate the definition for Fisher information matrix and define extra quantities for proving purposes.

$$I_s=-\mathbb{E}_{\theta_s^*}[\nabla^2\log P(X_s,Y_s|\theta_s^*)],\tag{135}$$
$$I_t=-\mathbb{E}_{\theta_t^*}[\nabla^2\log P(X_t|\theta_t^*)],\tag{136}$$
$$I_{t,X,Y}=-\mathbb{E}_{\theta_t^*}[\nabla^2\log P(X_t,Y_t|\theta_t^*)],\tag{137}$$
$$I_0=-\mathbb{E}_{\theta^*}[\nabla^2\log P(X_t,X_s,Y_s|\theta^*)],\tag{138}$$
$$I_{t,Y,U}=-\mathbb{E}_{\theta_t^*}[\nabla^2_{\theta_Y}\log P(X_t|\theta^{t*}_Y,\theta^{t*}_{X_Y})],\tag{139}$$
$$I_{t,Y}=-\mathbb{E}_{Y_t}[\nabla^2_{\theta_Y}\log P(Y_t|\theta^{t*}_Y)],\tag{140}$$
$$I_{s,Y}=-\mathbb{E}_{Y_s}[\nabla^2_{\theta_Y}\log P(Y_s|\theta^{s*}_Y)],\tag{141}$$
$$I_{t,X_Y,U}=-\mathbb{E}_{\theta_t^*}[\nabla^2_{\theta_{X_Y}}\log P(X_t|\theta^{t*}_Y,\theta^{t*}_{X_Y})],\tag{142}$$
$$I_{t,X_Y}=-\mathbb{E}_{\theta_t^*}[\nabla^2_{\theta_{X_Y}}\log P(X_t|\theta^{t*}_{X_{Y_t}})],\tag{143}$$
$$I_{s,X_Y}=-\mathbb{E}_{\theta_s^*}[\nabla^2_{\theta_{X_Y}}\log P(X_s|\theta^{s*}_{X_{Y_s}})].\tag{144}$$

Now we will firstly consider the case $P_S(Y)\neq P_S(Y)$ and $P_S(X|Y)\neq P_T(X|Y)$.

Proof Knowing the conditions $\theta^{s*}_Y\neq\theta^{t*}_Y$ and $\theta^{s*}_{X_{y_i}}\neq\theta^{t*}_{X_{y_i}}$ for every $i=1,2,\cdots,k'$, we then choose the prior distribution $\omega(\Theta_s,\Theta_t)$ as

$$\omega(\Theta_s,\Theta_t)=\omega(\Theta^t_Y)\omega(\Theta^s_Y)\omega(\Theta^s_{X_Y})\omega(\Theta^t_{X_Y}).\tag{145}$$

With such a prior distribution, we will arrive at the asymptotic estimation of the KL divergence as,

$$\begin{aligned}&D(P_{\theta_t^*,\theta_s^*}(D_t^{\mathrm{U},n},D_s^m,X'_t,Y'_t)\|Q(D_t^{\mathrm{U},n},D_s^m,X'_t,Y'_t))\\ &\qquad=\frac{1}{2}\log\det\mathbf{I}_\theta+\log\frac{1}{2\pi e}+\log\frac{1}{\omega(\theta_s^*,\theta_t^*)}+o(\frac{1}{m\vee n}),\end{aligned}\tag{146}$$

where

$$\mathbf{I}_\theta = \begin{bmatrix} nI_t + I_{t,X,Y} & \mathbf{0} \\ \mathbf{0} & mI_s \end{bmatrix}. \tag{147}$$

We also have,

$$\begin{aligned} &D(P_{\theta_t^*,\theta_s^*}(D_t^{\mathrm{U},n}, X_t', D_s^m) \| Q(D_t^{\mathrm{U},n}, X_t', D_s^m)) \\ &\qquad = \frac{1}{2}\log\det\widetilde{\mathbf{I}}_\theta + \log\frac{1}{2\pi e} + \log\frac{1}{\omega(\theta_s^*,\theta_t^*)} + o(\frac{1}{m\vee n}), \end{aligned} \tag{148}$$

where

$$\widetilde{\mathbf{I}}_\theta = \begin{bmatrix} (n+1)I_t & \mathbf{0} \\ \mathbf{0} & mI_s \end{bmatrix}. \tag{149}$$

