Why Stable Learning Works? A Theory of Covariate Shift Generalization

Renzhe Xu[1], Peng Cui*[1], Zheyan Shen[1], Xingxuan Zhang[1], and Tong Zhang[2]

[1]Department of Computer Science, Tsinghua University, Beijing, China
[2]Computer Science & Mathematics, The Hong Kong University of Science and Technology, Hong Kong, China

xrz199721@gmail.com, cuip@tsinghua.edu.cn, shenzy17@mails.tsinghua.edu.cn, xingxuanzhang@hotmail.com, tongzhang@tongzhang-ml.org

Abstract

Covariate shift generalization, a typical case in out-of-distribution (OOD) generalization, requires a good performance on the unknown testing distribution, which varies from the accessible training distribution in the form of covariate shift. Recently, stable learning algorithms have shown empirical effectiveness to deal with covariate shift generalization on several learning models involving regression algorithms and deep neural networks. However, the theoretical explanations for such effectiveness are still missing. In this paper, we take a step further towards the theoretical analysis of stable learning algorithms by explaining them as feature selection processes. We first specify a set of variables, named **minimal stable variable set**, that is minimal and optimal to deal with covariate shift generalization for common loss functions, including the mean squared loss and binary cross entropy loss. Then we prove that under ideal conditions, stable learning algorithms could identify the variables in this set. Further analysis on asymptotic properties and error propagation are also provided. These theories shed light on why stable learning works for covariate shift generalization.

1 Introduction

Although modern machine learning techniques have achieved great success in various areas, many researchers have revealed the vulnerability of machine learning models when exposed to data with distribution shifts. This issue arises from the violation of a fundamental assumption that training and testing data are independent and identically distributed (*i.i.d.* assumption) and stimulates recent research on out-of-distribution (OOD) generalization [Shen et al., 2021]. Among different types of distribution shifts considered in OOD literature, covariate shift, where the marginal distribution of variables shifts from the training phase to the testing phase while the label generation mechanism keeps unchanged, is the most common one [Shimodaira, 2000; Sugiyama et al., 2007; Ben-David et al., 2007]. With prior knowledge of testing distribution, traditional methods showed effectiveness in handling the covariate shift problem via importance sampling

*Corresponding author

[Huang et al., 2006; Storkey and Sugiyama, 2007; Bickel et al., 2007] or feature space transformation [Fernando et al., 2013; Long et al., 2017]. Nevertheless, covariate shift generalization is much more challenging, given that testing distribution remains unknown in the training phase.

Recently, stable learning algorithms [Shen et al., 2020b; Kuang et al., 2020b; Zhang et al., 2021] have shown empirical effectiveness to deal with covariate shift generalization on several learning tasks, involving regression algorithms and deep models. The framework of these stable learning algorithms usually consists of two steps, namely importance sampling [Koller and Friedman, 2009] and weighted regression. In the importance sampling step, the algorithms learn sample weights to ensure statistical independence between features in weighted distribution. Then typical regression models are adopted in the weighted regression step. Although the advantages of stable learning algorithms have been proved empirically, the theoretical explanations for these methods are missing. In this paper, we take a step towards the theoretical analysis of stable learning algorithms by explaining them as feature selection processes.

We first show that for common loss functions, including the mean squared loss and binary cross entropy loss, the covariate shift generalization problem can be tackled by a minimal set of variables $\mathbf{S}$ that satisfies the condition: $\mathbb{E}[Y|\mathbf{S}] = \mathbb{E}[Y|\mathbf{X}]$. Such minimal set of variables is named the **minimal stable variable set**. Afterward, we prove that stable learning algorithms could identify the minimal stable variable set. We analyze the typical algorithms [Kuang et al., 2020b; Shen et al., 2020b] where the weighted least squares (WLS) is adopted in the weighted regression step. Variables whose corresponding coefficients of WLS are not zero could be considered as chosen variables. Under ideal conditions, *i.e.*, perfectly learned sample weights and infinite samples, the selected variables are proved to be the minimal stable variable set. We further provide asymptotic properties and error analysis when the ideal conditions are not satisfied. We highlight that although a linear model (WLS) is adopted, these theoretical results hold for both linear and non-linear data-generating processes. Along with the optimality and minimality of the minimal stable variable set, these theories provide a way to explain why stable learning works for covariate shift generalization.

1.1 Overview of results

We begin with a simplified presentation of our results. Consider a set of variables $(\mathbf{X}, Y)$ where $\mathbf{X}$ are features and Y is the outcome that we try to predict from $\mathbf{X}$. We consider OOD problems with covariate shift, which is the most common one among the different distribution shifts [Shen et al., 2021]. Covariate shift considers the scenario where the marginal distribution of $\mathbf{X}$ shifts from training phase to testing phase while the label generation mechanism keeps unchanged.

Assumption 1. Suppose the testing distribution P^{te} differs from the training distribution P^{tr} in covariate shift only, *i.e.*,

$$P^{\text{te}}(\mathbf{X}, Y) = P^{\text{te}}(\mathbf{X})P^{\text{tr}}(Y|\mathbf{X}). \tag{1}$$

In addition, P^{te} has the same support of P^{tr}.

Problem 1 (Covariate shift generalization problem)**.** Given the samples from the training distribution P^{tr}, covariate shift generalization problem is to design an algorithm which can guarantee the performance on the unknown testing distribution P^{te} that satisfies Assumption 1.

We focus on several common loss functions, including the mean squared loss and binary cross entropy loss, where $\mathbb{E}_{P^{\text{te}}}[Y|\mathbf{X}]$ is the global optimum for the testing distribution P^{te}.

Theorem 1 (Informal version of Theorem 3)**.** *Let P^{te} be the unknown testing distribution in the covariate shift generalization problem defined in Problem 1. Then a subset of variables $\mathbf{S} \subseteq \mathbf{X}$ that can approximate the target $\mathbb{E}_{P^{te}}[Y|\mathbf{X}]$ if and only if it satisfies $\mathbb{E}_{P^{tr}}[Y|\mathbf{S}] = \mathbb{E}_{P^{tr}}[Y|\mathbf{X}]$.*

We define the minimal set of variables that satisfies $\mathbb{E}_{p^{\text{tr}}}[Y|\mathbf{S}] = \mathbb{E}_{p^{\text{tr}}}[Y|\mathbf{X}]$ as **the minimal stable variable set** (Definition 4). Under mild assumptions (Assumption 2), the existence and uniqueness of such variables are guaranteed (Theorem 4). As relationships between $\mathbf{X}$ are unstable, *i.e.*, $P^{\text{tr}}(\mathbf{X}) \neq P^{\text{te}}(\mathbf{X})$, it is reasonable to find the minimal set of variables to make predictions so that it can relieve the negative impact from other features in the testing distribution.

Now we consider stable learning algorithms. Typical stable learning algorithms [Shen et al., 2020b; Kuang et al., 2018] learn sample weights first and then adopt a weighted least squares regression step. The algorithms can be considered as processes of feature selection by examining the coefficients of WLS. In detail, the variables with non-zero coefficients are chosen. The variables chosen by stable learning algorithms have the following properties.

Theorem 2 (Informal version of Theorem 5 and Theorem 6)**.** *Under ideal conditions (perfectly learned sample weights and infinite samples),*
– if X_i is not in the minimal stable variable set, stable learning algorithms could filter it out, and
– if X_i is in the minimal stable variable set, there exists sample weighting functions with which stable learning algorithms could identify X_i.

We further analyze the error of coefficients if these ideal conditions are not satisfied (Corollary 8, Corollary 9, and Theorem 10).

Theorem 1 and Theorem 2 provide a general picture of the effectiveness of stable learning algorithms. To conclude, under ideal assumptions, stable learning algorithms could identify the minimal stable variable set, which is the minimal and optimal set of variables to deal with covariate shift generalization.

1.2 Related works

OOD and covariate shift generalization OOD generalization has raised great concerns. According to [Shen et al., 2021], OOD methods could be categorized into unsupervised representation learning methods [Bengio et al., 2013; Yang et al., 2021], supervised learning models [Peters et al., 2016; Zhou et al., 2021], and optimization methods [Duchi and Namkoong, 2021; Liu et al., 2021]. More thorough discussions could be found in [Shen et al., 2021].

There are many types of distribution shift, including covariate shift [Shimodaira, 2000], label shift [Garg et al., 2020], and concept shift [Gama et al., 2014]. Covariate shift is the most common distribution shift and stable learning algorithms mainly deal with it. Much of the work consider the domain adaptation (DA) setting where methods [Huang et al., 2006; Storkey and Sugiyama, 2007; Bickel et al., 2007; Gretton et al., 2009; Zhao et al., 2019] make importance sampling with the knowledge of the unlabeled testing distribution. To deal with unknown testing distribution under covariate shift, there are several methods recently including stable learning algorithms [Shen et al., 2020b; Kuang et al., 2020b; Zhang et al., 2021] and DRO [Duchi and Namkoong, 2021].

Stable learning Stable learning algorithms can be considered as a feature selection mechanism according to the regression coefficients [Shen et al., 2021]. Motivated by the literature of variable balancing methods [Hainmueller, 2012; Zubizarreta, 2015; Athey et al., 2016], Shen et al. [2018] proposed to consider all the variables as the treatment and learn a set of weights for all of available samples to remove the confounding bias from data distribution. Specifically, a global balancing loss is proposed as a regularizer which can be easily plugged into machine learning models. Kuang et al. [2018] managed to combine global balancing and unsupervised feature representation learning with auto-encoders [Bengio et al., 2007].

The following work [Shen et al., 2020b] proposed to address model misspecification scenarios with linear models via a sample reweighting strategy. Shen et al. [2020a] further proposed to recover the latent cluster structures among variables using unlabeled data and proved decorrelating the variables between clusters instead of each other sufficient to achieve a stable estimation while preventing the variance inflation. Recently, Zhang et al. [2021] proposed a framework named StableNet, which extends former linear stable learning frameworks [Shen et al., 2018; Kuang et al., 2018; Shen et al., 2020b] to incorporate deep models. StableNet adopted Random Fourier Features (RFF) [Rahimi et al., 2007] to eliminate non-linear dependences between features sufficiently. Moreover, Kuang et al. [2020b] attempted to reduce the effects of confounding bias by sampling the data and several extensions of stable learning algorithms are proposed for causal feature selection or out-of-distribution generalization [Zhang et al., 2020; Wang et al., 2020; Yuan et al., 2021].

Feature Selection Feature selection aims to construct a diagnostic or predictive model for a given regression or classification task via selecting a minimal-size subset of variables that show the best performance [Guyon and Elisseeff, 2003]. It is of great importance for learning trustworthy models especially when there are distribution shifts between training and testing data and, thus some variables can be uninformative or even misleading [Shen et al., 2020b; Kuang et al., 2020a].

Feature selection approaches can be broadly divided into four categories, namely filter methods, wrapper methods, embedded methods and others [Guyon and Elisseeff, 2003; Bolón-Canedo et al., 2013; Urbanowicz et al., 2018]. Filter methods adopt statistical criteria to rank and select features before building classifier with selected features [John et al., 1994; Langley et al., 1994; Guyon and Elisseeff, 2003; Law et al., 2004]. Given filter methods are usually independent from the learning of classifier, they show superiority in operating time and applicability over other methods [Kira and Rendell, 1992; Bolón-Canedo et al., 2013]. Wrapper methods heuristically search variable subsets via learning a predictive model, thus they can identify the best performing feature subsets for given modeling algorithm, but are typically computationally intensive [Menze et al., 2009; Bolón-Canedo et al., 2013; Urbanowicz et al., 2018]. Embedded methods seek to minimizing the size of selected feature subset while maximizing the classification performance simultaneously [Tibshirani, 1996; Rakotomamonjy, 2003; Zou and Hastie, 2005; Loh, 2011; Chen and Guestrin, 2016]. There are also some methods attempt to combine the advantages of wrapper methods and filter methods [Cortizo and Giraldez, 2006; Liu et al., 2014; Benoît et al., 2013].

