A Theoretical Analysis on Independence-driven Importance Weighting for Covariate-shift Generalization

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

Abstract

Covariate-shift generalization, a typical case in out-of-distribution (OOD) generalization, requires a good performance on the unknown test distribution, which varies from the accessible training distribution in the form of covariate shift. Recently, independence-driven importance weighting algorithms in stable learning literature have shown empirical effectiveness to deal with covariate-shift generalization on several learning models, including regression algorithms and deep neural networks, while their theoretical analyses are missing. In this paper, we theoretically prove the effectiveness of such algorithms by explaining them as feature selection processes. We first specify a set of variables, named **minimal stable variable set**, that is the minimal and optimal set of variables to deal with covariate-shift generalization for common loss functions, such as the mean squared loss and binary cross-entropy loss. Afterward, we prove that under ideal conditions, independence-driven importance weighting algorithms could identify the variables in this set. Analysis of asymptotic properties is also provided. These theories are further validated in several synthetic experiments. The source code is available at https://github.com/windxrz/independence-driven-IW.

1. Introduction

Although modern machine learning techniques have achieved great success in various areas, many researchers have demonstrated the vulnerability of machine learning models under distribution shifts (Shen et al., 2021). This issue arises from the violation of the *i.i.d.* assumption (*i.e.*, training and test data are independent and identically distributed) and stimulates recent research on out-of-distribution (OOD) generalization (Shen et al., 2021; Zhang et al., 2022a). Among different types of distribution shifts considered in OOD literature, covariate shift (Shimodaira, 2000; Sugiyama et al., 2007a; Ben-David et al., 2007), where the marginal distribution of variables shifts from the training data to the test data while the labeling function keeps unchanged, is the most common one (Shen et al., 2021). Further, covariate-shift generalization is much more challenging, given that the test distribution remains unknown in the training phase.

With the prior knowledge of the test distribution, importance weighting (IW) is common in dealing with covariate shift (Shimodaira, 2000; Sugiyama et al., 2007a;b; 2008; Fang et al., 2020). In detail, IW methods consist of two steps, namely weight estimation and weighted regression (Fang et al., 2020). The weight estimation step estimates sample weights that characterize the density ratio between the training and test distribution. The weighted regression step trains predictors after plugging the sample weights into loss functions. However, IW methods can not adapt to covariate-shift generalization problems directly because the test distribution is unknown.

Recently, independence-driven importance weighting methods (Shen et al., 2020; Kuang et al., 2020b; Zhang et al., 2021; 2022b) in stable learning literature (Cui & Athey, 2022) have shown empirical effectiveness to deal with covariate-shift generalization on several learning tasks involving regression algorithms and deep models. Without the knowledge of the test distribution, in the weight estimation step, they propose to learn sample weights that guarantee the statistical independence between features in the weighted distribution. Although the advantages of these 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 independence-driven IW methods on covariate-shift generalization problems by explaining them as feature selection processes.

[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China [2]Computer Science & Mathematics, The Hong Kong University of Science and Technology, Hong Kong, China. Emails: xrz199721@gmail.com, xingxuanzhang@hotmail.com, shenzy17@mails.tsinghua.edu.cn, tongzhang@tongzhang-ml.org, cuip@tsinghua.edu.cn. Correspondence to: Peng Cui <cuip@tsinghua.edu.cn>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

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 $\boldsymbol{S}$ that satisfies the condition: $\mathbb{E}[Y|\boldsymbol{S}] = \mathbb{E}[Y|\boldsymbol{X}]$. Such a minimal set of variables is named the **minimal stable variable set**. Afterward, we prove that independence-driven IW algorithms could identify the minimal stable variable set. We analyze the typical algorithms (Kuang et al., 2020b; Shen et al., 2020) 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 independence-driven IW methods work for covariate-shift generalization. These theories are further validated in several synthetic experiments.

1.1. Overview of Results

We begin with a simplified presentation of our results. Consider a set of variables $(\boldsymbol{X}, Y)$ where $\boldsymbol{X}$ represents features and Y represents the outcome that we try to predict from $\boldsymbol{X}$. We consider covariate-shift generalization problems, which is the most common one among the different distribution shifts (Shen et al., 2021). In detail, covariate shift considers the scenario where the marginal distribution of $\boldsymbol{X}$ shifts from the training phase to the test phase while the labeling function keeps unchanged.

Assumption 1.1. Suppose the test distribution P^{te} differs from the training distribution P^{tr} in covariate shift only, *i.e.*,

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

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

Problem 1.1 (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 test distribution P^{te} that satisfies Assumption 1.1.

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

Theorem 1.1 (Informal version of Theorem 3.1)**.** *Let P^{te} be the unknown test distribution in the covariate-shift generalization problem defined in Problem 1.1. Then a subset of variables $\boldsymbol{S} \subseteq \boldsymbol{X}$ that can fit the target $\mathbb{E}_{P^{te}}[Y|\boldsymbol{X}]$ if and only if it satisfies $\mathbb{E}_{P^{tr}}[Y|\boldsymbol{S}] = \mathbb{E}_{P^{tr}}[Y|\boldsymbol{X}]$.*

We define the minimal set of variables that satisfies $\mathbb{E}_{P^{\text{tr}}}[Y|\boldsymbol{S}] = \mathbb{E}_{P^{\text{tr}}}[Y|\boldsymbol{X}]$ as the **minimal stable variable set** (Definition 3.4). Under mild assumptions (Assumption 2.1), the existence and uniqueness of the variable set are guaranteed (Theorem 3.2). As relationships between $\boldsymbol{X}$ vary from the training phase to the test phase, *i.e.*, $P^{\text{tr}}(\boldsymbol{X}) \neq P^{\text{te}}(\boldsymbol{X})$, it is reasonable to find the minimal set of variables to make predictions so that it can relieve the negative impact of other features in the test distribution. We will show the optimality property of the minimal stable variable set empirically in Figure 2.

Now we consider independence-driven IW algorithms (The framework of such algorithms can be found in Section 4 and Algorithm 1). Typical independence-driven IW algorithms (Shen et al., 2020; Kuang et al., 2018) learn sample weights first to make features statistically independent in the weighted distribution 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 independence-driven IW algorithms have the following properties.

Theorem 1.2 (Informal version of Theorem 5.1 and Theorem 5.2)**.** *Under ideal conditions (perfectly learned sample weights and infinite samples),*

- *if a variable X_i is not in the minimal stable variable set, then independence-driven IW algorithms could filter it out with any weighting function that satisfies the independence condition, and*
- *if a variable X_i is in the minimal stable variable set, then there exist weighting functions with which independence-driven IW algorithms could identify X_i.*

We further analyze the error of coefficients if these ideal conditions are not satisfied (Theorem 5.3) under several mild assumptions.

Theorem 1.1 and Theorem 1.2 provide a general picture of the effectiveness of independence-driven IW algorithms. To conclude, under ideal assumptions, they 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; Zhang et al., 2022c), supervised learning

models (Peters et al., 2016; Zhou et al., 2021; Liu et al., 2021a;b; Zhou et al., 2022b; Lin et al., 2022a;b), and optimization methods (Duchi et al., 2020; Duchi & Namkoong, 2021; Zhou et al., 2022a). More thorough discussions could refer to (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) and covariate shift is the most common distribution shift (Shen et al., 2021). To deal with the covariate-shift generalization problem, there are several methods recently (Shen et al., 2020; Kuang et al., 2020b; Zhang et al., 2021; Duchi & Namkoong, 2021; Krueger et al., 2021; Ruan et al., 2021). In this paper, we focus on independence-driven IW algorithms (Shen et al., 2020; Kuang et al., 2020b; Zhang et al., 2021) and provide a theoretical analysis of them.

Importance weighting (IW) and independence-driven IW algorithms Importance weighting methods are common practices to tackle distribution shifts. In traditional domain adaptation (DA) problems (Daume III & Marcu, 2006; Ben-David et al., 2007), importance weighting methods assume the prior knowledge of the test distribution and they can estimate the density ratio between the training and test distributions directly (Shimodaira, 2000; Huang et al., 2006; Storkey & Sugiyama, 2007; Sugiyama et al., 2007a;b; Bickel et al., 2007; Sugiyama et al., 2008; Kanamori et al., 2009; Fang et al., 2020). As a result, the ERM training on the weighted distribution is unbiased in the test distribution (Fang et al., 2020).