Then the regret can be calculated by

$$\mathcal{R}(b) = \frac{1}{2}\frac{\log\det(\mathbf{I}_\theta)}{\log\det(\widetilde{\mathbf{I}}_\theta)} + o(\frac{1}{m\vee n}) \tag{150}$$

$$= \frac{1}{2}\log\det\left(\mathbf{I}_{k'+1} + \frac{1}{n+1}(I_{t,X,Y} - I_t)I_t^{-1}\right) + o(\frac{1}{m\vee n}) \tag{151}$$

$$\asymp \log\left(1 + \frac{\mathrm{Tr}((I_{t,X,Y} - I_t)I_t^{-1})}{n+1})\right) \tag{152}$$

$$\asymp \frac{k'+1}{n+1}, \tag{153}$$

which completes the proof. ■

Now we turn to conditional shifting case $P_S(Y) = P_S(Y)$ and $P_S(X|Y) \neq P_T(X|Y)$.

Proof In this section, we define,

$$I_{t,U} = -\mathbb{E}_{\theta_t^*}\left[\frac{\partial^2 \log P(X_t|\theta_t^*)}{\partial\theta_Y \partial\theta_{X_{y_i}}}\right] \text{ for } i = 1, 2, \cdots, k'. \tag{154}$$

Knowing the conditions $\theta_Y^{s*} = \theta_Y^{t*}$ and $\theta_{X_{y_i}}^{s*} \neq \theta_{X_{y_i}}^{t*}$ for every $i = 1, 2, \cdots, k'$, we then choose the prior distribution $\omega(\Theta_s, \Theta_t)$ as

$$\omega(\Theta_s, \Theta_t) = \omega(\Theta_Y^{st})\omega(\Theta_{X_Y}^s)\omega(\Theta_{X_Y}^t). \tag{155}$$

where we denote the random variable for estimating θ_Y^{st*} by Θ_Y^{st}. With such a prior distribution, we will arrive at the asymptotic estimation of the KL divergence as,

$$\begin{aligned} &D(P_{\theta_t^*,\theta_s^*}(D_t^{\mathrm{U},n}, D_s^m, X_t', Y_t') \| Q(D_t^{\mathrm{U},n}, D_s^m, X_t', Y_t')) \\ &\qquad = \frac{1}{2}\log\det\mathbf{I}_\theta + \log\frac{1}{2\pi e} + \log\frac{1}{\omega(\theta_s^*,\theta_t^*)} + o(\frac{1}{m\vee n}), \end{aligned} \tag{156}$$

where the joint Fisher information matrix $\mathbf{I}_\theta$ is defined as,

$$\mathbf{I}_\theta = \begin{bmatrix} nI_{t,Y,U} + mI_{s,Y} + I_{t,Y} & nI_{t,U} & \mathbf{0} \\ nI_{t,U}^T & nI_{t,X_Y,U} + I_{t,X_Y} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & mI_{s,X_Y} \end{bmatrix}. \tag{157}$$

Here zero vectors are due to the mutually independence assumption between the distribution parameters and i.i.d. assumption on the source and target samples. We also have,

$$D(P_{\theta_t^*,\theta_s^*}(D_t^{\mathrm{U},n}, X_t', D_s^m)\|Q(D_t^{\mathrm{U},n}, X_t', D_s^m)) = \frac{1}{2}\log\det\widetilde{\mathbf{I}}_\theta + \log\frac{1}{2\pi e} + \log\frac{1}{\omega(\theta_s^*,\theta_t^*)} + o(\frac{1}{m\vee n}), \tag{158}$$

where

$$\widetilde{\mathbf{I}}_\theta = \begin{bmatrix} (n+1)I_{t,Y,U} + mI_{s,Y} & (n+1)I_{t,U} & \mathbf{0} \\ (n+1)I_{t,U}^T & (n+1)I_{t,X_Y,U} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & mI_{s,X_Y} \end{bmatrix}. \tag{159}$$

Assume $m = cn^p$ for some $p > 0$, as n goes to infinity, we define the scalars $\Delta_U = I_{t,Y,U} - I_{t,U}I_{t,X_Y,U}^{-1}I_{t,U}^T$ and $\Delta_s = I_{s,Y}$, then the regret can be calculated by