Causal discovery and Markov boundary Causal literature can be categorized into two frameworks, namely the potential outcome [Rosenbaum and Rubin, 1983; Holland, 1986; Rubin, 2005; Imbens and Rubin, 2015] and structural causal model framework [Pearl, 2014]. The definition of the minimal stable variable set in this work is closely related to the Markov boundary, which falls into the structural causal model framework. Traditional causal discovery literature aims to discover the causal relationship between all variables. Typical methods include constraint-based methods [Spirtes et al., 2000, 2013], scored-based [Chickering, 2002; Huang et al., 2018], and learning-based [Zheng et al., 2018, 2020; He et al., 2021].

Markov blankets and Markov boundary [Pearl, 2014] are the cores of local causal discovery. Under the intersection assumption [Pearl, 2014], Markov boundary is proved unique and the discovery algorithms include [Tsamardinos and Aliferis, 2003; Tsamardinos et al., 2003a,b; Mani and Cooper, 2004; Aliferis et al., 2010a,b; Pena et al., 2007]. Moreover, Liu et al. [2010a,b]; Statnikov et al. [2013] studied the setting when multiple Markov boundaries exist. In this paper, we assume that the probability are strictly positive, which is a stronger assumption than intersection assumption [Pearl, 2014] but is also common in reality [Strobl and Visweswaran, 2016]. With this

assumption, we can guarantee the uniqueness of the Markov boundary and the minimal stable variable set proposed in this paper.

2 Preliminaries

2.1 Notations

Let $\mathbf{X} = (X_1, X_2, \dots X_d)^T \in \mathbb{R}^d$ denote the d-dimensional features and $Y \in \mathbb{R}$ denote the outcome. The data is from a joint training distribution $P^{\text{tr}}(\mathbf{X}, Y)$. Let $\mathcal{X}$, $\mathcal{X}_j$, and $\mathcal{Y}$ denote the support of $\mathbf{X}$, X_j, and Y, respectively. Suppose we get n *i.i.d.* samples, $\left\{\mathbf{X}^{(i)} = \left(x_1^{(i)}, \dots, x_d^{(i)}\right)^T, y^{(i)}\right\}_{i=1}^{n}$ sampled from the distribution. Let P^{te} denote the testing distribution.

We use $\mathbf{S} \subseteq \mathbf{X}$ to indicate that $\mathbf{S}$ is a subset of features $\mathbf{X}$ and $\subsetneq$ to mean proper subset. We write $\mathbf{A} \perp \mathbf{B} | \mathbf{C}$ when two sets of variables $\mathbf{A}, \mathbf{B} \subseteq \mathbf{X}$ are statistically independent given another set of variables $\mathbf{C} \subseteq \mathbf{X}$. We also adopt $\mathbf{A} \perp \mathbf{B}$ when conditioning set is empty to indicate that $\mathbf{A}$ and $\mathbf{B}$ are statistically independent.

We use $\mathbb{E}_{Q(\cdot)}[\cdot]$ and $\mathbb{E}_{Q(\cdot)}[\cdot|\cdot]$ to denote expectation and conditional expectation, respectively, under a distribution Q. For example, $\mathbb{E}_{Q(\mathbf{X})}[\mathbf{X}] = \int_{\mathcal{X}} \mathbf{x} Q(\mathbf{X} = \mathbf{x}) \mathrm{d}x$ represent the expectation of $\mathbf{X}$ and $\mathbb{E}_{Q(\mathbf{X},Y)}[Y|\mathbf{X}] = \int_{\mathcal{Y}} Q(Y = y|\mathbf{X}) y \mathrm{d}y$ represent the conditional expectation of Y given $\mathbf{X}$ under distribution Q. Q could be chosen as the training distribution P^{tr}, testing distribution P^{te}, or any other proper distributions. If not confusing, we will use $\mathbb{E}[\cdot]$ and $\mathbb{E}[\cdot|\cdot]$ to denote the expectation and conditional expectation under the training distribution P^{tr}.

2.2 Assumptions

Assumption 2 (Strictly positive density assumption)**.** $\forall x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2, \dots, x_d \in \mathcal{X}_d$, $P^{\text{tr}}(X_1 = x_1, X_2 = x_2, \dots, X_d = x_d) > 0$.

Remark. Assumption 2 is reasonable on the grounds that there always exists uncertainty in the data [Pearl, 2014; Strobl and Visweswaran, 2016]. Therefore, we suppose strictly positive density assumption in the whole paper for simplicity.

3 Minimal and optimal predictor for covariate shift generalization

In this section, we specify the set of variables that are suitable for covariate shift generalization problems. We first provide the definition of the minimal and optimal predictor.

Definition 1 (Optimal predictor [Statnikov et al., 2013])**.** Given a dataset sampled from $P^{\text{tr}}(\mathbf{X}, Y)$, a learning algorithm $\mathbb{L}$, and a performance metric $\mathbb{M}$ to assess learner's models, a variable set $\mathbf{S} \subseteq \mathbf{X}$ is an optimal predictor of Y if $\mathbf{S}$ maximizes the performance metric $\mathbb{M}$ for predicting Y using learner $\mathbb{L}$ in the dataset.

Definition 2 (Minimal and optimal predictor [Strobl and Visweswaran, 2016])**.** Let $\mathbf{S}$ be an optimal predictor of Y. If no proper subset of $\mathbf{S}$ satisfies the definition of an optimal predictor of Y, then $\mathbf{S}$ is a minimal and optimal predictor of Y.

The minimal and optimal predictor for covariate shift generalization can be given as follows.

Theorem 3. *Under Assumption 1 and Assumption 2, if* $\mathbb{M}$ *is a performance metric that is maximized only when* $\mathbb{E}_{P^{te}}[Y|\mathbf{X}]$ *is estimated accurately and* $\mathbb{L}$ *is a learning algorithm that can approximate any conditional expectation. Suppose* $\mathbf{S} \subseteq \mathbf{X}$ *is a subset of variables, then*

1. $\mathbf{S}$ *is an optimal predictor of* Y *under distribution* P^{te} *if and only if* $\mathbb{E}_{P^{tr}}[Y|\mathbf{X}] = \mathbb{E}_{P^{tr}}[Y|\mathbf{S}]$*, and*
2. $\mathbf{S}$ *is a minimal and optimal predictor of* Y *under distribution* P^{te} *if and only if* $\mathbb{E}_{P^{tr}}[Y|\mathbf{X}] = \mathbb{E}_{P^{tr}}[Y|\mathbf{S}]$ *and no proper subset* $\mathbf{S}' \subsetneq \mathbf{S}$ *satisfies* $\mathbb{E}_{P^{tr}}[Y|\mathbf{X}] = \mathbb{E}_{P^{tr}}[Y|\mathbf{S}']$.

Remark. To deal with covariate shift generalization, $\mathbb{M}$ should be measured on the unknown testing distribution P^{te} with common loss functions. In practice, researchers often adopt the mean squared loss in regression problems and the cross entropy loss in binary classification problems. It is easy to check that the global optimum for both loss functions are $\mathbb{E}_{P^{\text{te}}}[Y|\mathbf{X}]$ if applying the loss functions on the testing distribution P^{te}.

As a result, we provide the following definitions.

Definition 3 (Stable variable set)**.** A stable variable set of Y under distribution P is any subset $\mathbf{S}$ of $\mathbf{X}$ for which

$$\mathbb{E}_P[Y|\mathbf{S}] = \mathbb{E}_P[Y|\mathbf{X}]. \tag{2}$$

The set of all stable variable sets for Y is denoted as $\text{Stable}_P(Y)$. In addition, we use $\text{Stable}(Y)$ to denote the set under the training distribution P^{tr} for simplicity, *i.e.*, $\text{Stable}(Y) \triangleq \text{Stable}_{P^{\text{tr}}}(Y)$.

Definition 4 (Minimal stable variable set)**.** A minimal stable variable set of Y is a minimal set in $\text{Stable}(Y)$, *i.e.*, none of its proper subsets satisfy Equation 2.

With these definitions, the conclusions of Theorem 3 become: (1) $\mathbf{S}$ is an optimal predictor of Y under P^{te} if and only if it is a stable variable set under P^{tr}, and (2) $\mathbf{S}$ is a minimal and optimal predictor of Y under P^{te} if and only if it is a minimal stable variable set under P^{tr}. Furthermore, the existence and uniqueness of the minimal stable variable set are given by the following theorem.

Theorem 4. *Under Assumption 2, there exists a unique minimal stable variable set of* Y*, which can be denoted as* $\text{MinStable}(Y)$*. Furthermore, with the unique* $\text{MinStable}(Y)$*, the set of all stable variable sets of* Y *under the training distribution* P^{tr}*, i.e.,* $\text{Stable}(Y)$*, can be expressed as*

$$\text{Stable}(Y) = \{\mathbf{S} \subseteq \mathbf{X} \mid \text{MinStable}(Y) \subseteq \mathbf{S}\}. \tag{3}$$

Theorem 3 and Theorem 4 provides a way to ensure promising OOD performance for covariate shift generalization problems. The minimal stable variable set under the training distribution P^{tr} is a minimal and optimal predictor in the testing distribution P^{te}, with which we can learn reliable models [John et al., 1994; Guyon and Elisseeff, 2003]. As relationships between $\mathbf{X}$ are usually unstable and $P^{\text{tr}}(\mathbf{X}) \neq P^{\text{te}}(\mathbf{X})$, it is reasonable to find the minimal and optimal predictor, *i.e.*, $\text{MinStable}(Y)$, to make predictions so that it can relieve the negative impact from $\mathbf{X} \setminus \text{MinStable}(Y)$ under the testing distribution.

$\text{MinStable}(Y)$ could be explained as the direct causal variables in typical data-generating processes. Consider the following mechanism [Tibshirani, 1996; Ravikumar et al., 2009; Hastie and Tibshirani, 2017; Kuang et al., 2020a],

$$\mathbf{X} = (\mathbf{S}, \mathbf{V}), \quad Y = f(\mathbf{S}) + \epsilon, \quad \epsilon \perp \mathbf{X}. \tag{4}$$

The relationship between $\mathbf{S}$ and $\mathbf{V}$ is arbitrary. In such common cases, $\mathbf{S}$ includes all the direct causal variables and is the minimal stable variable set of Y.

The minimal stable variable set is also closely related to the Markov boundary [Pearl, 2014]. From a causal perspective, under the performance metric in Theorem 3, the minimal stable variable set shares the same prediction power of Y with the Markov boundary while the minimal stable variable set contains fewer variables and thus combats covariate shift generalization problems better. A detailed comparison between the minimal stable variable set and Markov boundary can be found in Section 6.

4 Stable learning algorithms

4.1 General framework

Algorithm 1 *Stable Learning Algorithm*

Input: Dataset $D = \left\{ \mathbf{x}^{(i)} = \left(x_1^{(i)}, \ldots, x_d^{(i)} \right)^T, y^{(i)} \right\}_{i=1}^{n}$

Output: Coefficients $\hat{\boldsymbol{\beta}}$ on each variables

1: /* Step I */
2: Learn weight $w(\mathbf{X})$ to make $\mathbf{X}$ are mutually independent of each other.
3: /* Step II */
4: Solve weighted least squares with weighting function $w(\mathbf{X})$. The solution is $\hat{\boldsymbol{\beta}}_w^{(n)}$.
5: Return $\hat{\boldsymbol{\beta}}_w^{(n)}$.