Compared to typical DA settings, covariate-shift generalization problems consider a much more challenging setting where the test distribution is unknown (Shen et al., 2021). Without the knowledge of the test distribution, independence-driven IW algorithms (Shen et al., 2018; Kuang et al., 2020a; Shen et al., 2020; Zhang et al., 2021) in stable learning literature (Cui & Athey, 2022) propose to learn sample weights that make features statistically independent in the weighted distribution. Although the effectiveness of such algorithms on covariate-shift generalization has been proved empirically, their detailed theoretical analysis is missing.

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 & Elisseeff, 2003). Feature selection approaches can be broadly divided into four categories, namely filter methods, wrapper methods, embedded methods, and others. Filter methods adopt statistical criteria to rank and select features before building classifiers with selected features (John et al., 1994; Langley et al., 1994; Guyon & Elisseeff, 2003; Law et al., 2004). Given filter methods are usually independent of the learning of the classifiers, they show superiority in operating time and applicability over other methods (Kira & 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 the 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 minimize the size of the selected feature subset while maximizing the classification performance simultaneously (Tibshirani, 1996; Rakotomamonjy, 2003; Zou & Hastie, 2005; Loh, 2011; Chen & Guestrin, 2016). Some methods attempt to combine the advantages of wrapper methods and filter methods (Cortizo & Giraldez, 2006; Liu et al., 2014; Benoît et al., 2013). However, discussions on feature selection problems under covariate-shift generalization settings are missing. In this paper, we specify the optimal and minimal set of variables to deal with covariate-shift generalization and prove that independence-driven IW algorithms could identify them.

2. Preliminaries

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

We use $\boldsymbol{S} \subseteq \boldsymbol{X}$ to indicate that $\boldsymbol{S}$ is a subset of features $\boldsymbol{X}$ and $\subsetneq$ to mean proper subset. We write $\boldsymbol{A} \perp \boldsymbol{B} \mid \boldsymbol{C}$ when two sets of variables $\boldsymbol{A}, \boldsymbol{B} \subseteq \boldsymbol{X}$ are statistically independent given another set of variables $\boldsymbol{C} \subseteq \boldsymbol{X}$. We also adopt $\boldsymbol{A} \perp \boldsymbol{B}$ when conditioning set is empty to indicate that $\boldsymbol{A}$ and $\boldsymbol{B}$ 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, $\mathbb{E}_{Q(\boldsymbol{X})}[\boldsymbol{X}] = \int_{\mathcal{X}} \boldsymbol{x} Q(\boldsymbol{X} = \boldsymbol{x}) \mathrm{d}x$ represent the expectation of $\boldsymbol{X}$ and $\mathbb{E}_{Q(\boldsymbol{X},Y)}[Y|\boldsymbol{X}] = \int_{\mathcal{Y}} Q(Y = y|\boldsymbol{X}) y \mathrm{d}y$ represent the conditional expectation of Y given $\boldsymbol{X}$ under distribution Q. Q could be chosen as the training distribution P^{tr}, test 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}. We use $\hat{\mathbb{E}}[\cdot]$ to denote the empirical expectation *w.r.t.* n samples.

Basic assumption We consider the following assumption.

Assumption 2.1 (Strictly positive density assumption)**.** $\forall x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2, \dots, x_d \in \mathcal{X}_d$, $P^{\text{tr}}(X_1 = x_1, X_2 =$

$x_2, \ldots, X_d = x_d) > 0$.

Remark 2.1. Assumption 2.1 is reasonable on the grounds that there always exists uncertainty in the data (Pearl, 2014; Strobl & Visweswaran, 2016). Therefore, we suppose the strictly positive density assumption in the whole paper for simplicity.

3. Minimal Stable Variable Set 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 3.1 (Optimal predictor (Statnikov et al., 2013))**.** Given a dataset sampled from $P^{\text{tr}}(\boldsymbol{X}, Y)$, a learning algorithm $\mathbb{L}$, and a performance metric $\mathbb{M}$ to assess learner's models, a variable set $\boldsymbol{S} \subseteq \boldsymbol{X}$ is an optimal predictor of Y if $\boldsymbol{S}$ maximizes the performance metric $\mathbb{M}$ for predicting Y using learner $\mathbb{L}$ in the dataset.

Definition 3.2 (Minimal and optimal predictor (Strobl & Visweswaran, 2016))**.** Let $\boldsymbol{S}$ be an optimal predictor of Y. If no proper subset of $\boldsymbol{S}$ satisfies the definition of the optimal predictor of Y, then $\boldsymbol{S}$ is a minimal and optimal predictor of Y.

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

Theorem 3.1. *Under Assumption 1.1 and Assumption 2.1, if $\mathbb{M}$ is a performance metric that is maximized only when $\mathbb{E}_{P^{te}}[Y|\boldsymbol{X}]$ is estimated accurately and $\mathbb{L}$ is a learning algorithm that can approximate any conditional expectation. Suppose $\boldsymbol{S} \subseteq \boldsymbol{X}$ is a subset of variables, then*

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

Remark 3.1. To deal with covariate-shift generalization, $\mathbb{M}$ should be measured on the unknown test distribution P^{te} with common loss functions. In practice, researchers often adopt the mean squared loss in regression problems and the binary cross-entropy loss in binary classification problems. It is easy to check that the global optimum for both loss functions is $\mathbb{E}_{P^{\text{te}}}[Y|\boldsymbol{X}]$ if applying the loss functions on the test distribution P^{te}.

As a result, we provide the following definitions.

Definition 3.3 (Stable variable set)**.** A stable variable set of Y under distribution P is any subset $\boldsymbol{S}$ of $\boldsymbol{X}$ for which

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

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 3.4 (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 satisfies Equation (2).

With these definitions, the conclusions of Theorem 3.1 become: (1) $\boldsymbol{S}$ is an optimal predictor of Y under P^{te} if and only if it is a stable variable set under P^{tr}, and (2) $\boldsymbol{S}$ 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 3.2. *Under Assumption 2.1, 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*

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

Theorem 3.1 and Theorem 3.2 provide 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 test distribution P^{te}, with which we can learn reliable models (John et al., 1994; Guyon & Elisseeff, 2003). As relationships between $\boldsymbol{X}$ are usually unstable and $P^{\text{tr}}(\boldsymbol{X}) \neq P^{\text{te}}(\boldsymbol{X})$, 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 $\boldsymbol{X} \setminus \text{MinStable}(Y)$ under the test distribution.

Comparing the minimal stable variable set with other variable sets $\text{MinStable}(Y)$ could be explained as the direct causal variables in typical data-generating processes. Consider the following mechanism (Tibshirani, 1996; Ravikumar et al., 2009; Hastie & Tibshirani, 2017; Kuang et al., 2020a),

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

Here variables $\boldsymbol{X}$ contain two kinds of variables ($\boldsymbol{S}$ and $\boldsymbol{V}$) while Y depends on $\boldsymbol{S}$ only. The relationship between $\boldsymbol{S}$ and $\boldsymbol{V}$ is arbitrary. In such common cases, $\boldsymbol{S}$ is the set of all the direct causal variables and is the minimal stable variable set of Y.

In addition, the minimal stable variable set has relationships with the stable blanket proposed by Pfister et al. (2021). However, the stable blankets are defined in causal graphs

over a set of interventions while the minimal stable variable set targets for the covariate-shift generalization.

Furthermore, the minimal stable variable set is closely related to the Markov boundary (Pearl, 2014). Under the performance metric in Theorem 3.1, 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 the Markov boundary can be found in Appendix A.

4. Independence-driven IW Algorithms

4.1. General Framework

The framework of typical independence-driven importance weighting algorithms (Shen et al., 2020; Kuang et al., 2020a) is shown in Algorithm 1. Similar to standard IW algorithms (Fang et al., 2020), independence-driven IW algorithms consist of two steps, which are independence-driven weight estimation and weighted least squares respectively.

4.1.1. Independence-driven Weight Estimation

Independence-driven IW algorithms consider weighting functions that depend on $\boldsymbol{X}$ only.

Definition 4.1 (Weighting function and weighted distribution)**.** Let $\mathcal{W}$ be the set of **weighting functions** that satisfies

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

Then $\forall w \in \mathcal{W}$, the corresponding **weighted distribution** $\tilde{P}_w$ can be determined by the following probability density function.

$$\tilde{P}_w(\boldsymbol{X}, Y) = w(\boldsymbol{X}) P^{\text{tr}}(\boldsymbol{X}, Y). \tag{6}$$

$\tilde{P}_w$ is well defined with the same support of P^{tr}.