$$\begin{aligned}
\mathcal{R}(b) =& \frac{1}{2}\frac{\log\det(\mathbf{I}_\theta)}{\log\det(\widetilde{\mathbf{I}}_\theta)} + o(\frac{1}{m\vee n}) && (160)\\
=& \frac{1}{2}\left(\log\det(\mathbf{I}_k + \frac{1}{n+1}(I_{t,X_Y} - I_{t,X_Y,U})I_{t,X_Y,U}^{-1}) + \log\det(1 + \frac{I_{t,Y} - I_{t,Y,U}}{(n+1)\Delta_U + cn^p\Delta_s})\right) && (161)\\
& + o(\frac{1}{m\vee n}) && (162)\\
\asymp& \frac{k'}{n+1} + \frac{1}{(n+1)\vee n^p} && (163)\\
\asymp& \frac{k'}{n} + \frac{1}{n\vee n^p}, && (164)
\end{aligned}$$

where $\mathbf{I}_k$ denotes the identity matrix with dimension of $k \times k$. Since we assume $I_t \succ 0$ and $I_s \succ 0$, we have that $\Delta_U > 0$ and $\Delta_s > 0$. From the information processing perspective, the labelled target data always contains more information than unlabelled target data, hence we have both $I_{t,X_Y} - I_{t,X_Y,U} \succ 0$ and $I_{t,Y} - I_{t,Y,U} \succ 0$, which completes the proof. ■

Regarding the target shift scenario $P_S(Y) \neq P_S(Y)$ and $P_S(X|Y) = P_T(X|Y)$, we could follow the similar procedures as the label drifting case.

Proof Knowing the conditions $\theta_Y^{s*} = \theta_Y^{t*}$ and $\theta_{X_{y_i}}^{s*} \neq \theta_{X_{y_i}}^{t*}$ for every $i = 1, 2, \cdots, k'$, we choose the prior distribution as,

$$\omega(\Theta_s, \Theta_t) = \omega(\Theta_Y^s)\omega(\Theta_Y^t)\omega(\Theta_{X_Y}^{st}). \tag{165}$$

where we denote the random variables for estimating $\theta_{X_Y}^{t*}$ by $\Theta_{X_Y}^{st}$. Following the similar procedure as shown in the proof of conditional shift case, we can write,

$$\mathbf{I}_\theta = \begin{bmatrix} nI_{t,X_Y,U} + mI_{s,X_Y} + I_{t,X_Y} & nI_{t,U}^T & \mathbf{0} \\ nI_{t,U} & nI_{t,Y,U} + I_{t,Y} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & mI_{s,Y} \end{bmatrix} \tag{166}$$

and

$$\widetilde{\mathbf{I}}_\theta = \begin{bmatrix} (n+1)I_{t,X_Y,U} + mI_{s,X_Y} & (n+1)I_{t,U}^T & \mathbf{0} \\ (n+1)I_{t,U} & (n+1)I_{t,Y,U} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & mI_{s,Y} \end{bmatrix}. \tag{167}$$

where $I_{t,U}$ is defined in (154). We first consider the case where $m = cn^p$ for some $p \geq 1$, as n goes to infinity, we define the matrices $\Delta_U = I_{t,X_Y,U} - I_{t,U}^T I_{t,Y,U}^{-1} I_{t,U}$ and $\Delta_s = I_{s,X_Y}$, then the expected regret can be calculated by using the following argument

$$\det(\mathbf{I} + \frac{1}{n} A) = 1 + \frac{1}{n} \operatorname{Tr}(A) + o(1/n). \tag{168}$$

Then,

$$\begin{aligned}
\mathcal{R}(b) =& \frac{1}{2} \log\det(1 + \frac{1}{n+1}(I_{t,Y} - I_{t,Y,U}) I_{t,Y,U}^{-1}) && (169)\\
&+ \frac{1}{2} \log\det(I_{t,X_Y} - I_{t,X_Y,U} + (n+1)\Delta_U + cn^p \Delta_s) && (170)\\
&- \frac{1}{2} \log\det((n+1)\Delta_U + cn^p \Delta_s) + o(\frac{1}{m \vee n}) && (171)\\
=& \frac{1}{2} \log\det(1 + \frac{1}{n+1}(I_{t,Y} - I_{t,Y,U}) I_{t,Y,U}^{-1}) && (172)\\
&+ \frac{1}{2} \log(\mathbf{I}_{k'} + \frac{1}{cn^p}(I_{t,X_Y} - I_{t,X_Y,U} + (n+1)\Delta_U)\Delta_s^{-1}) && (173)\\
&- \frac{1}{2} \log(\mathbf{I}_{k'} + \frac{1}{cn^p}((n+1)\Delta_U \Delta_s^{-1})) + o(\frac{1}{m \vee n}) && (174)\\
\asymp& \frac{(I_{t,Y} - I_{t,Y,U}) I_{t,Y,U}^{-1}}{n+1} + \frac{\operatorname{Tr}((I_{t,X_Y} - I_{t,X_Y,U} + (n+1)\Delta_U)\Delta_s^{-1}))}{cn^p} - \frac{\operatorname{Tr}((n+1)\Delta_U \Delta_s^{-1})}{cn^p} && (175)\\
\asymp& \frac{1}{n} + \frac{k'}{cn^p} && (176)
\end{aligned}$$