The framework of typical stable learning algorithms [Shen et al., 2020b; Kuang et al., 2020a] is shown in Algorithm 1. The algorithms usually consist of two steps, which are importance sampling and weighted least squares respectively.

A) Importance sampling

Importance sampling [Koller and Friedman, 2009, Section 12.2.2] is a general approach for estimating the expectation of a function $f(\mathbf{X}, Y)$ relative to some distribution $\tilde{P}(\mathbf{X}, Y)$, typically called the target distribution. If samples are generated from P^{tr} instead of $\tilde{P}$, one needs to adjust the estimator to compensate for the incorrect sampling distribution. We have $\mathbb{E}_{\tilde{P}(\mathbf{X},Y)}[f(\mathbf{X}, Y)] = \mathbb{E}_{P^{\text{tr}}(\mathbf{X},Y)}\left[\frac{\tilde{P}(\mathbf{X},Y)}{P^{\text{tr}}(\mathbf{X},Y)} f(\mathbf{X}, Y) \right]$. Stable learning algorithms consider a weighting function that depends on $\mathbf{X}$ only.

Definition 5 (Weighting function)**.** Let $\mathcal{W}$ be the set of weighting functions that satisfies

$$\mathcal{W} = \left\{ w : \mathcal{X} \to \mathbb{R}^+ \mid \mathbb{E}_{P^{\text{tr}}(\mathbf{X})}[w(\mathbf{X})] = 1 \right\}. \tag{5}$$

Then $\forall w \in \mathcal{W}$, the corresponding weighted distribution is $\tilde{P}_w(\mathbf{X}, Y) = w(\mathbf{X}) P^{\text{tr}}(\mathbf{X}, Y)$. $\tilde{P}_w$ is well defined with the same support of P^{tr}.

Instead of the whole set $\mathcal{W}$, stable learning algorithms consider a subset $\mathcal{W}_\perp \subseteq \mathcal{W}$. The weighting functions in $\mathcal{W}_\perp$ satisfies that $\mathbf{X}$ are mutually independent of each other in the corresponding distribution, *i.e.*,

$$\mathcal{W}_\perp = \left\{ w \in \mathcal{W} \mid \mathbf{X} \text{ are mutually independent in distribution } \tilde{P}_w \right\}. \tag{6}$$

B) Weighted least squares

Let $w \in \mathcal{W}$ be a weighing function. With n samples sampled from $P^{\text{tr}}(\mathbf{X}, Y)$, the weighted least squares solves the following equation

$$\hat{\beta}_w^{(n)} = \arg\min_{\beta} \sum_{i=1}^{n} w\left(\mathbf{x}^{(i)}\right)\left(\beta_{1\ldots d}^T \mathbf{x}^{(i)} + \beta_0 - y^{(i)}\right)^2. \tag{7}$$

And we denote the solution to population-level weighted least squares under distribution $P^{\text{tr}}(\mathbf{X}, Y)$ as

$$\beta_w = \arg\min_{\beta} \mathbb{E}_{P^{\text{tr}}(\mathbf{X},Y)}\left[w(\mathbf{X})\left(\beta_{1\ldots d}^T \mathbf{X} + \beta_0 - Y\right)^2\right]. \tag{8}$$

We use $\beta_w(X_i)$ and $\hat{\beta}_w^{(n)}(X_i)$ to denote the corresponding coefficient of X_i.

4.2 Two specific stable learning algorithms

Algorithm 1 has two typical implementations, namely DWR [Kuang et al., 2020a] and SRDO [Shen et al., 2020b]. They differ mainly in the way to learn sample weights.

DWR Kuang et al. [2020a] proposed to decorrelate the every two features, *i.e.*,

$$w(\mathbf{X}) = \arg\min_{w_0(\mathbf{X})} \sum_{1 \le i,j \le d, i \ne j} \left(\text{Cov}(X_i, X_j; w_0)\right)^2, \tag{9}$$

where $\text{Cov}(X_i, X_j; w_0)$ represents the covariance of feature X_i and X_j under weighted distribution $\tilde{P}_{w_0}$. The loss function in Equation 9 focuses on the linear correlation only and is used as an approximation for statistical independence. They proved that linear decorrelation suffices to generate good prediction models under simple models. Recently, Zhang et al. [2021] combined DWR with random fourier features [Rahimi et al., 2007] to achieve the statistical independence and showed that deep models could perform better if the representations are statistically independent instead of linearly decorrelated.

SRDO Shen et al. [2020b] proposed to learn $w(\mathbf{X})$ by estimating the density ratio of the training distribution P^{tr} and a specific target distribution $\tilde{P}$. The target distribution $\tilde{P}$ is determined by performing random resampling on each feature so that $\tilde{P}(X_1, X_2, \ldots, X_d) = \prod_{i=1}^{d} P^{\text{tr}}(X_i)$. And the weighting function $w(\mathbf{X})$ is given by

$$w(\mathbf{X}) = \frac{P(Z = 1|\mathbf{X})}{1 - P(Z = 1|\mathbf{X})}. \tag{10}$$

Here $P(Z = 1|\mathbf{X})$ means the probability of a sample $\mathbf{X}$, which is drawn from the balanced mixture of P^{tr} and $\tilde{P}$, belonging to P^{tr}. It can be learned by several different methods such as the cross entropy loss, the LSIF loss [Kanamori et al., 2009], and the KLIEP loss [Sugiyama et al., 2009]. A thorough review of density ratio estimation methods can be found in [Menon and Ong, 2016]. SRDO method can guarantee the statistical independence between variables $\mathbf{X}$ if the density ratio is estimated accurately.

5 Theoretical analysis of stable learning algorithms

In this section, we will show that stable learning algorithms as shown in Algorithm 1 can be considered as a process of feature selection according to the coefficients of weighted least squares. The chosen features are the minimal stable variable set in Definition 4. We first show the identifiability result with perfectly learned weighting functions and infinite samples in Section 5.1. These theoretical results, along with Theorem 3 could prove the effectiveness of stable learning algorithms for the covariate shift generalization problem (Problem 1). In Section 5.2, we relax the assumption to finite samples. Finally in Section 5.3, we consider the scenario when weighting functions could only be learned nearly perfectly.

5.1 Population level properties

Generally speaking, with infinite samples, for any perfectly learned proper weighting function $w \in \mathcal{W}_\perp$ adopted by the algorithms, the coefficient on variables that do not belong to the minimal stable variable set will be zero (Theorem 5). In addition, there exists proper weighting function such that the coefficients on the minimal stable variable set would not be zero (Theorem 6).

Theorem 5. *Under Assumption 2, suppose $X_i \notin \text{MinStable}(Y)$. Let w be any weighting function in $\mathcal{W}_\perp$. Suppose $\mathbb{E}_{P^{tr}(\mathbf{X})}\left[w(\mathbf{X})\|\mathbf{X}\|_2^2\right] < \infty$ and $\mathbb{E}_{P^{tr}(\mathbf{X},Y)}\left[w(\mathbf{X})Y^2\right] < \infty$. Then the population level solution $\boldsymbol{\beta}_w$ of weighted least squares under w satisfies*

$$\boldsymbol{\beta}_w(X_i) = 0. \tag{11}$$

Here $\boldsymbol{\beta}_w(X_i)$ means the corresponding coefficient on X_i.

Theorem 6. *Under Assumption 2, suppose $X_i \in \text{MinStable}(Y)$. Then there exists $w \in \mathcal{W}_\perp$ and constant $\alpha \neq 0$, such that the population-level solution $\boldsymbol{\beta}_w$ satisfies*

$$\boldsymbol{\beta}_w(X_i) = \alpha. \tag{12}$$

Here $\boldsymbol{\beta}_w(X_i)$ means the corresponding coefficient on X_i.

Remark. In very rare cases, stable learning algorithms may fail to identify the minimal stable variable set if X_i is not independent of Y but is linearly decorrelated with Y in the weighted distribution $\tilde{P}_w$.

Shen et al. [2020b] analyzed the theoretical properties of stable learning algorithms under a specific form of data-generating process, *i.e.*, linear model with a bounded bias term. However, besides Assumption 2, Theorem 5 and Theorem 6 do not need further assumptions on label generation mechanism $P^{\text{tr}}(Y|\mathbf{X})$, which implies that stable learning algorithms could be applied to both linear and non-linear data-generating processes.

These two theorems, along with Theorem 3, prove the effectiveness of stable learning algorithms for the covariate shift generalization problem (Problem 1). In detail, under ideal conditions, *i.e.*, perfectly learned sample weights and infinite samples, stable learning algorithms could find the minimal stable variable set of Y, which are minimal and optimal predictor under the testing distribution P^{te} according to Theorem 3.

5.2 Asymptotic properties of finite samples

With the asymptotic property of weighted least squares (Theorem 7), we can show the asymptotic properties (Corollary 8 and Corollary 9) of stable learning algorithms with perfectly learned weighting functions.

Theorem 7. *$\forall w \in \mathcal{W}$, suppose $\mathbb{E}_{P(\mathbf{X})}\left[w(\mathbf{X})\|\mathbf{X}\|_2^2\right] < \infty$, $\mathbb{E}_{P(\mathbf{X},Y)}\left[w(\mathbf{X})Y^2\right] < \infty$, and covariance matrix $\mathrm{Cov}_{\tilde{P}_w}[\mathbf{X}]$ is invertible. Let $\hat{\boldsymbol{\beta}}_w^{(n)}$ be the solution to weighted least squares under w with n samples, and $\boldsymbol{\beta}_w$ be the solution to population level weighted least squares. Then*

$$\hat{\boldsymbol{\beta}}_w^{(n)} \xrightarrow{p} \boldsymbol{\beta}_w, \quad \text{when } n \to \infty. \tag{13}$$

Here $\xrightarrow{p}$ means convergence in probability.

Combining Theorem 7 with Theorem 5, Theorem 6, we can get the following corollaries.

Corollary 8. *Under Assumption 2, suppose $X_i \notin \mathrm{MinStable}(Y)$. Let w be any weighting function in $\mathcal{W}_\perp$. Suppose $\mathbb{E}_{P^{tr}(\mathbf{X})}\left[w(\mathbf{X})\|\mathbf{X}\|_2^2\right] < \infty$ and $\mathbb{E}_{P^{tr}(\mathbf{X},Y)}\left[w(\mathbf{X})Y^2\right] < \infty$. Then*

$$\hat{\boldsymbol{\beta}}_w^{(n)}(X_i) \xrightarrow{p} 0, \quad \text{when } n \to \infty. \tag{14}$$

Corollary 9. *Under Assumption 2, suppose $X_i \in \mathrm{MinStable}(Y)$. Then there exists $w \in \mathcal{W}_\perp$ and constant $\alpha \neq 0$ such that*

$$\hat{\boldsymbol{\beta}}_w^{(n)}(X_i) \xrightarrow{p} \alpha, \quad \text{when } n \to \infty. \tag{15}$$

5.3 Error analysis with imperfectly learned weights

In this subsection, we further analyze the relative error of $\boldsymbol{\beta}$ if weighting function w is not perfectly learned.