Furthermore, instead of the whole set $\mathcal{W}$, independence-driven IW algorithms consider a subset $\mathcal{W}_\perp \subseteq \mathcal{W}$. The weighting functions in $\mathcal{W}_\perp$ satisfies that $\boldsymbol{X}$ are mutually independent of each other in the corresponding weighted distribution $\tilde{P}_w$, *i.e.*,

$$\mathcal{W}_\perp \triangleq \left\{ w \in \mathcal{W} \mid \boldsymbol{X} \text{ are statistically independent in } \tilde{P}_w \right\}. \tag{7}$$

4.1.2. Weighted Least Squares

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

$$\hat{\boldsymbol{\beta}}_w = \arg\min_{\boldsymbol{\beta}} \hat{\mathbb{E}} \left[w(\boldsymbol{X}) \left(\boldsymbol{\beta}^T \boldsymbol{X} - Y \right)^2 \right] = \hat{\Sigma}_w^{-1} \hat{\mathbb{E}}[w(\boldsymbol{X}) \boldsymbol{X} Y]. \tag{8}$$

Algorithm 1 Independence-driven IW Algorithm

1: **Input:** Dataset $\left\{ \boldsymbol{x}^{(i)} = \left(x_1^{(i)}, \ldots, x_d^{(i)} \right)^T, y^{(i)} \right\}_{i=1}^n$
2: Learn sample weight $w \in \mathcal{W}_\perp$ so that $\boldsymbol{X}$ are statistically independent in the weighted distribution $\tilde{P}_w$.
3: Solve weighted least squares with weighting function $w(\boldsymbol{X})$. The solution is $\hat{\boldsymbol{\beta}}_w$.
4: **Output:** Cofficients of weighted least squares $\hat{\boldsymbol{\beta}}_w$.

Here $\hat{\Sigma}_w \triangleq \hat{\mathbb{E}}[w(\boldsymbol{X}) \boldsymbol{X} \boldsymbol{X}^T]$ represents the empirical covariance matrix with sample weights w. Furthermore, we denote the solution to population level weighted least squares under distribution $P^{\text{tr}}(\boldsymbol{X}, Y)$ as

$$\boldsymbol{\beta}_w = \arg\min_{\boldsymbol{\beta}} \mathbb{E} \left[w(\boldsymbol{X}) \left(\boldsymbol{\beta}^T \boldsymbol{X} - Y \right)^2 \right] = \Sigma_w^{-1} \mathbb{E}[w(\boldsymbol{X}) \boldsymbol{X} Y]. \tag{9}$$

Here $\Sigma_w \triangleq \mathbb{E}[w(\boldsymbol{X}) \boldsymbol{X} \boldsymbol{X}^T]$ represents the population level covariance matrix. In addition, we use $\boldsymbol{\beta}_w(X_i)$ and $\hat{\boldsymbol{\beta}}_w(X_i)$ to denote the corresponding coefficient of $\boldsymbol{\beta}_w$ and $\hat{\boldsymbol{\beta}}_w$ on the i-th feature X_i.

4.2. Two Specific Implementations

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

DWR Kuang et al. (2020a) propose to decorrelate every two features, *i.e.*,

$$w(\boldsymbol{X}) = \arg\min_{w_0(\boldsymbol{X})} \sum_{1 \le i, j \le d, i \ne j} \left(\text{Cov}(X_i, X_j; w_0) \right)^2, \tag{10}$$

where $\text{Cov}(X_i, X_j; w_0)$ represents the covariance of features X_i and X_j in the weighted distribution $\tilde{P}_{w_0}$. The loss function in Equation (10) 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 statistical independence and showed that deep models could perform better if the representations are statistically independent instead of linearly decorrelated.

SRDO Shen et al. (2020) propose to learn $w(\boldsymbol{X})$ by estimating the density ratio of the training distribution P^{tr} and a specific weighted distribution $\tilde{P}$. The weighted distribution $\tilde{P}$ is determined by performing random resampling on each feature so that $\tilde{P}(X_1, X_2, \ldots, X_d) = \prod_{i=1}^d P^{\text{tr}}(X_i)$. As a

result, the weighting function $w(\boldsymbol{X})$ is given by

$$w(\boldsymbol{X}) = \frac{\tilde{P}(\boldsymbol{X})}{P^{\text{tr}}(\boldsymbol{X})} = \frac{P^{\text{tr}}(X_1)P^{\text{tr}}(X_2)\cdots P^{\text{tr}}(X_d)}{P^{\text{tr}}(X_1, X_2, \ldots, X_d)}. \quad (11)$$

The density ratio in Equation (11) can be tackled by class-probability estimation problems and can be learned by several methods such as the binary 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 is presented by Menon & Ong (2016). As a result, SRDO can guarantee statistical independence between variables $\boldsymbol{X}$ if the density ratio is estimated accurately.

5. Theoretical Analysis of Independence-driven IW Algorithms

In this section, we will show that independence-driven IW 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 3.4. We first show the identifiability result with perfectly learned weighting functions and infinite samples in Section 5.1. Afterward, we relax the assumptions and study the asymptotic properties in Section 5.2. These theoretical results, along with Theorem 3.1 could prove the effectiveness of independence-driven IW algorithms for the covariate-shift generalization problem (Problem 1.1).

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 5.1). In addition, there exist proper weighting functions with which the coefficients on the minimal stable variable set would not be zero (Theorem 5.2).

Theorem 5.1. *Under Assumption 2.1, suppose $X_i \notin \text{MinStable}(Y)$. Let w be any weighting function in $\mathcal{W}_\perp$. Suppose $\mathbb{E}_{P^{tr}(\boldsymbol{X})}\left[w(\boldsymbol{X})\|\boldsymbol{X}\|_2^2\right] < \infty$ and $\mathbb{E}_{P^{tr}(\boldsymbol{X},Y)}\left[w(\boldsymbol{X})Y^2\right] < \infty$. Then the population level solution $\boldsymbol{\beta}_w$ of weighted least squares under w satisfies $\boldsymbol{\beta}_w(X_i) = 0$.*

Theorem 5.2. *Under Assumption 2.1, 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 $\boldsymbol{\beta}_w(X_i) = \alpha$.*

Remark 5.1. In very rare cases, independence-driven IW 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 $\tilde{P}_w$.

These two theorems, along with Theorem 3.1, prove the effectiveness of independence-driven IW algorithms for the covariate-shift generalization problem (Problem 1.1). In detail, under ideal conditions, *i.e.*, perfectly learned sample weights and infinite samples, independence-driven IW algorithms could find the minimal stable variable set of Y, which is the minimal and optimal predictor under the test distribution P^{te} according to Theorem 3.1.

5.2. Asymptotic Properties

We further analyze the asymptotic properties of independence-driven IW algorithms in this subsection. Given a weighting function $w \in \mathcal{W}$, let

$$\begin{cases} \text{approx}_w(\boldsymbol{X}) \triangleq \mathbb{E}[Y|\boldsymbol{X}] - \langle \boldsymbol{\beta}_w, \boldsymbol{X} \rangle, \\ \text{noise}(\boldsymbol{X}) \triangleq Y - \mathbb{E}[Y|\boldsymbol{X}]. \end{cases} \quad (12)$$

Here $\text{approx}_w(\boldsymbol{X})$ denotes the model misspecification term *w.r.t.* linear models and $\text{noise}(\boldsymbol{X})$ represents the noise term of Y. For a non-trivial asymptotic property of the independence-driven IW algorithms, similar to Zhang (2005); Hsu et al. (2014), we first make the assumptions about the data-generating process between $\boldsymbol{X}$ and Y.

Assumption 5.1 (Bounded covariate)**.** There exists a finite constant $B > 0$ such that, in the training distribution P^{tr}, almost surely, $\|\boldsymbol{X}\|_2 \leq B$.

Assumption 5.2 (Bounded approximation error)**.** There exists a finite constant $C_w > 0$ such that, in the training distribution P^{tr}, almost surely, $|\text{approx}_w(\boldsymbol{X})| \leq C_w$.

Assumption 5.3 (Sub-gaussian noise)**.** There exists a finite constant $\sigma \geq 0$ such that, in the training distribution P^{tr}, almost surely, $\forall \eta \in \mathbb{R}$, $\mathbb{E}\left[\exp\left(\eta \cdot \text{noise}(\boldsymbol{X})\right) | \boldsymbol{X}\right] \leq \exp\left(\eta^2\sigma^2/2\right)$.

Furthermore, we assume that the chosen weighting function is non-degenerate.

Assumption 5.4 (Non-degenerate weighting function)**.** The minimal eigenvalue of Σ_w is greater than 0, *i.e.*, $\lambda_{\min}(\Sigma_w) \triangleq \Lambda_w > 0$.