the last asymptotic relationship is due to that $I_{t,Y} \succ I_{t,Y,U}$ and $I_{t,X_Y} \succ I_{t,X_Y,U}$ as mentioned in the conditional shift case. For the case $0 < p < 1$, similarly we arrive at,

$$\begin{aligned}
\mathcal{R}(b) \asymp& \frac{1}{n+1} + \frac{\operatorname{Tr}((I_{t,X_Y} - I_{t,X_Y,U} + cn^p \Delta_s)\Delta_U^{-1}))}{n+1} && (177)\\
\asymp& \frac{1}{n} + \frac{k'}{n}. && (178)\\
& && (179)
\end{aligned}$$

which completes the proof. ■

In the following, we consider the semi-supervised learning scenario as $P_S(Y) = P_S(Y)$ and $P_S(X|Y) = P_T(X|Y)$.

Proof Since the source and the target have the same distribution, we choose the prior distribution as,

$$\omega(\Theta_s, \Theta_t) = \omega(\Theta_Y^{st}) \omega(\Theta_{X_Y}^{st}). \tag{180}$$

Combining the proofs of labelling drift and target shift cases, we arrive at,

$$\mathbf{I}_\theta = \begin{bmatrix} nI_{t,X_Y,U} + mI_{s,X_Y} + I_{t,X_Y} & nI_{t,U}^T \\ nI_{t,U} & nI_{t,Y,U} + I_{t,Y} + mI_{s,Y} \end{bmatrix} = nI_t + mI_s + I_{t,X,Y} \tag{181}$$

and

$$\widetilde{\mathbf{I}}_\theta = \begin{bmatrix} (n+1)I_{t,X_Y,U} + mI_{s,X_Y} & (n+1)I_{t,U}^T \\ (n+1)I_{t,U} & (n+1)I_{t,Y,U} + mI_{s,Y} \end{bmatrix} = (n+1)I_t + mI_s. \tag{182}$$

We first consider $m = cn^p$ for some $p \geq 1$, as n goes to infinity, we define the matrices $\Delta_U = I_{t,X_Y,U}^T - I_{t,U}I_{t,Y,U}^{-1}I_{t,U}$ and $\Delta_s = I_{s,X_Y}$, then the regret can be calculated by,

$$\begin{aligned}
\mathcal{R}(b) =& \frac{1}{2}\frac{\log\det(\mathbf{I}_\theta)}{\log\det(\widetilde{\mathbf{I}}_\theta)} + o(\frac{1}{m\vee n}) && (183)\\
=& \frac{1}{2}\log\det\left(\mathbf{I}_{k'+1} + (I_{t,X,Y} - I_t)((n+1)I_t + cn^pI_s)^{-1}\right) + o(\frac{1}{m\vee n}) && (184)\\
\asymp& \frac{\mathrm{Tr}((I_{t,X,Y} - I_t)(\frac{n+1}{cn^p}I_t + I_s)^{-1})}{cn^p} && (185)\\
\asymp& \frac{k'+1}{n^p} && (186)
\end{aligned}$$

due to that $I_{t,X,Y} \succ I_t$. Similarly for the case where $0 < p < 1$, we have,

$$\begin{aligned}
\mathcal{R}(b) &\asymp \frac{\mathrm{Tr}((I_{t,X,Y} - I_t)(\frac{cn^p}{n+1}I_s + I_t)^{-1})}{n+1} && (187)\\
&\asymp \frac{k'+1}{n}. && (188)
\end{aligned}$$