Theorem 10. *Suppose $C = \max\left\{\sqrt{\mathbb{E}_{P^{tr}(\mathbf{X})}\left[\|\mathbf{X}\|_2^4\right]}, \sqrt{\mathbb{E}_{P^{tr}(\mathbf{X},Y)}\left[\|\mathbf{X}Y\|_2^2\right]}\right\} < \infty$. Let $w \in \mathcal{W}$ and $\hat{w}$ be an estimation of w. Suppose $\mathbb{E}_{P^{tr}(\mathbf{X})}\left[w(\mathbf{X})\mathbf{X}\mathbf{X}^T\right]$ and $\mathbb{E}_{P^{tr}(\mathbf{X})}\left[\hat{w}(\mathbf{X})\mathbf{X}\mathbf{X}^T\right]$ is invertible. Let $\mathbb{E}_{P^{tr}(\mathbf{X})}[(w(\mathbf{X}) - \hat{w}(\mathbf{X}))^2] = \epsilon^2$, $A = \mathbb{E}_{P^{tr}(\mathbf{X})}\left[w(\mathbf{X})\mathbf{X}\mathbf{X}^T\right]$, and $b = \mathbb{E}_{P^{tr}(\mathbf{X},Y)}[w(\mathbf{X})\mathbf{X}Y]$. Suppose $\epsilon C\|A^{-1}\|_2 < 1$, then*

$$\frac{\left\|\boldsymbol{\beta}_{\hat{w}} - \boldsymbol{\beta}_w\right\|_2}{\left\|\boldsymbol{\beta}_w\right\|_2} \leq \frac{\epsilon C\|A^{-1}\|_2}{1 - \epsilon C\|A^{-1}\|_2} \cdot \left(1 + \frac{\|A\|_2}{\|b\|_2}\right). \tag{16}$$

Here $\boldsymbol{\beta}_{\hat{w}}$ and $\boldsymbol{\beta}_w$ represents the population-level solution to weighted linear squares under w and $\hat{w}$ respectively.

Remark. Theorem 10 ensures that if the error ϵ^2, of weighting function is small enough, the relative error of estimated coefficients $\boldsymbol{\beta}$ will also be small. To ensure a small ϵ^2, we adopt LSIF [Kanamori et al., 2009] to optimize $\mathbb{E}_{P^{tr}(\mathbf{X})}[(w(\mathbf{X}) - \hat{w}(\mathbf{X}))^2]$ directly. If we know a target distribution Q and want to learn a weighting function $w(\mathbf{X}) = \frac{Q(\mathbf{X})}{P^{tr}(\mathbf{X})}$. According to Menon and Ong [2016], the loss of LSIF is

$$L(w) = \mathbb{E}_{Q(\mathbf{X})}[-w(\mathbf{X})] + \mathbb{E}_{P^{tr}(\mathbf{X})}\left[\frac{1}{2}w(\mathbf{X})^2\right]. \tag{17}$$

It is easy to see that $w^*(\mathbf{X}) = \min_w L(w) = \frac{Q(\mathbf{X})}{P^{tr}(\mathbf{X})}$ and

$$L(w) - L(w^*) = \frac{1}{2}\mathbb{E}_{P^{tr}(\mathbf{X})}\left[(w^*(\mathbf{X}) - w(\mathbf{X}))^2\right]. \tag{18}$$

As a result, minimizing the loss of LSIF will meet the assumption which requires that $\mathbb{E}_{P^{tr}(\mathbf{X})}[(w(\mathbf{X}) - \hat{w}(\mathbf{X}))^2] = \epsilon^2$ be small enough.

6 Relationships between minimal stable variable set and Markov boundary

The minimal stable variable set is closely related to the Markov boundary and stable learning may help identify the Markov boundary to some extent. In addition, if setting covariate shift generalization as the goal, the Markov boundary is not necessary while the minimal stable variable set is sufficient and optimal.

Definition and basic property of the Markov blankets and boundary According to [Statnikov et al., 2013; Pearl, 2014], Markov blankets and Markov boundary are defined as follows.

Definition 6 (Markov blanket)**.** A Markov blanket of Y under distribution P is any subset $\mathbf{S}$ of $\mathbf{X}$ for which

$$Y \perp (\mathbf{X} \backslash \mathbf{S}) \mid \mathbf{S}. \tag{19}$$

The set of all Markov blankets for Y is denoted as $\mathrm{BL}_P(Y)$. In addition, we use $\mathrm{BL}(Y)$ to denote the set under the training distribution P^{tr} for simplicity, *i.e.*, $\mathrm{BL}(Y) \triangleq \mathrm{BL}_{P^{\mathrm{tr}}}(Y)$.

Definition 7 (Markov boundary)**.** A Markov Boundary of Y is a minimal Markov blanket of Y, *i.e.*, none of its proper subsets satisfy Equation 19.

The existence of Markov blankets and Markov boundaries are given by the following proposition.

Proposition 11. *Under Assumption 2, there exists a unique Markov boundary of Y, which can be denoted as* $\mathrm{BD}(Y)$*. Furthermore, with the unique Markov boundary* $\mathrm{BD}(Y)$*, the set of all Markov blankets of Y,* $\mathrm{BL}(Y)$*, can be expressed as*

$$\mathrm{BL}(Y) = \{\mathbf{S} \subseteq \mathbf{X} \mid \mathrm{BD}(Y) \subseteq \mathbf{S}\}. \tag{20}$$

Comparing the minimal stable variable set and the Markov boundary Besides the similarities in mathematical forms, there exist some connections between the stable variable set and the Markov blanket, and between the minimal stable variable set and the Markov boundary.

Theorem 12. *Under Assumption 2, a stable variable set is also a Markov blanket and the minimal stable variable set is a subset of the Markov boundary, i.e.,*

$$\mathrm{BL}(Y) \subseteq \mathrm{Stable}(Y), \quad \mathrm{MinStable}(Y) \subseteq \mathrm{BD}(Y). \tag{21}$$

The above theorem shows the inclusion relations between those two concepts, and the following example further illustrates an proper inclusion case.

Example 1 (from Strobl and Visweswaran [2016])**.** Let $\mathbf{X} = (X_1, X_2)$ and the data-generating process is given as follows.

$$X_1, X_2 \sim N(0,1), \quad Y = f(X_1) + N\left(0, \rho(X_2)^2\right), \tag{22}$$

where $f(\cdot)$ and $\rho(\cdot)$ are fixed functions. Then

$$\begin{aligned} &\{X_1\} = \mathrm{MinStable}(Y) \subsetneq \mathrm{BD}(Y) = \{X_1, X_2\}, \\ &\{\{X_1, X_2\}\} = \mathrm{BL}(Y) \subsetneq \mathrm{Stable}(Y) = \{\{X_1\}, \{X_1, X_2\}\}. \end{aligned} \tag{23}$$

The following proposition provides the property of the Markov boundary on covariate shift generalization.

Theorem 13. *Under Assumption 1 and Assumption 2, if* $\mathbb{M}$ *is a performance metric that is maximized only when* $P^{te}(Y|\mathbf{X})$ *is estimated accurately and* $\mathbb{L}$ *is a learning algorithm that can approximate any conditional expectation, then*

1. $\mathbf{S}$ *is an optimal predictor of* Y *under the testing distribution* P^{te} *if and only if it is a Markov blanket of* Y *under the training distribution* P^{tr}*, and*
2. $\mathbf{S}$ *is a minimal and optimal predictor of* Y *under the testing distribution* P^{te} *if and only if it is a Markov boundary of* Y *under the training distribution* P^{tr}*.*

Remark. The main difference of Theorem 3 and Theorem 13 is the requirement on the performance metric $\mathbb{M}$. The Markov boundary is minimal and optimal predictor if $\mathbb{M}$ is chosen as maximizing $P^{\text{te}}(Y|\mathbf{X})$. However, for regression problems with the mean squared loss and binary classification problems with the cross entropy loss, $\mathbb{E}_{P^{\text{te}}}[Y|\mathbf{X}]$ is optimal in the testing distribution P^{te}.

As a result, compared with the Markov boundary, the minimal stable variable set can bring two advantages.

1. The conditional independence test is the crux to the precise discovery of the Markov boundary. Shah and Peters [2020] have shown that conditional independence is a particularly difficult hypothesis to test for, which highlights the challenges of discovering the Markov boundary in real-world tasks. However, discovering the minimal stable variable set is relatively easier and proved possible in this paper.
2. In several common machine learning tasks, including regression and binary classification, not all variables in the Markov boundary are necessary. As shown in Example 1, if a variable only affects the variance of the response variable Y, it would not be useful to predict Y when adopting mean squared loss. The minimal stable variable set is proved to be a subset of the Markov boundary and it excludes useless variables in the Markov boundary for covariate shift generalization.

7 Discussions

To conclude, in this paper, we theoretically prove the effectiveness of stable learning algorithms. We show that under ideal conditions, *i.e.* perfectly learned sample weights and infinite samples, the algorithms could identify the minimal stable variable set, which is the minimal set of variables that could provide good predictions under covariate shift. We further provide asymptotic properties and error analysis when the two conditions are not satisfied.

We should notice that the definitions are applicable only when $\mathbb{E}[Y|\mathbf{X}]$ is well defined. This implies that the definitions could be applied to typical regression and binary classification settings, but they may not be applicable in multi-class classification settings. In addition, under regression settings, $\mathbb{E}[Y|\mathbf{X}]$ will not be the solution in other forms of losses. For example, consider the Minkowski loss [Bishop, 2006, Section 1.5.5] given as $L_q = \mathbb{E}[|Y - f(\mathbf{X})|^q]$. It reduces to the expected squared loss when $q = 2$. The minimum of L_q is given by the conditional mean $\mathbb{E}[Y|\mathbf{X}]$ for $q = 2$, which is our case. But the solution becomes the conditional median for $q = 1$ and the conditional mode for $q \to 0$. Nevertheless, we do highlight that the squared loss under regression settings and the cross-entropy loss under binary classification settings are general enough for most potential applications. We leave the theoretical analysis and applications of stable learning algorithms on multi-class classification settings as future work.

References

Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11(1), 2010a.

Constantin F Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification part ii: analysis and extensions. *Journal of Machine Learning Research*, 11(1), 2010b.

S Athey, GW Imbens, and S Wager. Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. arxiv e-prints. *arXiv preprint arXiv:1604.07125*, 25, 2016.

Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007.

Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In *Advances in neural information processing systems*, pages 153–160, 2007.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Frénay Benoît, Mark Van Heeswijk, Yoan Miche, Michel Verleysen, and Amaury Lendasse. Feature selection for nonlinear models with extreme learning machines. *Neurocomputing*, 102:111–124, 2013.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, pages 81–88, 2007.

Christopher M Bishop. Pattern recognition. *Machine learning*, 128(9), 2006.

Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. A review of feature selection methods on synthetic data. *Knowledge and information systems*, 34(3):483–519, 2013.

Shivkumar Chandrasekaran and Ilse CF Ipsen. On the sensitivity of solution components in linear systems of equations. *SIAM Journal on Matrix Analysis and Applications*, 16(1):93–112, 1995.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

José Carlos Cortizo and Ignacio Giraldez. Multi criteria wrapper improvements to naive bayes learning. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 419–427. Springer, 2006.

John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE international conference on computer vision*, pages 2960–2967, 2013.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary C Lipton. A unified view of label shift estimation. *Advances in Neural Information Processing Systems*, 33, 2020.

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.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, 20(1):25–46, 2012.

Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*. Routledge, 2017.

Yue He, Peng Cui, Zheyan Shen, Renzhe Xu, Furui Liu, and Yong Jiang. Daring: Differentiable causal discovery with residual independence. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 596–605, 2021.