In practice, we can not obtain the true weighting function w and we need to estimate it from finite samples. The estimated weighting function is denoted as $\hat{w}$. We further provide assumptions about it.

Assumption 5.5 (Small estimation error of the weighting function)**.** The estimation error of the estimated weighting function $\hat{w}$ is small. In detail, $\mathbb{E}\left[\left(w(\boldsymbol{X}) - \hat{w}(\boldsymbol{X})\right)^2\right] \triangleq \epsilon^2 < \Lambda_w^2/\mathbb{E}[\|\boldsymbol{X}\|_2^4]$.

Assumption 5.6 (Bounded estimated weighting function)**.** There exists a finite constant $\delta_{\hat{w}} > 0$ such that, in the training distribution P^{tr}, almost surely, $\hat{w}(\boldsymbol{X}) < \delta_{\hat{w}}$.

Remark 5.2. To ensure a small ϵ^2 in Assumption 5.5, we can adopt LSIF (Kanamori et al., 2009) to optimize $\mathbb{E}_{P^{\text{tr}}(\boldsymbol{X})}[(w(\boldsymbol{X})-\hat{w}(\boldsymbol{X}))^2]$ directly. If we know a weighted distribution Q and want to learn a weighting function $w(\boldsymbol{X}) = Q(\boldsymbol{X})/P^{\text{tr}}(\boldsymbol{X})$. According to Menon & Ong (2016), the loss of LSIF is $L(w) = \mathbb{E}_{Q(\boldsymbol{X})}[-w(\boldsymbol{X})] + \mathbb{E}_{P^{\text{tr}}(\boldsymbol{X})}\left[w(\boldsymbol{X})^2/2\right]$. It is easy to see that $w^*(\boldsymbol{X}) = \min_w L(w) = Q(\boldsymbol{X})/P^{\text{tr}}(\boldsymbol{X})$ and $L(w) - L(w^*) = \mathbb{E}_{P^{\text{tr}}(\boldsymbol{X})}\left[(w^*(\boldsymbol{X}) - w(\boldsymbol{X}))^2\right]/2$. As a result, minimizing the loss of LSIF will meet the assumption which requires that $\mathbb{E}_{P^{\text{tr}}(\boldsymbol{X})}[(w(\boldsymbol{X}) - \hat{w}(\boldsymbol{X}))^2] = \epsilon^2$ be small enough.

Remark 5.3. The difference between SRDO and DWR lies in Assumption 5.5 due to the way of learning sample weights. Specifically, it is harder for DWR to satisfy Assumption 5.5 because DWR focuses more on the linear correlation. As a result, its performance may drop when Y has a complex non-linear relationship with $\mathbf{X}$, which is further validated by our experiments as shown in the fourth point in Section 6.3.

With the assumptions, we can provide the asymptotic property of independence-driven IW algorithms.

Theorem 5.3. *Let $w \in \mathcal{W}$ be a weighting function. Suppose Assumptions 2.1, 5.1-5.6 (with parameters B, C_w, σ, Λ_w, ϵ, $\delta_{\hat{w}}$) hold. Pick any $t > \max\{0, 2.6 - \log d\}$, let*

$$n \geq \frac{6\delta_{\hat{w}} B^2(\log d + t)}{\Lambda_w - \epsilon\sqrt{\mathbb{E}\left[\|\boldsymbol{X}\|_2^4\right]}}. \tag{13}$$

Then with probability at least $1 - 3e^{-t}$,

$$\begin{aligned}
&\left\|\hat{\boldsymbol{\beta}}_{\hat{w}} - \boldsymbol{\beta}_w\right\|_2^2 \\
&\leq \underbrace{\frac{4\delta_{\hat{w}}\sigma^2(d + 2\sqrt{td} + 2t)}{n\left(\Lambda_w - \epsilon\sqrt{\mathbb{E}\left[\|\boldsymbol{X}\|_2^4\right]}\right)} + \frac{8\delta_{\hat{w}} B^2 C_w^2(1+\epsilon)\left(1+\sqrt{8t}\right)^2}{n\left(\Lambda_w - \epsilon\sqrt{\mathbb{E}\left[\|\boldsymbol{X}\|_2^4\right]}\right)^2}}_{\text{error caused by WLS from finite samples}} \\
&+ \underbrace{\frac{4\epsilon^2 M_w}{\left(\Lambda_w - \epsilon\sqrt{\mathbb{E}\left[\|\boldsymbol{X}\|_2^4\right]}\right)^2}}_{\text{error caused by imperfectly learned weights}} + o(1/n).
\end{aligned} \tag{14}$$

Here $M_w = \|\Sigma_w\|_2^2\|\boldsymbol{\beta}_w\|_2^2\left(\frac{\mathbb{E}\left[\|\boldsymbol{X}\|_2^4\right]}{\|\Sigma_w\|_2^2} + \frac{\mathbb{E}\left[\|\boldsymbol{X}Y\|_2^2\right]}{\|\mathbb{E}[w(\boldsymbol{X})\boldsymbol{X}Y]\|_2^2}\right)$ is a constant when w is fixed. In particular, if $w \in \mathcal{W}_\perp$, then

$$\left\|\hat{\boldsymbol{\beta}}_{\hat{w}} - \boldsymbol{\beta}_w\right\|_2^2 = \left\|\hat{\boldsymbol{\beta}}_{\hat{w}}(\boldsymbol{V})\right\|_2^2 + \left\|\hat{\boldsymbol{\beta}}_{\hat{w}}(\boldsymbol{S}) - \boldsymbol{\beta}_w(\boldsymbol{S})\right\|_2^2, \tag{15}$$

where $\hat{\boldsymbol{\beta}}_{\hat{w}}(\boldsymbol{V})$ represents the coefficients of $\hat{\boldsymbol{\beta}}_{\hat{w}}$ on $\boldsymbol{X}\setminus\text{MinStable}(Y)$ and $\hat{\boldsymbol{\beta}}_{\hat{w}}(\boldsymbol{S})$, $\boldsymbol{\beta}_{\hat{w}}(\boldsymbol{S})$ represent the coefficients of $\hat{\boldsymbol{\beta}}_{\hat{w}}$, $\boldsymbol{\beta}_w$ on $\text{MinStable}(Y)$. As a result, $\left\|\hat{\boldsymbol{\beta}}_{\hat{w}}(\boldsymbol{V})\right\|_2^2$ and $\left\|\hat{\boldsymbol{\beta}}_{\hat{w}}(\boldsymbol{S}) - \boldsymbol{\beta}_w(\boldsymbol{S})\right\|_2^2$ are also bounded by the RHS of Equation (14).

Remark 5.4. Equation (14) applies for any weighting function $w \in \mathcal{W}$ that satisfies the listed assumptions. Excluding the high-order term of $o(1/n)$, the RHS of Equation (14) consists of two parts. The first part is caused by WLS from finite samples and it vanishes when $n \to \infty$. The second part is caused by the error between the estimated function $\hat{w}$ and the true weighting function w and it also vanishes when $\epsilon \to 0$.

In particular, let $w \in \mathcal{W}_\perp$ be a weighting function adopted by independence-driven IW algorithms. According to Theorems 5.1 and 5.2, the coefficients on $\boldsymbol{X}\setminus\text{MinStable}(Y)$ (*i.e.*, $\|\hat{\boldsymbol{\beta}}_{\hat{w}}(\boldsymbol{V})\|_2^2$) and the error of coefficients on $\text{MinStable}(Y)$ (*i.e.*, $\|\hat{\boldsymbol{\beta}}_{\hat{w}}(\boldsymbol{S}) - \boldsymbol{\beta}_w(\boldsymbol{S})\|_2^2$) will be bounded by the RHS of Equation (14) and become zero when $n \to \infty$ and $\epsilon \to 0$. This property guarantees that we could eliminate $\boldsymbol{X}\setminus\text{MinStable}(Y)$ and find $\text{MinStable}(Y)$ with finite samples and imperfectly learned sample weights.

6. Synthetic Experiments

We run various experiments on synthetic data to verify the effectiveness of independence-driven IW algorithms in discovering the minimal stable variable set in covariate-shift generalization problems. We consider the following data-generating process similar to Kuang et al. (2020a).