As a consequence,

$$\mathcal{R}(b) \asymp \frac{k'+1}{n\vee n^p}. \tag{189}$$

■

A.6 Proof of Lemma 13

Proof We write the minimax expected regret as,

$$\begin{aligned}
&\min_b \max_{\theta_s^*,\theta_t^*} R(b) = \min_Q \left\{\max_{\theta_s,\theta_t}\left\{D\left(P_{\theta_s,\theta_t}\|Q(\theta_s,\theta_t)\right)\right\}\right\}\\
&= \min_b\left\{\max_{\theta_s,\theta_t}\left\{\int P_{\theta_s,\theta_t}\left(D_t^{U,m}, D_s^m, X_t', Y_t'\right)\log\left(\frac{P_{\theta_t}\left(Y_t'|X_t'\right)}{Q\left(Y_t'|D_t^{U,m}, D_s^m, X_t'\right)}\right)dD_t^{U,m}dD_s^m dX_t' dY_t'\right\}\right\}\\
&\overset{(a)}{=} \min_b\left\{\max_{\omega(\theta_s,\theta_t)}\left\{\int P_{\theta_s,\theta_t}\left(D_t^{U,m}, D_s^m, X_t', Y_t'\right)\log\left(\frac{P_{\theta_t}\left(Y_t'|X_t'\right)}{Q\left(Y_t'|D_t^{U,m}, D_s^m, X_t'\right)}\right)\omega(\theta_s,\theta_t)d\theta_s d\theta_t dD_t^{U,m}dD_s^m dX_t' dY_t'\right\}\right\}\\
&\overset{(b)}{=} \max_{\omega(\theta_s,\theta_t)}\left\{\min_b\left\{\int D\left(P_{\theta_t}\left(Y_t'|X_t'\right)\|Q\left(Y_t'|D_t^{U,m}, D_s^m, X_t'\right)\right)\omega(\theta_s,\theta_t)d\theta_s d\theta_t dD_t^{U,m}dD_s^m dX_t' dY_t'\right\}\right\}\\
&= \max_{\omega(\theta_s,\theta_t)} I(Y_t';\theta_s,\theta_t|D_s^m, D_t^{U,n}, X_t'),
\end{aligned}$$

where (a) follows as maximizing over θ_s and θ_t and is equivalent to maximizing over a distribution over them and (b) follows from the minimax theorem, e.g., see Du and Pardalos (2013) for proof. ■

References

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

Ole Barndorff-Nielsen. Identifiability of mixtures of exponential families. *Journal of Mathematical Analysis and Applications*, 12(1):115–121, 1965.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.

Irineo Cabreros and John Storey. Causal models on probability spaces. *arXiv preprint arXiv:1907.01672*, 2019.

T. Tony Cai and Hongji Wei. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1):100–128, 2021.

Bradley P. Carlin and Thomas A. Louis. *Bayesian methods for data analysis*. CRC press, 2008.

Vittorio Castelli and Thomas M. Cover. The relative value of labelled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on Information Theory*, 42(6):2102–2117, 1996.

Minghui Chen and Joseph G. Ibrahim. Conjugate priors for generalized linear models. *Statistica Sinica*, pages 461–476, 2003.

Yuansi Chen and Peter Bühlmann. Domain adaptation under structural causal models. *Journal of Machine Learning Research*, 22:1–80, 2021.

Bertrand S. Clarke. Asymptotic normality of the posterior in relative entropy. *IEEE Transactions on Information Theory*, 45(1):165–176, 1999.

Bertrand S. Clarke and Andrew R. Barron. Information-theoretic asymptotics of bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471, 1990.

Bertrand S. Clarke and Andrew R. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41(1):37–60, 1994.

Spencer Compton, Murat Kocaoglu, Kristjan Greenewald, and Dmitriy Katz. Entropic causal inference: Identifiability and finite sample results. *Advances in Neural Information Processing Systems*, 33:14772–14782, 2020.

Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *International Conference on Algorithmic Learning Theory*, pages 38–53. Springer, 2008.

Thomas M. Cover and Erik Ordentlich. Universal portfolios with side information. *IEEE Transactions on Information Theory*, 42(2):348–363, 1996.

Thomas M. Cover and Joy A. Thomas. Elements of information theory, 2006.

Peng Cui and Susan Athey. Stable learning establishes some common ground between causal inference and machine learning. *Nature Machine Intelligence*, 4:110–115, 2022.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

Persi Diaconis and Donald Ylvisaker. Conjugate priors for exponential families. *The Annals of Statistics*, pages 269–281, 1979.

Dingzhu Du and Panos M. Pardalos. *Minimax and applications*, volume 4. Springer Science & Business Media, 2013.

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

Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

Meir Feder, Neri Merhav, and Michael Gutman. Universal prediction of individual sequences. *IEEE Transactions on Information Theory*, 38(4):1258–1270, 1992.

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

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset Shift in Machine Learning*, 3(4):5, 2009.

Bettina Grün and Friedrich Leisch. Dealing with label switching in mixture models under genuine multimodality. *Journal of Multivariate Analysis*, 100(5):851–861, 2009.