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

Biwei Huang, Kun Zhang, Yizhu Lin, Bernhard Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1551–1560, 2018.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19:601–608, 2006.

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

George H John, Ron Kohavi, and Karl Pfleger. Irrelevant features and the subset selection problem. In *Machine learning proceedings 1994*, pages 121–129. Elsevier, 1994.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445, 2009.

Kenji Kira and Larry A Rendell. A practical approach to feature selection. In *Machine learning proceedings 1992*, pages 249–256. Elsevier, 1992.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

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.

Kun Kuang, Ruoxuan Xiong, Peng Cui, Susan Athey, and Bo Li. Stable prediction with model misspecification and agnostic distribution shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4485–4492, 2020a.

Kun Kuang, Hengtao Zhang, Fei Wu, Yueting Zhuang, and Aijun Zhang. Balance-subsampled stable prediction. *arXiv preprint arXiv:2006.04381*, 2020b.

Pat Langley et al. Selection of relevant features in machine learning. In *Proceedings of the AAAI Fall symposium on relevance*, volume 184, pages 245–271, 1994.

Martin HC Law, Mario AT Figueiredo, and Anil K Jain. Simultaneous feature selection and clustering using mixture models. *IEEE transactions on pattern analysis and machine intelligence*, 26(9):1154–1166, 2004.

Chao Liu, Dongxiang Jiang, and Wenguang Yang. Global geometric similarity scheme for feature selection in fault diagnosis. *Expert Systems with Applications*, 41(8):3585–3595, 2014.

Huawen Liu, Lei Liu, and Huijie Zhang. Ensemble gene selection by grouping for microarray data classification. *Journal of biomedical informatics*, 43(1):81–87, 2010a.

Huawen Liu, Lei Liu, and Huijie Zhang. Ensemble gene selection for cancer classification. *Pattern Recognition*, 43(8):2763–2772, 2010b.

Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Heterogeneous risk minimization. In *International Conference on Machine Learning*. PMLR, 2021.

Wei-Yin Loh. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(1):14–23, 2011.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017.

Subramani Mani and Gregory F Cooper. Causal discovery using a bayesian local causal discovery algorithm. In *MEDINFO 2004*, pages 731–735. IOS Press, 2004.

Aditya Menon and Cheng Soon Ong. Linking losses for density ratio and class-probability estimation. In *International Conference on Machine Learning*, pages 304–313. PMLR, 2016.

Bjoern H Menze, B Michael Kelm, Ralf Masuch, Uwe Himmelreich, Peter Bachert, Wolfgang Petrich, and Fred A Hamprecht. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics*, 10(1):1–16, 2009.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

Jose M Pena, Roland Nilsson, Johan Björkegren, and Jesper Tegnér. Towards scalable and data efficient learning of markov boundaries. *International Journal of Approximate Reasoning*, 45(2): 211–232, 2007.

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

Ali Rahimi, Benjamin Recht, et al. Random features for large-scale kernel machines. In *NIPS*, volume 3, page 5. Citeseer, 2007.

Alain Rakotomamonjy. Variable selection using svm-based criteria. *Journal of machine learning research*, 3(Mar):1357–1370, 2003.

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

Zheyan Shen, Peng Cui, Kun Kuang, Bo Li, and Peixuan Chen. Causally regularized learning with agnostic data selection bias. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 411–419, 2018.

Zheyan Shen, Peng Cui, Jiashuo Liu, Tong Zhang, Bo Li, and Zhitang Chen. Stable learning via differentiated variable decorrelation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2185–2193, 2020a.

Zheyan Shen, Peng Cui, Tong Zhang, and Kun Kuang. Stable learning via sample reweighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5692–5699, 2020b.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

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

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Peter L Spirtes, Christopher Meek, and Thomas S Richardson. Causal inference in the presence of latent variables and selection bias. *arXiv preprint arXiv:1302.4983*, 2013.

Alexander Statnikov, Jan Lemeir, and Constantin F Aliferis. Algorithms for discovery of multiple markov boundaries. *The Journal of Machine Learning Research*, 14(1):499–566, 2013.

Amos J Storkey and Masashi Sugiyama. Mixture regression for covariate shift. *Advances in neural information processing systems*, 19:1337, 2007.

Eric V Strobl and Shyam Visweswaran. Markov boundary discovery with ridge regularized linear models. *Journal of Causal inference*, 4(1):31–48, 2016.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

Masashi Sugiyama, Takafumi Kanamori, Taiji Suzuki, Shohei Hido, Jun Sese, Ichiro Takeuchi, and Liwei Wang. A density-ratio framework for statistical data processing. *IPSJ Transactions on Computer Vision and Applications*, 1:183–208, 2009.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Ioannis Tsamardinos and Constantin F Aliferis. Towards principled feature selection: Relevancy, filters and wrappers. In *International Workshop on Artificial Intelligence and Statistics*, pages 300–307. PMLR, 2003.

Ioannis Tsamardinos, Constantin F Aliferis, and Alexander Statnikov. Time and sample efficient discovery of markov blankets and direct causal relations. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 673–678, 2003a.

Ioannis Tsamardinos, Constantin F Aliferis, Alexander R Statnikov, and Er Statnikov. Algorithms for large scale markov blanket discovery. In *FLAIRS conference*, volume 2, pages 376–380, 2003b.

Ryan J Urbanowicz, Melissa Meeker, William La Cava, Randal S Olson, and Jason H Moore. Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, 85:189–203, 2018.

Xiao Wang, Shaohua Fan, Kun Kuang, Chuan Shi, Jiawei Liu, and Bai Wang. Decorrelated clustering with data selection bias. *arXiv preprint arXiv:2006.15874*, 2020.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9593–9602, 2021.

Zhaoquan Yuan, Xiao Peng, Xiao Wu, Bing-kun Bao, and Changsheng Xu. Meta-learning causal feature selection for stable prediction. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. Devlbert: Learning deconfounded visio-linguistic representations. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4373–4382, 2020.

Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5372–5382, 2021.

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

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. PMLR, 2020.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *arXiv preprint arXiv:2103.02503*, 2021.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

A Proofs

A.1 Proof of Theorem 3

Lemma A.1. *Under Assumption 2, if* $\mathbb{M}$ *is a performance metric that is maximized only when* $\mathbb{E}_{P^{tr}}[Y|\mathbf{X}]$ *is estimated accurately and* $\mathbb{L}$ *is a learning algorithm that can approximate any conditional expectation, then*

1. $\mathbf{S}$ *is an optimal predictor of* Y *if and only if it is a stable variable set of* Y *under distribution* P^{tr}*, and*
2. $\mathbf{S}$ *is a minimal and optimal predictor of* Y *if and only if it is a minimal stable variable set of* Y *under distribution* P^{tr}*.*

Proof. We omit the subscript of $\mathbb{E}_{P^{\text{tr}}}[\cdot|\cdot]$ for simplicity.

Consider the first part. On the one hand, if $\mathbf{S}$ is a stable variable set of Y, then $\mathbb{E}[Y|\mathbf{X}] = \mathbb{E}[Y|\mathbf{S}]$ by definition. Hence $\mathbf{S}$ is an optimal predictor because $\mathbb{E}[Y|\mathbf{X}] = \mathbb{E}[Y|\mathbf{S}]$ can be approximated perfectly by $\mathbb{L}$ and $\mathbb{M}$ will be maximized. On the other hand, assume $\mathbf{S}$ is an optimal predictor but not a stable variable set, which implies that $\mathbb{E}[Y|\mathbf{S}] \neq \mathbb{E}[Y|\mathbf{X}]$. $\mathbf{X}$ is a stable variable set by definition. Hence, By first part of the proof, $\mathbf{X}$ is an optimal predictor of Y, similar to $\mathbf{S}$. Therefore, the following should hold: $\mathbb{E}[Y|\mathbf{X}] = \mathbb{E}[Y|\mathbf{S}]$, which contradicts the assumption that $\mathbf{S}$ is not a stable variable set. As a result, $\mathbf{S}$ is a stable variable set of Y.

Consider the second part. On the one hand, if $\mathbf{S}$ is the minimal stable variable set of Y, then it is also a stable variable set of Y. So $\mathbf{S}$ is an optimal predictor. Moreover, by the definition of the minimal stable variable set, no proper subset of $\mathbf{S}$ is a stable variable set of Y. Therefore, no proper subset of $\mathbf{S}$ satisfies the definition of an optimal predictor. Thus, $\mathbf{S}$ is a minimal and optimal predictor of Y. On the other hand, assume $\mathbf{S}$ is a minimal and optimal predictor of Y. Then, $\mathbf{S}$ is also an optimal predictor of Y, which implies that $\mathbf{S}$ is a stable variable set of Y. By the definition of minimality, no proper subset of $\mathbf{S}$ is a minimal and optimal predictor. Hence, no proper subset of $\mathbf{S}$ is a stable variable set of Y. As a result, $\mathbf{S}$ is the minimal stable variable of Y. □

Now we prove the original theorem.

Proof. It is obvious that $\mathbb{E}_{P^{\text{tr}}}[Y|\mathbf{X}] = \mathbb{E}_{P^{\text{te}}}[Y|\mathbf{X}]$ from Assumption 1. As a result, the original theorem is proved according to Lemma A.1. □

A.2 Proof of Theorem 4

The proof is based on the following intersection property.

Lemma A.2. *Under Assumption 2, if* $\mathbf{S}_1, \mathbf{S}_2 \in \text{Stable}(Y)$*, then* $\mathbf{S}_1 \cap \mathbf{S}_2 \in \text{Stable}(Y)$*.*

Proof. Let $\mathbf{S} = \mathbf{S}_1 \cap \mathbf{S}_2$, $\bar{\mathbf{S}}_1 = \mathbf{S}_1 \backslash \mathbf{S}$, $\bar{\mathbf{S}}_2 = \mathbf{S}_2 \backslash \mathbf{S}$, and $\bar{\mathbf{X}} = \mathbf{X} \backslash (\mathbf{S}_1 \cup \mathbf{S}_2)$. Then $\mathbf{X} = (\mathbf{S}, \bar{\mathbf{S}}_1, \bar{\mathbf{S}}_2, \bar{\mathbf{X}})$.