6.1. Data-generating Process

Data Let $\boldsymbol{X} = (\boldsymbol{S}, \boldsymbol{V})$ and the dimension of $\boldsymbol{X}$ is fixed to $d = 10$. In our experiments, the dimensions of S and V are specified as $d_s = d_v = 0.5 \cdot d = 5$. Covariate $\boldsymbol{X}$ is generated by the following process.

$$\begin{aligned}
&Z_1, Z_2, \ldots, Z_{d+1} \sim N(0,1), V_1, V_2, \ldots, V_{d_v} \sim N(0,1), \\
&S_i = 0.8Z_i + 0.2Z_{i+1}, i = 1, 2, \ldots, d_s.
\end{aligned} \tag{16}$$

We further clip features $\boldsymbol{X}$ into $[-2, 2]$ by letting $\boldsymbol{X} = \max(\min(\boldsymbol{X}, 2), -2)$. The outcome Y is generated through $Y = f(\boldsymbol{S}) + \epsilon$, in which $f(\boldsymbol{S})$ may contain both linear and non-linear transformations. To test the performance with different forms of non-linear terms in $f(\boldsymbol{S})$, we generate the outcome Y from an MLP non-linear function (Y_{MLP}) and a polynomial one (Y_{poly}), respectively:

$$\begin{cases}
Y_{\text{MLP}} = f(\boldsymbol{S}) + \epsilon = \boldsymbol{\beta}^T\boldsymbol{S} + \text{MLP}([S_1, S_2, S_3]; \theta) + \epsilon, \\
Y_{\text{poly}} = f(\boldsymbol{S}) + \epsilon = \boldsymbol{\beta}^T\boldsymbol{S} + S_1S_2S_3/4 + \epsilon.
\end{cases} \tag{17}$$

Here $\boldsymbol{\beta} = \{1/3, -2/3, 1, -1/3, 2/3\}$, $\epsilon \sim N(0, 0.3^2)$, and $\text{MLP}([S_1, S_2, S_3]; \theta)$ represents the transformation of MLP with two hidden layers (sizes 3 and 3, respectively) parametrized by randomly generated $\theta \sim U(-1, 1)$.

Generating various environments We generate various environments by constructing spurious correlations between

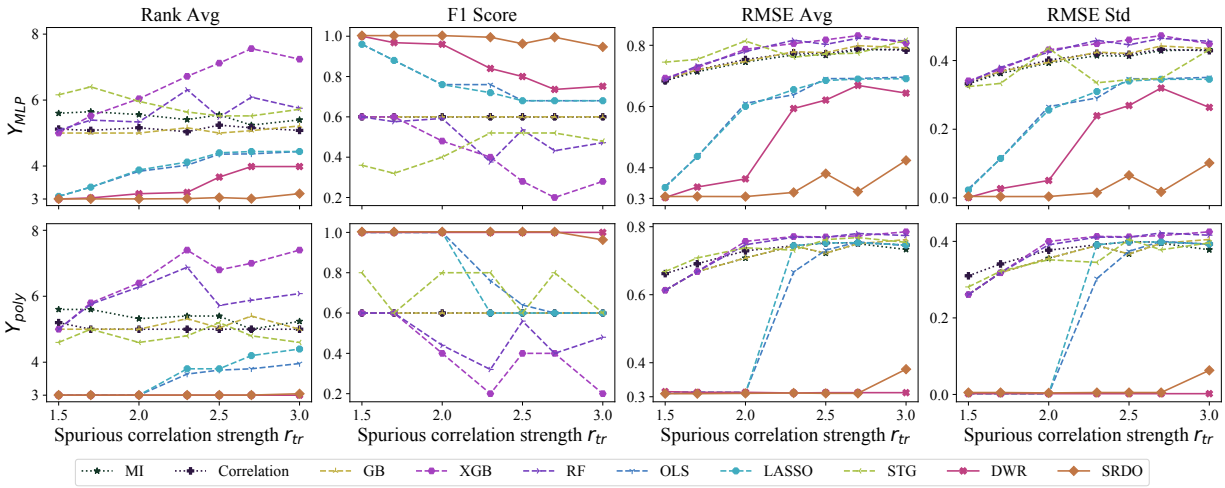


Figure 1. Experimental results on synthetic data (MLP non-linear function Y_{MLP} and polynomial function Y_{poly} from top to bottom). Varying the spurious correlation strength r_{tr}, we compare independence-driven IW algorithms (DWR and SRDO shown in solid lines) with several baselines (shown in dashed lines) on both feature selection (Rank average and F1 score) and covariate-shift generalization (RMSE average and standard deviation) metrics. Independence-driven IW algorithms outperform other methods in the synthetic experiments.

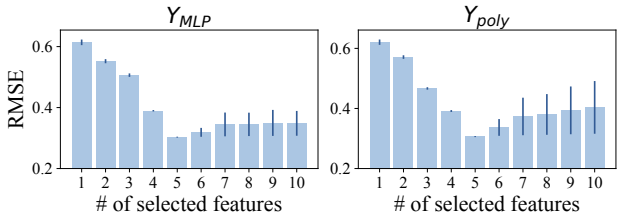


Figure 2. The covariate-shift generalization metrics (RMSE average and standard deviation) *w.r.t.* the number of selected features. Fix $r_{\text{tr}} = 2.5$ here and the feature ranking lists are provided by SRDO. The minimal stable variable set (5 features) achieves the optimal performance.

Y and V_4, V_5. Specifically, we fix a bias rate $r \in \mathbb{R}\,(|r| > 1)$ in each generated environment. For each sample $\boldsymbol{x}^{(i)} = (\boldsymbol{s}^{(i)}, \boldsymbol{v}^{(i)})$, we select it to the corresponding environment with probability $\Pr\left(\text{select}|\boldsymbol{x}^{(i)}; r\right) = \prod_{j=4}^{5} |r|^{-10D_j^{(i)}}$, where $D_j^{(i)} = \left| f(\boldsymbol{S}) - \text{sgn}(r) * \boldsymbol{v}_j^{(i)} \right|$ and $\text{sgn}(r)$ is the indicator function on whether $r > 0$. Intuitively, r controls the strength and direction of spurious correlations. Specifically, $r > 1$ corresponds to the positive spurious correlation between Y and $\boldsymbol{V}$ and $r < -1$ corresponds to the negative spurious correlation. In addition, the higher $|r|$ is, the stronger correlation between Y and $\boldsymbol{V}$ becomes.

Here $P(Y|\boldsymbol{X})$ is obviously invariant across different environments and the data-generating process satisfies the covariate-shift condition. Moreover, the minimal stable variable set $\text{MinStable}(Y)$ is $\boldsymbol{S}$ in each environment.

Experimental setting In the Y_{MLP} setting, we randomly generate 5 different MLPs and report the results averaged over the 5 MLPs. We train feature selection models on one training dataset with a specific bias rate r_{tr} and $n = 10,000$ samples. We then choose the top 5 features selected by each model and further train an MLP regressor on them. The regressor is then evaluated on 10 test environments with corresponding $r_{\text{te}} = -3.0, -2.5, -2.0, -1.5, -1.3, 1.3, 1.5, 2.0, 2.5, 3.0$. To test the effect of spurious correlation strength on feature selection models, we vary $r_{\text{tr}} = 1.5, 1.7, 2.0, 2.3, 2.5, 2.7, 3.0$.

6.2. Baselines and Evaluation Metrics

Baselines We compare independence-driven IW algorithms (including **DWR** (Kuang et al., 2020a) and **SRDO** (Shen et al., 2020)) with filter methods (including mutual information based (**MI**) and correlation based (**Correlation**) methods), wrapper methods (including gradient boosting (**GB**) (Friedman, 2001), XGBoost (**XGB**) (Chen & Guestrin, 2016), and random forests (**RF**) (Díaz-Uriarte & De Andres, 2006)), and embedded methods (including **OLS**, **LASSO** (Tibshirani, 1996), and **STG** (Yamada et al., 2020)). More details on baseline implementations can be found in Appendix B.

Evaluation metrics On the one hand, to test the performances on feature selection, we report the rank average and F1 score of selected features. To compute the rank average, we utilize the scores that each model assigns to the features and then rank all features according to the scores. The rank average is calculated as the mean of the ranks of the minimal stable variable set $\text{MinStable}(Y) = \boldsymbol{S}$. The F1 score is defined as the harmonic mean of the precision and recall, where precision and recall are computed by comparing the selected features to the true features, *i.e.*, the minimal stable variable set. On the other hand, to test the performances on covariate-shift generalization, we calculate the root mean squared error (RMSE) in each test environment and report the mean and standard deviation of RMSE in various test

environments.

6.3. Experimental Results and Analysis

The results are shown in Figure 1 and Figure 2 and we have the following observations.