Mats Gyllenberg, Timo Koski, Edwin Reilink, and Martin Verlaan. Non-uniqueness in probabilistic numerical identification of bacteria. *Journal of Applied Probability*, 31(2): 542–548, 1994.

David Haussler and Manfred Opper. General bounds on the mutual information between a parameter and n conditionally independent observations. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, pages 402–411, 1995.

Miguel A. Hernán and James M. Robins. Causal inference, 2010.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994. doi: 10.1109/34.291440.

Guido W. Imbens and Donald B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Niki Kilbertus, Giambattista Parascandolo, and Bernhard Schölkopf. Generalization in anti-causal learning. *arXiv preprint arXiv:1812.00524*, 2018.

Murat Kocaoglu, Alexandros G. Dimakis, Sriram Vishwanath, and Babak Hassibi. Entropic causal inference. In *The Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Samory Kpotufe and Guillaume Martinet. Marginal singularity, and the benefits of labels in covariate-shift. In *Conference on Learning Theory*, pages 1882–1886. PMLR, 2018.

Kun Kuang, Peng Cui, Susan Athey, Ruoxuan Xiong, and Bo Li. Stable prediction across unknown environments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1617–1626, 2018.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yufeng Li and Zhihua Zhou. Towards making unlabeled data never hurt. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):175–188, 2014.

Feng Liang, Sayan Mukherjee, and Mike West. The use of unlabeled data in predictive modelling. *Statistical Science*, 22(2):189–205, 2007.

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. *Advances in neural information processing systems*, 31, 2018.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *Proceedings of the 38th International Conference on Machine Learning*, pages 7313–7324. PMLR, 2021.

Subha Maity, Yuekai Sun, and Moulinath Banerjee. Minimax optimal approaches to the label shift problem in non-parametric settings. *Journal of Machine Learning Research*, 23 (346):1–45, 2022.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*, 2020.

Jean Michel Marin, Kerrie Mengersen, and Christian P. Robert. Bayesian modelling and inference on mixtures of distributions. *Handbook of Statistics*, 25:459–507, 2005.

Geoffrey J. McLachlan, Sharon X. Lee, and Suren I. Rathnayake. Finite mixture models. *Annual Review of Statistics and Its Application*, 6:355–378, 2019.

Neri Merhav and Meir Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998.

Alexander Mey and Marco Loog. Improvability through semi-supervised learning: a survey of theoretical results. *arXiv preprint arXiv:1908.09574*, 2019.

Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(1):18–33, 2020.

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.

Judea Pearl. Graphs, causality, and structural equation models. *Sociological Methods & Research*, 27(2):226–284, 1998.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

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.

Bernhard Schölkopf. Causality for machine learning. In *Probabilistic and causal inference: The works of Judea Pearl*, pages 765–804. 2022.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

Matthias Seeger. Input-dependent regularization of conditional density models. Technical report, 2000.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *The Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

Henry Teicher. Identifiability of finite mixtures. *The Annals of Mathematical Statistics*, pages 1265–1269, 1963.

D. Michael Titterington, Smith Afm, Adrian F.M. Smith, Udi Makov, et al. *Statistical analysis of finite mixture distributions*, volume 198. John Wiley & Sons Incorporated, 1985.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7167–7176, 2017.

Xuetong Wu, Jonathan H Manton, Uwe Aickelin, and Jingge Zhu. Online transfer learning: Negative transfer and effect of prior knowledge. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 1540–1545. IEEE, 2021.

Qun Xie and Andrew R. Barron. Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Transactions on Information Theory*, 46(2):431–445, 2000.

Oren Yuval and Saharon Rosset. Semi-supervised empirical risk minimization: Using unlabeled data to improve prediction. *Electronic Journal of Statistics*, 16(1):1434–1460, 2022.

Yusen Zhan and Mattew E Taylor. Online transfer learning in reinforcement learning domains. In *2015 AAAI Fall Symposium Series*, 2015.

Jing Zhang, Wanqing Li, and Philip Ogunbona. Joint geometrical and statistical alignment for visual domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1859–1867, 2017.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827. PMLR, 2013.

Tong Zhang and Frank J. Oles. The value of unlabeled data for classification problems. In *Proceedings of the 17th International Conference on Machine Learning*, volume 20, page 0. Citeseer, 2000.

Jingge Zhu. Semi-supervised learning: the case when unlabeled data is equally useful. In *Conference on Uncertainty in Artificial Intelligence*, pages 709–718. PMLR, 2020.