By definition, $\forall \mathbf{s} \in \mathcal{S}, \bar{\mathbf{s}}_1 \in \bar{\mathcal{S}}_1, \bar{\mathbf{s}}_2 \in \bar{\mathcal{S}}_2, \bar{\mathbf{x}} \in \bar{\mathcal{X}}$, $\mathbb{E}[Y|\mathbf{S} = \mathbf{s}, \bar{\mathbf{S}}_1 = \bar{\mathbf{s}}_1] = \mathbb{E}[Y|\mathbf{S} = \mathbf{s}, \bar{\mathbf{S}}_2 = \bar{\mathbf{s}}_2] = \mathbb{E}[Y|\mathbf{S} =$

$\mathbf{s}, \bar{\mathbf{S}}_1 = \bar{\mathbf{s}}_1, \bar{\mathbf{S}}_2 = \bar{\mathbf{s}}_2, \bar{\mathbf{X}} = \bar{\mathbf{x}}]$. Let $\mathbf{x} = (\mathbf{s}, \bar{\mathbf{s}}_1, \bar{\mathbf{s}}_2, \bar{\mathbf{x}})$. Under Assumption 2,

$$
\begin{aligned}
&\mathbb{E}[Y|\mathbf{S} = \mathbf{s}] \\
=&\int_{\mathcal{Y}} y P^{\mathrm{tr}}(Y = y|\mathbf{S} = \mathbf{s}) \mathrm{d}y \\
=&\int_{\mathcal{Y}} \int_{\bar{\mathcal{S}}_1} y P^{\mathrm{tr}}(Y = y|\mathbf{S} = \mathbf{s}, \bar{\mathbf{S}}_1 = \bar{\mathbf{s}}_1) P^{\mathrm{tr}}(\bar{\mathbf{S}}_1 = \bar{\mathbf{s}}_1|\mathbf{S} = \mathbf{s}) \mathrm{d}\bar{\mathbf{s}}_1 \mathrm{d}y \\
=&\int_{\bar{\mathcal{S}}_1} \mathbb{E}[Y|\mathbf{S} = \mathbf{s}, \bar{\mathbf{S}}_1 = \bar{\mathbf{s}}_1] P^{\mathrm{tr}}(\bar{\mathbf{S}}_1 = \bar{\mathbf{s}}_1|\mathbf{S} = \mathbf{s}) \mathrm{d}\bar{\mathbf{s}}_1 \\
=&\mathbb{E}[Y|\mathbf{S} = \mathbf{s}, \bar{\mathbf{S}}_2 = \bar{\mathbf{s}}_2] \int_{\bar{\mathcal{S}}_1} P^{\mathrm{tr}}(\bar{\mathbf{S}}_1 = \bar{\mathbf{s}}_1|\mathbf{S} = \mathbf{s}) \mathrm{d}\bar{\mathbf{s}}_1 \\
=&\mathbb{E}[Y|\mathbf{S} = \mathbf{s}, \bar{\mathbf{S}}_2 = \bar{\mathbf{s}}_2] = \mathbb{E}[Y|\mathbf{S} = \mathbf{s}, \bar{\mathbf{S}}_1 = \bar{\mathbf{s}}_1, \bar{\mathbf{S}}_2 = \bar{\mathbf{s}}_2, \bar{\mathbf{X}} = \bar{\mathbf{x}}] = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}].
\end{aligned}
$$

As a result, $\mathbf{S} \in \mathrm{Stable}(Y)$. □

Now we prove the original theorem.

Proof. We first prove the uniqueness of the minimal stable variable set. Suppose there are two minimal stable variable sets w.r.t. Y, denoted as $\mathrm{MinStable}_1(Y)$ and $\mathrm{MinStable}_2(Y)$. By definition, $\mathrm{MinStable}_1(Y), \mathrm{MinStable}_2(Y) \in \mathrm{Stable}(Y)$. Under Assumption 2, according to Lemma A.2, $\mathrm{MinStable}_1(Y) \cap \mathrm{MinStable}_2(Y) \in \mathrm{Stable}(Y)$. Because $\mathrm{MinStable}_1(Y)$ has no proper subset that is in $\mathrm{Stable}(Y)$, we have $\mathrm{MinStable}_1(Y) \cap \mathrm{MinStable}_2(Y) = \mathrm{MinStable}_1(Y)$. Similarly, $\mathrm{MinStable}_1(Y) \cap \mathrm{MinStable}_2(Y) = \mathrm{MinStable}_2(Y)$, which means $\mathrm{MinStable}_1(Y) = \mathrm{MinStable}_2(Y)$.

Next, we prove the exact form of the stable variable sets. Let

$$\Omega = \{\mathbf{S} \subseteq \mathbf{X} \mid \mathrm{MinStable}(Y) \subseteq \mathbf{S}\}.$$

On the one hand, $\forall \mathbf{S} \in \mathrm{Stable}(Y)$, according to Lemma A.2, $\mathbf{S} \cap \mathrm{MinStable}(Y) \in \mathrm{Stable}(Y)$. Because of the minimality of $\mathrm{MinStable}(Y)$, $|\mathbf{S} \cap \mathrm{MinStable}(Y)| \geq |\mathrm{MinStable}(Y)|$. As a result, $\mathrm{MinStable}(Y) \subseteq \mathbf{S}$ and $\mathbf{S} \in \Omega$. Hence $\mathrm{Stable}(Y) \subseteq \Omega$.

On the other hand, $\forall \mathbf{S} \in \Omega$, let $\mathbf{D} = \mathrm{MinStable}(Y)$, $\mathbf{W} = \mathbf{S} \backslash \mathbf{D}$, and $\bar{\mathbf{X}} = \mathbf{X} \backslash \mathbf{S}$. Then $\forall \mathbf{d} \in \mathcal{D}, \mathbf{w} \in \mathcal{W}$, $\mathbf{s} = (\mathbf{d}, \mathbf{w})$, we can get

$$
\begin{aligned}
&\mathbb{E}[Y|\mathbf{S} = \mathbf{s}] = \int_{\mathcal{Y}} y P^{\mathrm{tr}}(Y = y|\mathbf{D} = \mathbf{d}, \mathbf{W} = \mathbf{w}) \mathrm{d}y \\
=&\int_{\mathcal{Y}} \int_{\bar{\mathcal{X}}} y P^{\mathrm{tr}}(Y = y|\mathbf{D} = \mathbf{d}, \mathbf{W} = \mathbf{w}, \bar{\mathbf{X}} = \bar{\mathbf{x}}) P^{\mathrm{tr}}(\bar{\mathbf{X}} = \bar{\mathbf{x}}|\mathbf{D} = \mathbf{d}, \mathbf{W} = \mathbf{w}) \mathrm{d}\bar{\mathbf{x}} \mathrm{d}y \\
=&\int_{\bar{\mathcal{X}}} P^{\mathrm{tr}}(\bar{\mathbf{X}} = \bar{\mathbf{x}}|\mathbf{D} = \mathbf{d}, \mathbf{W} = \mathbf{w}) \mathbb{E}[Y|\mathbf{D} = \mathbf{d}, \mathbf{W} = \mathbf{w}, \bar{\mathbf{X}} = \bar{\mathbf{x}}] \mathrm{d}\bar{\mathbf{x}} \\
=&\int_{\bar{\mathcal{X}}} P^{\mathrm{tr}}(\bar{\mathbf{X}} = \bar{\mathbf{x}}|\mathbf{D} = \mathbf{d}, \mathbf{W} = \mathbf{w}) \mathbb{E}[Y|\mathbf{D} = \mathbf{d}] \mathrm{d}\bar{\mathbf{x}} \\
=&\mathbb{E}[Y|\mathbf{D} = \mathbf{d}] \int_{\bar{\mathcal{X}}} P^{\mathrm{tr}}(\bar{\mathbf{X}} = \bar{\mathbf{x}}|\mathbf{D} = \mathbf{d}, \mathbf{W} = \mathbf{w}) \mathrm{d}\bar{\mathbf{x}} \\
=&\mathbb{E}[Y|\mathbf{D} = \mathbf{d}].
\end{aligned}
$$

As a result, $\mathbf{S}$ satisfies the requirement of stable variable sets and $\mathbf{S} \in \mathrm{Stable}(Y)$. Hence $\Omega \subseteq \mathrm{Stable}(Y)$.

To conclude, $\mathrm{Stable}(Y) \subseteq \Omega$ and $\Omega \subseteq \mathrm{Stable}(Y)$, which results in $\Omega = \mathrm{Stable}(Y)$. □

A.3 Proof of Theorem 5

We need the following lemma first.

Lemma A.3. *Let $w \in \mathcal{W}$ be a weighting function, and $\tilde{P}_w$ be the corresponding weighted distribution. Then $\tilde{P}_w(Y|\mathbf{X}) = P^{tr}(Y|\mathbf{X})$.*

Proof. $\forall \mathbf{x} \in \mathcal{X}, y \in \mathcal{Y}$,

$$\tilde{P}_w(Y = y|\mathbf{X} = \mathbf{x}) = \frac{\tilde{P}_w(Y = y, \mathbf{X} = \mathbf{x})}{\tilde{P}_w(\mathbf{X} = \mathbf{x})} = \frac{P^{\text{tr}}(Y = y, \mathbf{X} = \mathbf{x})w(\mathbf{x})}{\int_{y'} \tilde{P}_w(\mathbf{X}, Y = y')\mathrm{d}y'}$$
$$= \frac{P^{\text{tr}}(Y = y, \mathbf{X} = \mathbf{x})w(\mathbf{x})}{w(\mathbf{x})\int_{y'} P^{\text{tr}}(\mathbf{X} = x, Y = y')\mathrm{d}y'} = \frac{P^{\text{tr}}(Y = y, \mathbf{X} = \mathbf{x})}{P^{\text{tr}}(\mathbf{X} = \mathbf{x})} = P^{\text{tr}}(Y = y|\mathbf{X} = \mathbf{x}).$$

□

Now we prove the original theorem.

Proof. Let $\mathbf{X}_{-i}$ denote variables other than X_i and $\mathcal{X}_{-i}$ denote the support of $\mathbf{X}_{-i}$.

Given $X_i \notin \text{MinStable}(Y)$, there exists a function $f : \mathcal{X}_{-i} \to \mathcal{Y}$ such that $\mathbb{E}_{P^{\text{tr}}(\mathbf{X},Y)}[Y|\mathbf{X}] = f(\mathbf{X}_{-i})$. According to Lemma A.3, $\mathbb{E}_{\tilde{P}_w(\mathbf{X},Y)}[Y|\mathbf{X}] = \mathbb{E}_{P^{\text{tr}}(\mathbf{X},Y)}[Y|\mathbf{X}] = f(\mathbf{X}_{-i})$. As a result, because $\mathbb{E}_{P^{\text{tr}}(\mathbf{X})}\left[w(\mathbf{X})\|\mathbf{X}\|_2^2\right] < \infty$ and $\mathbb{E}_{P^{\text{tr}}(\mathbf{X},Y)}\left[w(\mathbf{X})Y^2\right] < \infty$, the covariance between X_i and Y under $\tilde{P}_w$ is

$$\begin{aligned}\text{Cov}_{\tilde{P}_w}[X_iY] =& \mathbb{E}_{\tilde{P}_w(X_i,Y)}[X_iY] - \mathbb{E}_{\tilde{P}_w(X_i)}[X_i]\mathbb{E}_{\tilde{P}_w(Y)}[Y] \\ =& \mathbb{E}_{\tilde{P}_w(\mathbf{X})}\left[X_i\mathbb{E}_{\tilde{P}_w(\mathbf{X},Y)}[Y|\mathbf{X}]\right] - \mathbb{E}_{\tilde{P}_w(X_i)}[X_i]\mathbb{E}_{\tilde{P}_w(\mathbf{X})}\left[\mathbb{E}_{\tilde{P}_w(\mathbf{X},Y)}[Y|\mathbf{X}]\right] \\ =& \mathbb{E}_{\tilde{P}_w(\mathbf{X})}[X_if(\mathbf{X}_{-i})] - \mathbb{E}_{\tilde{P}_w(X_i)}[X_i]\mathbb{E}_{\tilde{P}_w(\mathbf{X}_{-i})}[f(\mathbf{X}_{-i})] = 0.\end{aligned}$$

The last equation is due to the independence between X_i and $\mathbf{X}_{-i}$. As a result, the coefficient $\boldsymbol{\beta}_w(X_i)$ on X_i is

$$\boldsymbol{\beta}_w(X_i) = \text{Var}_{\tilde{P}_w}(X_i)^{-1}\,\text{Cov}_{\tilde{P}_w}[X_iY] = 0.$$

□

A.4 Proof of Theorem 6

Proof. Let $\mathbf{X}_{-i}$ denote the rest variable except X_i and P^{tr}_{-i} denote the marginal distribution of P^{tr} on $\mathbf{X}_{-i}$. Because $X_i \in \text{MinStable}(Y)$, $\mathbb{E}_{P^{\text{tr}}(\mathbf{X},Y)}[Y|\mathbf{X}]$ depends on X_i. Hence, there exists a probability density function $\tilde{P}_{-i}$ with the same support of P^{tr}_{-i} that satisfies

1. $\mathbf{X}_{-i}$ are mutually independent under $\tilde{P}_{-i}$, and
2. $g(X_i) \triangleq \mathbb{E}_{\tilde{P}_{-i}(\mathbf{X}_{-i})}[\mathbb{E}_{P^{\text{tr}}(\mathbf{X},Y)}[Y|\mathbf{X}_{-i}, X_i]]$ depends on X_i.