1. We first validate the optimality property of the minimal stable variable set on covariate-shift generalization proposed in Section 3. With fixed $r_{\text{tr}} = 2.5$ and predicted feature ranking by SRDO, we vary the number of top selected features and train an MLP on them. Afterward, we test the performances of the MLP on test distributions and show the results in Figure 2. The results demonstrate that the minimal stable variable set (5 features) achieves the optimal performance under covariate-shift generalization. The figures with different r_{tr} and more experimental details are provided in Appendix B.
2. Independence-driven IW algorithms perform much better on the discovery of the minimal stable variable set than other feature selection methods. As shown in Figure 1, SRDO and DWR achieve the minimal rank average and maximal F1 score for both data-generating processes Y_{MLP} and Y_{poly}. As a result, with the accurate discovery of the minimal stable variable set, SRDO and DWR further achieve the best covariate-shift generalization metrics (RMSE average and standard deviation). This experiment result validates the theories in Section 5.
3. The discovery of the minimal stable variable set becomes progressively challenging as spurious correlation strength r_{tr} increases. As shown in Figure 1, the rank average tends to increase while the F1 score tends to decrease for all methods as r_{tr} increases. This phenomenon makes sense on the grounds that V_1 and V_2 become strongly correlated with Y and models tend to select them when r_{tr} is large.
4. SRDO outperforms DWR in most settings, especially when Y has a complex non-linear relationship with $\boldsymbol{X}$. As discussed in Section 4.2, DWR aims to decorrelate the linear relationships between features and can not guarantee strict statistical independence. In the Y_{MLP} setting of our experiment, DWR fails to discover the minimal stable variable set when r_{tr} is large while SRDO performs much better in the setting. However, DWR could effectively handle the polynomial Y_{poly} setting, which is also suggested by Kuang et al. (2020a).

7. Discussions

In this paper, we theoretically prove the effectiveness of independence-driven IW 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 these two conditions are not satisfied. Empirical results also demonstrate the superiority of these methods in selecting target variables.

Relationships between the minimal stable variable set and the Markov boundary The minimal stable variable set has close relationships with the Markov boundary, which we will further demonstrate in Appendix A. Here we provide a brief discussion.

Firstly, we can easily verify that the minimal stable variable set is a subset of the Markov boundary (Theorem A.2 and Example A.1) by definition (Definition 3.4 and Definition A.2). However, not all variables in the Markov boundary are necessary for the covariate-shift generalization problem with common loss functions while the minimal stable variable set could provide the minimal set of variables (comparing Theorem 3.1 and Theorem A.3).

In addition, traditional Markov boundary discovery algorithms mainly adopt the conditional independence test (Fukumizu et al., 2007; Sejdinovic et al., 2013; Strobl et al., 2019), which is a particularly challenging hypothesis to test for (Shah & Peters, 2020) though. As a result, independence-driven IW algorithms would hopefully provide a proper approximation of the Markov boundary, which could be of independent interest.

Applicable scenarios and limitations We should notice that the definition of the minimal stable variable set is applicable only when $\mathbb{E}[Y|\boldsymbol{X}]$ 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, $\mathbb{E}[Y|\boldsymbol{X}]$ will not be the solution in other forms of losses. For example, consider the Minkowski loss (Bishop, 2006, Section 1.5.5) given as $L_q = \mathbb{E}[|Y - f(\boldsymbol{X})|^q]$. It reduces to the expected squared loss when $q = 2$. The minimum of L_q is given by the conditional mean $\mathbb{E}[Y|\boldsymbol{X}]$ 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 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 independence-driven IW algorithms on multi-class classification settings as future work.

Acknowledgments

This work was supported in part by National Key R&D Program of China (No. 2018AAA0102004), National Natural Science Foundation of China (No. 62141607, U1936219), and Beijing Academy of Artificial Intelligence (BAAI).

References

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., and Koutsoukos, X. D. 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.

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., and Koutsoukos, X. D. 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.

Ben-David, S., Blitzer, J., Crammer, K., Pereira, F., et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007.

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

Benoît, F., Van Heeswijk, M., Miche, Y., Verleysen, M., and Lendasse, A. Feature selection for nonlinear models with extreme learning machines. *Neurocomputing*, 102: 111–124, 2013.

Bickel, S., Brückner, M., and Scheffer, T. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, pp. 81–88, 2007.

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

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

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

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

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

Cortizo, J. C. and Giraldez, I. Multi criteria wrapper improvements to naive bayes learning. In *International Conference on Intelligent Data Engineering and Automated Learning*, pp. 419–427. Springer, 2006.

Cui, P. and Athey, S. Stable learning establishes some common ground between causal inference and machine learning. *Nature Machine Intelligence*, 4(2):110–115, 2022.

Daume III, H. and Marcu, D. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126, 2006.

Díaz-Uriarte, R. and De Andres, S. A. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):1–13, 2006.

Duchi, J., Hashimoto, T., and Namkoong, H. Distributionally robust losses for latent covariate mixtures. *arXiv preprint arXiv:2007.13982*, 2020.

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

Fang, T., Lu, N., Niu, G., and Sugiyama, M. Rethinking importance weighting for deep learning under distribution shift. *Advances in Neural Information Processing Systems*, 33, 2020.

Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel measures of conditional dependence. In *NIPS*, volume 20, pp. 489–496, 2007.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

Garg, S., Wu, Y., Balakrishnan, S., and Lipton, Z. C. A unified view of label shift estimation. *Advances in Neural Information Processing Systems*, 33, 2020.

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

Hastie, T. J. and Tibshirani, R. J. *Generalized additive models*. Routledge, 2017.

He, Y., Cui, P., Shen, Z., Xu, R., Liu, F., and Jiang, Y. Daring: Differentiable causal discovery with residual independence. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 596–605, 2021.

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

Horn, R. A. and Johnson, C. R. *Matrix analysis*. Cambridge university press, 2012.

Hsu, D., Kakade, S., and Zhang, T. Tail inequalities for sums of random matrices that depend on the intrinsic dimension. *Electronic Communications in Probability*, 17:1–13, 2012a.

Hsu, D., Kakade, S., and Zhang, T. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17:1–6, 2012b.

Hsu, D., Kakade, S. M., and Zhang, T. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600, 2014.

Huang, B., Zhang, K., Lin, Y., Schölkopf, B., and Glymour, C. Generalized score functions for causal discovery. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1551–1560, 2018.

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19:601–608, 2006.

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

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029. PMLR, 2016.

John, G. H., Kohavi, R., and Pfleger, K. Irrelevant features and the subset selection problem. In *Machine learning proceedings 1994*, pp. 121–129. Elsevier, 1994.

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

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

Kira, K. and Rendell, L. A. A practical approach to feature selection. In *Machine learning proceedings 1992*, pp. 249–256. Elsevier, 1992.

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

Kuang, K., Cui, P., Athey, S., Xiong, R., and Li, B. Stable prediction across unknown environments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1617–1626, 2018.

Kuang, K., Xiong, R., Cui, P., Athey, S., and Li, B. Stable prediction with model misspecification and agnostic distribution shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 4485–4492, 2020a.

Kuang, K., Zhang, H., Wu, F., Zhuang, Y., and Zhang, A. Balance-subsampled stable prediction. *arXiv preprint arXiv:2006.04381*, 2020b.

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

Law, M. H., Figueiredo, M. A., and Jain, A. K. Simultaneous feature selection and clustering using mixture models. *IEEE transactions on pattern analysis and machine intelligence*, 26(9):1154–1166, 2004.

Lin, Y., Dong, H., Wang, H., and Zhang, T. Bayesian invariant risk minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16021–16030, 2022a.

Lin, Y., Zhu, S., and Cui, P. Zin: When and how to learn invariance by environment inference? *arXiv preprint arXiv:2203.05818*, 2022b.

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

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

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

Liu, J., Hu, Z., Cui, P., Li, B., and Shen, Z. Heterogeneous risk minimization. In *International Conference on Machine Learning*. PMLR, 2021a.

Liu, J., Hu, Z., Cui, P., Li, B., and Shen, Z. Kernelized heterogeneous risk minimization. *arXiv preprint arXiv:2110.12425*, 2021b.

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

Mani, S. and Cooper, G. F. Causal discovery using a bayesian local causal discovery algorithm. In *MEDINFO 2004*, pp. 731–735. IOS Press, 2004.

Menon, A. and Ong, C. S. Linking losses for density ratio and class-probability estimation. In *International Conference on Machine Learning*, pp. 304–313. PMLR, 2016.

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., and Hamprecht, F. A. 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.

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

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

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

Pfister, N., Williams, E. G., Peters, J., Aebersold, R., and Bühlmann, P. Stabilizing variable selection and regression. *The Annals of Applied Statistics*, 15(3):1220–1246, 2021.