Moreover, there exist a probability density function $\tilde{P}_i$ with the same support of P^{tr}_i that satisfies $g(X_i)$ is linearly correlated with X_i under $\tilde{P}_i$.

Let $\tilde{P}$ be the joint distribution on $(\mathbf{X}, Y)$ and $\tilde{P}(\mathbf{X}_{-i}, X_i, Y) = \tilde{P}_{-i}(\mathbf{X}_{-i})\tilde{P}_i(X_i)P^{\text{tr}}(Y|X)$. Hence,

$$\mathbb{E}_{\tilde{P}(X_i,Y)}[Y|X_i] = \mathbb{E}_{\tilde{P}_{-i}(\mathbf{X}_{-i})}[\mathbb{E}_{P^{\text{tr}}(\mathbf{X},Y)}[Y|\mathbf{X}_{-i}, X_i]] = g(X_i).$$

Let $w(\mathbf{X}) = \tilde{P}(\mathbf{X})/P^{\text{tr}}(\mathbf{X})$. Because $\mathbb{E}_{P^{\text{tr}}(\mathbf{X},Y)}[Y|\mathbf{X}]$ depends on X_i, $\text{Var}_{P^{\text{tr}}}(X_i) > 0$. Hence, $\text{Var}_{\tilde{P}}(X_i) > 0$. As a result, the coefficient on X_i is

$$
\begin{aligned}
&\beta_w(X_i) \\
=&\frac{1}{\text{Var}_{\tilde{P}_i}(X_i)}\left(\mathbb{E}_{P^{\text{tr}}(\mathbf{X},Y)}[w(\mathbf{X})X_iY] - \mathbb{E}_{P^{\text{tr}}(\mathbf{X})}[w(\mathbf{X})X_i]\mathbb{E}_{P^{\text{tr}}(\mathbf{X},Y)}[w(\mathbf{X})Y]\right) \\
=&\frac{1}{\text{Var}_{\tilde{P}_i}(X_i)}\left(\mathbb{E}_{\tilde{P}(X_i,Y)}[X_iY] - \mathbb{E}_{\tilde{P}(X_i)}[X_i]\mathbb{E}_{\tilde{P}(Y)}[Y]\right) \\
=&\frac{1}{\text{Var}_{\tilde{P}_i}(X_i)}\left(\mathbb{E}_{\tilde{P}(X_i)}\left[X_i\mathbb{E}_{\tilde{P}(X_i,Y)}[Y|X_i]\right] - \mathbb{E}_{\tilde{P}_i(X_i)}[X_i]\mathbb{E}_{\tilde{P}_i(X_i)}\left[\mathbb{E}_{\tilde{P}(X_i,Y)}[Y|X_i]\right]\right) \\
=&\frac{1}{\text{Var}_{\tilde{P}_i}(X_i)}\left(\mathbb{E}_{\tilde{P}_i(X_i)}[X_ig(X_i)] - \mathbb{E}_{\tilde{P}_i(X_i)}[X_i]\mathbb{E}_{\tilde{P}_i(X_i)}[g(X_i)]\right) \neq 0.
\end{aligned}
$$

□

A.5 Proof of Theorem 7

Proof. For convenience, we append an 1 in the front of feature variable, i.e., $\ddot{\mathbf{X}} = (1, \mathbf{X})$ and $\ddot{\mathbf{x}}^{(i)} = \left(1, \mathbf{x}^{(i)}\right)$. Then the loss of weighted least squares becomes

$$
\hat{\beta}_w^{(n)} = \arg\min_{\beta} \sum_{i=1}^{n} w\left(\mathbf{x}^{(i)}\right)\left(\beta^T\ddot{\mathbf{x}}^{(i)} - y^{(i)}\right)^2.
$$

Hence,

$$
\hat{\beta}_w^{(n)} = \left(\frac{1}{n}\sum_{i=1}^{n} w\left(\mathbf{x}^{(i)}\right)\ddot{\mathbf{x}}^{(i)}\left(\ddot{\mathbf{x}}^{(i)}\right)^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} w\left(\mathbf{x}^{(i)}\right)\ddot{\mathbf{x}}^{(i)}y^{(i)}\right).
$$

By the weak law of large numbers,

$$
\begin{cases}
\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} w\left(\mathbf{x}^{(i)}\right)\ddot{\mathbf{x}}^{(i)}\left(\ddot{\mathbf{x}}^{(i)}\right)^T \xrightarrow{p} \mathbb{E}_{P(\mathbf{X})}\left[w(\mathbf{X})\ddot{\mathbf{X}}\ddot{\mathbf{X}}^T\right] = \mathbb{E}_{\tilde{P}_w(\mathbf{X})}\left[\ddot{\mathbf{X}}\ddot{\mathbf{X}}^T\right], \\
\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} w\left(\mathbf{x}^{(i)}\right)\ddot{\mathbf{x}}^{(i)}y^{(i)} \xrightarrow{p} \mathbb{E}_{P(\mathbf{X},Y)}\left[w(\mathbf{X})\ddot{\mathbf{X}}Y\right] = \mathbb{E}_{\tilde{P}_w(\mathbf{X},Y)}\left[\ddot{\mathbf{X}}Y\right].
\end{cases}
$$

Because covariance matrix $\text{Cov}_{\tilde{P}_w}[\mathbf{X}]$ is invertible,

$$
\begin{aligned}
&\left(\mathbb{E}_{\tilde{P}_w(\mathbf{X})}\left[\ddot{\mathbf{X}}\ddot{\mathbf{X}}^T\right]\right)^{-1} \\
=&\begin{pmatrix} 1 & \mathbb{E}_{\tilde{P}_w(\mathbf{X})}[\mathbf{X}]^T \\ \mathbb{E}_{\tilde{P}_w(\mathbf{X})}[\mathbf{X}] & \mathbb{E}_{\tilde{P}_w(\mathbf{X})}\left[\mathbf{X}\mathbf{X}^T\right] \end{pmatrix}^{-1} \\
=&\begin{pmatrix} 1 + \mathbb{E}_{\tilde{P}_w(\mathbf{X})}[\mathbf{X}]^T\,\text{Cov}_{\tilde{P}_w}[\mathbf{X}]^{-1}\mathbb{E}_{\tilde{P}_w(\mathbf{X})}[\mathbf{X}] & -\mathbb{E}_{\tilde{P}_w(\mathbf{X})}[\mathbf{X}]^T\,\text{Cov}_{\tilde{P}_w}[\mathbf{X}]^{-1} \\ -\text{Cov}_{\tilde{P}_w}[\mathbf{X}]^{-1}\mathbb{E}_{\tilde{P}_w(\mathbf{X})}[\mathbf{X}] & \text{Cov}_{\tilde{P}_w}[\mathbf{X}]^{-1} \end{pmatrix}.
\end{aligned}
$$

Hence, $\mathbb{E}_{\tilde{P}_w(\mathbf{X})}\left[\ddot{\mathbf{X}}\ddot{\mathbf{X}}^T\right]$ is invertible.

Finally, because function $g(A, b) = A^{-1}b$ is continuous at $\left(\mathbb{E}_{\tilde{P}_w(\mathbf{X})}\left[\ddot{\mathbf{X}}\ddot{\mathbf{X}}^T\right], \mathbb{E}_{\tilde{P}_w(\mathbf{X})}\left[\ddot{\mathbf{X}}Y\right]\right)$, by continuous mapping theorem, we have

$$
\hat{\beta}_w^{(n)} \xrightarrow{p} \left(\mathbb{E}_{\tilde{P}_w(\mathbf{X})}\left[\ddot{\mathbf{X}}\ddot{\mathbf{X}}^T\right]\right)^{-1}\left(\mathbb{E}_{\tilde{P}_w(\mathbf{X})}\left[\ddot{\mathbf{X}}Y\right]\right) = \beta_w.
$$

□

A.6 Proof of Theorem 10

The theorem is inspired by the following lemma.

Lemma A.4 (Chandrasekaran and Ipsen [1995]). *Suppose $Ax = b$ and $\hat{A}\hat{x} = \hat{b}$. Suppose $\|A^{-1}\|\|A-\hat{A}\| < 1$, then*

$$\frac{\|x-\hat{x}\|}{\|x\|} \le \frac{\|A\|\|A^{-1}\|}{1-\|A^{-1}\|\|A-\hat{A}\|}\left(\frac{\|A-\hat{A}\|}{\|A\|}+\frac{\|b-\hat{b}\|}{\|b\|}\right).$$

Now we prove the original theorem.

Proof. Let $\Delta w(\mathbf{X}) = \hat{w}(\mathbf{X}) - w(\mathbf{X})$ and

$$\begin{cases} \Delta A = \mathbb{E}_{P^{\mathrm{tr}}(\mathbf{X})}\left[\Delta w(\mathbf{X})\mathbf{X}\mathbf{X}^T\right], \\ \Delta b = \mathbb{E}_{P^{\mathrm{tr}}(\mathbf{X},Y)}\left[\Delta w(\mathbf{X})\mathbf{X}Y\right]. \end{cases}$$

Then

$$\begin{aligned}
&\|\Delta A\|_2 \\
=&\sup_{\|c\|_2=1}\|\Delta A\cdot c\|_2 = \sup_{\|c\|_2=1}\left\|\mathbb{E}_{P^{\mathrm{tr}}(\mathbf{X})}\left[\Delta w(\mathbf{X})\mathbf{X}\mathbf{X}^T c\right]\right\|_2 \\
\le&\sup_{\|c\|_2=1}\mathbb{E}_{P^{\mathrm{tr}}(\mathbf{X})}\left[\Delta w(\mathbf{X})\|\mathbf{X}\mathbf{X}^T c\|_2\right] && \text{(triangle inequality of norms)} \\
\le&\sup_{\|c\|_2=1}\sqrt{\mathbb{E}_{P^{\mathrm{tr}}(\mathbf{X})}\left[\Delta w(\mathbf{X})^2\right]\mathbb{E}_{P^{\mathrm{tr}}(\mathbf{X})}\left[\|\mathbf{X}\mathbf{X}^T c\|_2^2\right]} && \text{(Cauchy–Schwarz inequality)} \\
=&\epsilon\sup_{\|c\|_2=1}\sqrt{\mathbb{E}_{P^{\mathrm{tr}}(\mathbf{X})}\left[\|\mathbf{X}\mathbf{X}^T c\|_2^2\right]} && \left(\mathbb{E}\left[\Delta w(\mathbf{X})^2\right]=\epsilon\right) \\
\le&\epsilon\sqrt{\mathbb{E}_{P^{\mathrm{tr}}(\mathbf{X})}\left[\sup_{\|c\|_2=1}\|\mathbf{X}\mathbf{X}^T c\|_2^2\right]} && (\sup\mathbb{E}[\cdot]\le\mathbb{E}[\sup\cdot]) \\
=&\epsilon\sqrt{\mathbb{E}_{P^{\mathrm{tr}}(\mathbf{X})}\left[\|\mathbf{X}\|_2^4\right]} && \text{(see below)} \\
\le&\epsilon\cdot C.
\end{aligned}$$