Rahimi, A., Recht, B., et al. Random features for large-scale kernel machines. In *NIPS*, volume 3, pp. 5. Citeseer, 2007.

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

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

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

Ruan, Y., Dubois, Y., and Maddison, C. J. Optimal representations for covariate shift. *arXiv preprint arXiv:2201.00057*, 2021.

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

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, pp. 2263–2291, 2013.

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

Shen, Z., Cui, P., Kuang, K., Li, B., and Chen, P. Causally regularized learning with agnostic data selection bias. In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 411–419, 2018.

Shen, Z., Cui, P., Zhang, T., and Kuang, K. Stable learning via sample reweighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5692–5699, 2020.

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

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

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

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

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

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

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

Strobl, E. V., Zhang, K., and Visweswaran, S. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1), 2019.

Sugiyama, M., Krauledat, M., and Müller, K.-R. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007a.

Sugiyama, M., Nakajima, S., Kashima, H., Von Buenau, P., and Kawanabe, M. Direct importance estimation with model selection and its application to covariate shift

adaptation. In *NIPS*, volume 7, pp. 1433–1440. Citeseer, 2007b.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., and Kawanabe, M. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

Sugiyama, M., Kanamori, T., Suzuki, T., Hido, S., Sese, J., Takeuchi, I., and Wang, L. A density-ratio framework for statistical data processing. *IPSJ Transactions on Computer Vision and Applications*, 1:183–208, 2009.

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

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

Tsamardinos, I., Aliferis, C. F., and Statnikov, A. 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*, pp. 673–678, 2003a.

Tsamardinos, I., Aliferis, C. F., Statnikov, A. R., and Statnikov, E. Algorithms for large scale markov blanket discovery. In *FLAIRS conference*, volume 2, pp. 376–380, 2003b.

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., and Moore, J. H. Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, 85:189–203, 2018.

Yamada, Y., Lindenbaum, O., Negahban, S., and Kluger, Y. Feature selection using stochastic gates. In *International Conference on Machine Learning*, pp. 10648–10659. PMLR, 2020.

Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J. Causalvae: disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9593–9602, 2021.

Zhang, T. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17 (9):2077–2098, 2005.

Zhang, X., Cui, P., Xu, R., Zhou, L., He, Y., and Shen, Z. Deep stable learning for out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5372–5382, 2021.

Zhang, X., He, Y., Xu, R., Yu, H., Shen, Z., and Cui, P. Nico++: Towards better benchmarking for domain generalization. *arXiv preprint arXiv:2204.08040*, 2022a.

Zhang, X., Xu, Z., Xu, R., Liu, J., Cui, P., Wan, W., Sun, C., and Li, C. Towards domain generalization in object detection. *arXiv preprint arXiv:2203.14387*, 2022b.

Zhang, X., Zhou, L., Xu, R., Cui, P., Shen, Z., and Liu, H. Towards unsupervised domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4910–4920, 2022c.

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

Zheng, X., Dan, C., Aragam, B., Ravikumar, P., and Xing, E. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pp. 3414–3425. PMLR, 2020.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Loy, C. C. Domain generalization: A survey. *arXiv preprint arXiv:2103.02503*, 2021.

Zhou, X., Lin, Y., Pi, R., Zhang, W., Xu, R., Peng, C., and Zhang, T. Model agnostic sample reweighting for out-of-distribution learning. In *International Conference on Machine Learning*. PMLR, 2022a.

Zhou, X., Lin, Y., Zhang, W., and Zhang, T. Sparse invariant risk minimization. In *International Conference on Machine Learning*. PMLR, 2022b.

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

Zou, H., Kuang, K., Chen, B., Chen, P., and Cui, P. Focused context balancing for robust offline policy evaluation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 696–704, 2019.

Zou, H., Cui, P., Li, B., Shen, Z., Ma, J., Yang, H., and He, Y. Counterfactual prediction for bundle treatment. *Advances in Neural Information Processing Systems*, 33: 19705–19715, 2020.

A. Relationships between the Minimal Stable Variable Set and the Markov Boundary

A.1. Main Results

Generally speaking, the minimal stable variable set is closely related to the Markov boundary and independence-driven IW algorithms would hopefully provide a proper approximation of the Markov boundary. 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. Details are provided as follows.

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 A.1 (Markov blanket)**.** A Markov blanket of Y under distribution P is any subset $\boldsymbol{S}$ of $\boldsymbol{X}$ for which

$$Y \perp (\boldsymbol{X} \backslash \boldsymbol{S}) \mid \boldsymbol{S}. \tag{18}$$

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 A.2 (Markov boundary)**.** A Markov Boundary of Y is a minimal Markov blanket of Y, *i.e.*, none of its proper subsets satisfy Equation (18).

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

Proposition A.1. *Under Assumption 2.1, 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*

$$\mathrm{BL}(Y) = \{\boldsymbol{S} \subseteq \boldsymbol{X} \mid \mathrm{BD}(Y) \subseteq \boldsymbol{S}\}. \tag{19}$$

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 A.2. *Under Assumption 2.1, a stable variable set is also a Markov blanket and the minimal stable variable set is a subset of the Markov boundary, i.e.,*

$$\mathrm{BL}(Y) \subseteq \mathrm{Stable}(Y), \quad \mathrm{MinStable}(Y) \subseteq \mathrm{BD}(Y). \tag{20}$$

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

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

$$X_1, X_2 \sim N(0,1), \quad Y = f(X_1) + N\left(0, g(X_2)^2\right), \tag{21}$$

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

$$\begin{aligned} &\{X_1\} = \mathrm{MinStable}(Y) \subsetneq \mathrm{BD}(Y) = \{X_1, X_2\}, \\ &\{\{X_1, X_2\}\} = \mathrm{BL}(Y) \subsetneq \mathrm{Stable}(Y) = \{\{X_1\}, \{X_1, X_2\}\}. \end{aligned} \tag{22}$$

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

Theorem A.3. *Under Assumption 1.1 and Assumption 2.1, suppose $\mathbb{M}$ is a performance metric that is maximized only when $P^{te}(Y|\boldsymbol{X})$ is estimated accurately and $\mathbb{L}$ is a learning algorithm that can approximate any conditional probability distribution. Suppose $\boldsymbol{S} \subseteq \boldsymbol{X}$ is a subset of variables, then*

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

Remark A.1. The main difference between Theorem 3.1 and Theorem A.3 is the requirement on the performance metric $\mathbb{M}$. The Markov boundary is the minimal and optimal predictor if $\mathbb{M}$ is chosen as maximizing $P^{\text{te}}(Y|\boldsymbol{X})$. However, for regression problems with the mean squared loss and binary classification problems with the cross-entropy loss, $\mathbb{E}_{P^{\text{te}}}[Y|\boldsymbol{X}]$ is optimal in the test 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 & Peters (2020) have shown that conditional independence is a particularly challenging 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 A.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.

In addition, since the precise discovery of the Markov boundary is challenging, independence-driven IW algorithms would hopefully provide a proper approximation of the Markov boundary, which could be of independent interest.

A.2. Related Works on Causal Discovery and Markov Boundary

Causal literature can be categorized into two frameworks, namely the potential outcome (Rosenbaum & Rubin, 1983; Holland, 1986; Rubin, 2005; Imbens & Rubin, 2015; Johansson et al., 2016; Zou et al., 2019; 2020) and the 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 (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) methods.

Markov blankets and Markov boundary (Pearl, 2014) are the cores of local causal discovery. Under the intersection assumption (Pearl, 2014), the Markov boundary is proved unique and the discovery algorithms include (Tsamardinos & Aliferis, 2003; Tsamardinos et al., 2003a;b; Mani & 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 probabilities are strictly positive, which is a stronger assumption than the intersection assumption (Pearl, 2014) but is also common in reality (Strobl & Visweswaran, 2016). With this assumption, we can guarantee the uniqueness of the Markov boundary and the minimal stable variable set proposed in this paper.

Traditional discovery algorithms mainly use the conditional independence test (Fukumizu et al., 2007; Sejdinovic et al., 2013; Strobl et al., 2019). However, Shah & Peters (2020) proved that conditional independence is indeed a particularly difficult hypothesis to test for and there is no free lunch in conditional independence testing, which limits the application of these methods in reality. (Strobl & Visweswaran, 2016) propose a regression-based method to discover Markov boundaries, which is mostly closed to us. They proved that their method could find a subset of the Markov boundary but did not discuss the detailed properties of the subset. Here we further demonstrate that under ideal conditions, independence-driven IW algorithms could identify the exact subset of variables, *i.e.*, the minimal stable variable set.