Here

$$\begin{aligned}
&\sup_{\|c\|_2=1}\|\mathbf{X}\mathbf{X}^T c\|_2^2 = \sup_{\|c\|_2=1} c^T\mathbf{X}\mathbf{X}^T\mathbf{X}\mathbf{X}^T c = \sup_{\|c\|_2=1}\det\left(\mathbf{X}^T\mathbf{X}\mathbf{X}^T cc^T\mathbf{X}\right) \\
=&\mathbf{X}^T\mathbf{X}\sup_{\|c\|_2=1}\det\left(\mathbf{X}^T cc^T\mathbf{X}\right) = \|\mathbf{X}\|_2^2\sup_{\|c\|_2=1}(\mathbf{X}^T c)^2 = \|\mathbf{X}\|_2^4.
\end{aligned}$$

And

$$\begin{aligned}
\|\Delta b\|_2 &= \|\mathbb{E}_{P^{\mathrm{tr}}(\mathbf{X},Y)}\left[\Delta w(\mathbf{X})\mathbf{X}Y\right]\|_2 \\
&\le \mathbb{E}_{P^{\mathrm{tr}}(\mathbf{X},Y)}\left[\Delta w(\mathbf{X})\|\mathbf{X}Y\|_2\right] && \text{(triangle inequality of norms)} \\
&\le \sqrt{\mathbb{E}_{P^{\mathrm{tr}}(\mathbf{X})}\left[\Delta w(\mathbf{X})^2\right]\mathbb{E}_{P^{\mathrm{tr}}(\mathbf{X},Y)}\left[\|\mathbf{X}Y\|_2^2\right]} && \text{(Cauchy–Schwarz inequality)} \\
&= \epsilon\sqrt{\mathbb{E}_{P^{\mathrm{tr}}(\mathbf{X},Y)}\left[\|\mathbf{X}Y\|_2^2\right]} && \left(\mathbb{E}_{P^{\mathrm{tr}}}\left[\Delta w(\mathbf{X})^2\right]=\epsilon\right) \\
&\le \epsilon\cdot C.
\end{aligned}$$

In addition, $A\boldsymbol{\beta}_w = b$ and $(A+\Delta A)\boldsymbol{\beta}_{\hat{w}} = b+\Delta b$. As a result, according to Lemma A.4,

$$\begin{aligned}\frac{\|\boldsymbol{\beta}_{\hat{w}}-\boldsymbol{\beta}_w\|_2}{\|\boldsymbol{\beta}_w\|_2} &\le \frac{\|A\|_2\|A^{-1}\|_2}{1-\|A^{-1}\|_2\|\Delta A\|_2}\left(\frac{\|\Delta A\|_2}{\|A\|_2}+\frac{\|\Delta b\|_2}{\|b\|_2}\right)\\ &\le \frac{\|A\|_2\|A^{-1}\|_2}{1-\epsilon C\|A^{-1}\|_2}\left(\frac{\epsilon C}{\|A\|_2}+\frac{\epsilon C}{\|b\|_2}\right)\\ &= \frac{\epsilon C\|A^{-1}\|_2}{1-\epsilon C\|A^{-1}\|_2}\cdot\left(1+\frac{\|A\|_2}{\|b\|_2}\right).\end{aligned}$$

□

A.7 Proof of Proposition 11

The proof is based on the following lemma.

Lemma A.5 (Intersection Property)**.** *Under Assumption 2, let* $\mathbf{V}_1$, $\mathbf{V}_2$, *and* $\mathbf{S}$ *be subset of* $\mathbf{X}$. *Then,*

$$Y \perp \mathbf{V}_1 \mid (\mathbf{S}\cup\mathbf{V}_2) \;\&\; Y \perp \mathbf{V}_2 \mid (\mathbf{S}\cup\mathbf{V}_1) \Longrightarrow Y \perp (\mathbf{V}_1\cup\mathbf{V}_2)\mid \mathbf{S}.$$

The proof of Lemma A.5 can be found in [Pearl, 2014, Section 3.1.2]. Now we prove the original theorem.

Proof. According to Statnikov et al. [2013], if the distribution P^{tr} satisfies the intersetion property, then there exists a unique Markov boundary of Y.

Next we prove the exact form of the Markov blankets. On the one hand, from Lemma A.5, we can know that under Assumption 2, if $\mathbf{S}_1$ and $\mathbf{S}_2$ are Markov blankets of Y, so does $\mathbf{S}_1\cap\mathbf{S}_2$. As a result, for any $\mathbf{S}\in\text{BL}(Y)$, $\mathbf{S}\cap\text{BD}(Y)\in\text{BL}(Y)$. Because $\text{BD}(Y)$ is the minimal element in $\text{BL}(Y)$, we have $|\mathbf{S}\cap\text{BD}(Y)|\ge|\text{BD}(Y)|$. Hence, $\text{BD}(Y)\subseteq\mathbf{S}$.

On the other hand, for any $\mathbf{S}$ that $\text{BD}(Y)\subseteq\mathbf{S}\subseteq\mathbf{X}$. Let $\mathbf{V}=\mathbf{X}\backslash\mathbf{S}$ and $\mathbf{W}=\mathbf{S}\backslash\text{BD}(Y)$. Then

$$\begin{aligned}P^{\text{tr}}(Y,\mathbf{V}|\mathbf{S}) =& \frac{P^{\text{tr}}(Y,\mathbf{V},\text{BD}(Y),\mathbf{W})}{P^{\text{tr}}(\mathbf{S})} = \frac{P^{\text{tr}}(Y,\mathbf{V},\mathbf{W}|\text{BD}(Y))P^{\text{tr}}(\text{BD}(Y))}{P^{\text{tr}}(\mathbf{S})}\\ =& \frac{P^{\text{tr}}(Y|\text{BD}(Y))P^{\text{tr}}(\mathbf{V},\mathbf{W}|\text{BD}(Y))P^{\text{tr}}(\text{BD}(Y))}{P^{\text{tr}}(\mathbf{S})}\\ =& \frac{P^{\text{tr}}(Y|\text{BD}(Y))P^{\text{tr}}(\mathbf{V},\mathbf{W},\text{BD}(Y))}{P^{\text{tr}}(\mathbf{S})}\\ =& \frac{P^{\text{tr}}(Y|\text{BD}(Y))P^{\text{tr}}(\mathbf{V},\mathbf{S})}{P^{\text{tr}}(\mathbf{S})} = P^{\text{tr}}(Y|\mathbf{S})P^{\text{tr}}(\mathbf{V}|\mathbf{S}).\end{aligned}$$

As a result, $Y\perp\mathbf{V}\mid\mathbf{S}$ and $\mathbf{S}$ is a Markov blanket of Y. To conclude, $\text{BL}(Y)=\{\mathbf{S}\subseteq\mathbf{X}\mid\text{BD}(Y)\subseteq\mathbf{S}\}$. □

A.8 Proof of Theorem 12

Proof. $\forall\mathbf{S}\in\text{BL}(Y)$, $Y\perp(\mathbf{X}\backslash\mathbf{S})\mid\mathbf{S}$. Hence $\mathbb{E}[Y|\mathbf{X}]=\mathbb{E}[Y|\mathbf{S}]$ and $\mathbf{S}\in\text{Stable}(Y)$, which implies $\text{BL}(Y)\subseteq\text{Stable}(Y)$.

Therefore, $\forall\mathbf{S}\in\text{BL}(Y)$, $\mathbf{S}\in\text{Stable}(Y)$. According to Theorem 4, $\text{MinStable}(Y)\subseteq\mathbf{S}$. In particular, let $\mathbf{S}=\text{BD}(Y)\in\text{BL}(Y)$ and we have $\text{MinStable}(Y)\subseteq\text{BD}(Y)$. □

A.9 Proof of Theorem 13

The proof is based on the following proposition.

Proposition A.6 (Statnikov et al. [2013]; Strobl and Visweswaran [2016]). *If $\mathbb{M}$ is a performance metric that is maximized only when $P(Y|\mathbf{X})$ is estimated accurately and $\mathbb{L}$ is a learning algorithm that can approximate any conditional probability distribution, then*

1. *$\mathbf{S}$ is a Markov blanket of Y if and only if it is an optimal predictor of Y, and*
2. *$\mathbf{S}$ is a Markov boundary of Y if and only if it is a minimal and optimal predictor of Y.*

Now we can prove the orginal proposition.

Proof. We use $\mathrm{BL}^{\mathrm{test}}$ and $\mathrm{BD}^{\mathrm{test}}$ to denote the Markov blankets and Markov boundary in the testing distribution. We first prove that $\mathrm{BL}^{\mathrm{test}}(Y) = \mathrm{BL}(Y)$ and $\mathrm{BD}^{\mathrm{test}}(Y) = \mathrm{BD}(Y)$.

Suppose $\mathbf{S}$ is a Markov blanket under the training distribution P^{tr}. Let $\mathbf{V} = \mathbf{X}\backslash\mathbf{S}$. Under Assumption 2 and Assumption 1, $\forall \mathbf{v} \in \mathcal{V}, \mathbf{s} \in \mathcal{S}, y \in \mathcal{Y}$,

$$P^{\mathrm{te}}(Y = y|\mathbf{V} = \mathbf{v}, \mathbf{S} = \mathbf{s}) = P^{\mathrm{tr}}(Y = y|\mathbf{V} = \mathbf{v}, \mathbf{S} = \mathbf{s}) = P^{\mathrm{tr}}(Y = y|\mathbf{S} = \mathbf{s}).$$

Hence,

$$\begin{aligned}
&P^{\mathrm{te}}(Y = y|\mathbf{S} = \mathbf{s}) \\
=&\int_{\mathcal{V}} P^{\mathrm{te}}(Y = y|\mathbf{V} = \mathbf{v}', \mathbf{S} = \mathbf{s}) P^{\mathrm{te}}(\mathbf{V} = \mathbf{v}'|\mathbf{S} = \mathbf{s}) \mathrm{d}\mathbf{v}' \\
=&\int_{\mathcal{V}} P^{\mathrm{tr}}(Y = y|\mathbf{S} = \mathbf{s}) P^{\mathrm{te}}(\mathbf{V} = \mathbf{v}'|\mathbf{S} = \mathbf{s}) \mathrm{d}\mathbf{v}' \\
=&P^{\mathrm{tr}}(Y = y|\mathbf{S} = \mathbf{s}) = P^{\mathrm{te}}(Y = y|\mathbf{V} = \mathbf{v}, \mathbf{S} = \mathbf{s}).
\end{aligned}$$

As a result, $\mathbf{S}$ is a Markov blanket under P^{te}, which implies $\mathrm{BL}(Y) \subseteq \mathrm{BL}^{\mathrm{test}}(Y)$. With similar calculation, we can show that $\mathrm{BL}^{\mathrm{test}}(Y) \subseteq \mathrm{BL}(Y)$, which finally shows that $\mathrm{BL}^{\mathrm{test}}(Y) = \mathrm{BL}(Y)$. Because Markov boundary is the minimal element of the set of Markov blankets, we can get that $\mathrm{BD}^{\mathrm{test}}(Y) = \mathrm{BD}(Y)$.

Now the original proposition is straightforward with Proposition A.6. □