B. More Experimental Details

Implementation details We use scikit-learn[1] for mutual information based (MI), correlation based (Correlation), gradient boosting (GB), random forest (RF), and LASSO, XGBoost package[2] for XGBoost (XGB), and original implementation[3] of STG. The hyperparameter search ranges for these baselines are shown in Table 1.

For independence-driven IW algorithms, *i.e.*, DWR, and SRDO, the feature scores are calculated as the absolute values of WLS coefficients. To be specific, for the DWR algorithm, following Kuang et al. (2020a), we learn sample weights

[1] https://scikit-learn.org
[2] https://xgboost.readthedocs.io/en/stable/
[3] https://github.com/runopti/stg

$\boldsymbol{w} = \{w_i\}_{i=1}^n$ by

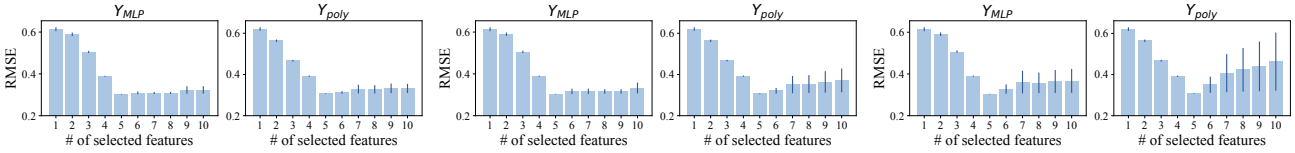
$$\boldsymbol{w} = \underset{\boldsymbol{w}\in\mathbb{R}^n}{\arg\min} \sum_{1\le i,j\le d, i\neq j} \left(\widetilde{\mathrm{Cov}}(X_i, X_j; \boldsymbol{w})\right)^2 + \lambda_1 \cdot \left(\sum_{i=1}^n w_i - 1\right)^2 + \lambda_2 \cdot \sum_{i=1}^n w_i^2. \tag{23}$$

Here $\widetilde{\mathrm{Cov}}(X_i, X_j; \boldsymbol{w})$ denotes the empirical covariance of X_i and X_j with sample weights $\boldsymbol{w}$. Equation (23) is optimized by the Adam algorithm (Kingma & Ba, 2014) with a learning rate of 0.001. For the SRDO algorithm, we implement it according to the official code[4]. In detail, an MLP classifier (two hidden layers with sizes 30 and 10, respectively) is utilized to discriminate between the training distribution $P^{\mathrm{tr}}(\boldsymbol{X})$ and the weighted distribution $P^{\mathrm{tr}}(X_1)P^{\mathrm{tr}}(X_2)\dots P^{\mathrm{tr}}(X_d)$. We adopt the binary cross-entropy loss and the Adam algorithm with a learning rate of 0.001. To further restrict the range of learned sample weights, we clip the weights to $[1/\gamma, \gamma]$. The search ranges of the hyperparameters are shown in Table 1.

We run each model 5 times on various training datasets. In each run, we train feature selection models and get the top 5 features selected by the algorithms. We then train an MLP regressor on these features to predict Y. The MLP adopted here has two hidden layers with sizes 5 and 5, respectively. The optimizer is the Adam algorithm with a learning rate of 0.001. All MLPs in our paper adopt the ReLU activation function.

Table 1. Hyperparameter search ranges of the methods.

Method	Hyperparameters
Mutual information based (MI)	$\mathrm{n\_neighbors} \in \{3, 5, 10, 20\}$
Correlation based (Correlation)	N/A
Gradient boosting (GB)	$\mathrm{n\_estimators} \in \{50, 100, 200\}$, $\mathrm{max\_depth} \in \{6, 8, 10\}$
XGBoost (XGB)	$\mathrm{n\_estimators} \in \{50, 100, 200\}$, $\mathrm{max\_depth} \in \{6, 8, 10\}$
Random forests (RF)	$\mathrm{n\_estimators} \in \{50, 100, 200\}$, $\mathrm{max\_depth} \in \{6, 8, 10\}$
OLS	N/A
LASSO	$\alpha \in \{0.0003, 0.001, 0.01, 0.1\}$
STG	$\lambda \in \{0.001, 0.01, 0.1, 1.0, 10.0\}$
DWR	$\lambda_1 \in \{0.02, 0.05, 0.1\}$, $\lambda_2 \in \{0.02, 0.05, 0.1\}$
SRDO	$\gamma \in \{5, 10, 20\}$



(a) $r_{\mathrm{tr}} = 1.5$ (b) $r_{\mathrm{tr}} = 2.0$ (c) $r_{\mathrm{tr}} = 3.0$

Figure 3. The covariate-shift generalization metrics (RMSE average and standard deviation) *w.r.t.* the number of selected features. Fix $r_{\mathrm{tr}} = 1.5, 2.0, 3.0$ and the feature ranking lists are provided by SRDO. The minimal stable variable set (5 features) achieves the optimal performance.

Table 2. Feature rankings in Figure 2 and Figure 3.

r_{train}	Y_{MLP}	Y_{poly}
1.5	$S_3, S_5, S_2, S_1, S_4, V_5, V_4, V_2, V_1, V_3$	$S_3, S_2, S_5, S_1, S_4, V_4, V_3, V_1, V_5, V_2$
2.0	$S_3, S_5, S_2, S_1, S_4, V_5, V_4, V_3, V_2, V_1$	$S_3, S_2, S_5, S_1, S_4, V_4, V_3, V_1, V_2, V_5$
2.5	$S_3, S_2, S_5, S_1, S_4, V_5, V_4, V_2, V_3, V_1$	$S_3, S_2, S_5, S_1, S_4, V_4, V_5, V_2, V_3, V_1$
3.0	$S_3, S_5, S_2, S_1, S_4, V_4, V_5, V_3, V_1, V_2$	$S_3, S_2, S_5, S_1, S_4, V_5, V_4, V_2, V_1, V_3$

[4] https://github.com/Silver-Shen/Stable_Linear_Model_Learning

The optimality property of the minimal stable variable set We adopt the SRDO method to generate feature rankings in both the Y_{MLP} (We only sample one MLP in this experiment.) and Y_{poly} settings. The feature rankings in different settings are shown in Table 2.

We vary r_{tr} to test the covariate-shift generalization metrics *w.r.t.* the number of selected features. The results are shown in Figure 3. We observe a similar phenomenon as that shown in Figure 2. These empirical results validate the advantage of the minimal stable variable set on covariate-shift generalization.

C. Omitted Proofs

C.1. Proof of Theorem 3.1

Lemma C.1. *Under Assumption 2.1, suppose $\mathbb{M}$ is a performance metric that is maximized only when $\mathbb{E}_{P^{tr}}[Y|\boldsymbol{X}]$ is estimated accurately and $\mathbb{L}$ is a learning algorithm that can approximate any conditional expectation. Suppose $\boldsymbol{S} \subseteq \boldsymbol{X}$ is a subset of variables, then*

1. *$\boldsymbol{S}$ is an optimal predictor of Y if and only if it is a stable variable set of Y under distribution P^{tr}, and*
2. *$\boldsymbol{S}$ 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 of Lemma C.1. We omit the subscript of $\mathbb{E}_{P^{\text{tr}}}[\cdot|\cdot]$ for simplicity.

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

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

Now we prove the original theorem.

Proof of Theorem 3.1. It is obvious that $\mathbb{E}_{P^{\text{tr}}}[Y|\boldsymbol{X}] = \mathbb{E}_{P^{\text{te}}}[Y|\boldsymbol{X}]$ from Assumption 1.1. As a result, the original theorem is proved according to Lemma C.1. □

C.2. Proof of Theorem 3.2

The proof is based on the following intersection property.

Lemma C.2. *Under Assumption 2.1, if $\boldsymbol{S}_1, \boldsymbol{S}_2 \in \text{Stable}(Y)$, then $\boldsymbol{S}_1 \cap \boldsymbol{S}_2 \in \text{Stable}(Y)$.*

Proof of Lemma C.2. Let $\boldsymbol{S} = \boldsymbol{S}_1 \cap \boldsymbol{S}_2$, $\bar{\boldsymbol{S}}_1 = \boldsymbol{S}_1 \backslash \boldsymbol{S}$, $\bar{\boldsymbol{S}}_2 = \boldsymbol{S}_2 \backslash \boldsymbol{S}$, and $\bar{\boldsymbol{X}} = \boldsymbol{X} \backslash (\boldsymbol{S}_1 \cup \boldsymbol{S}_2)$. Then $\boldsymbol{X} = (\boldsymbol{S}, \bar{\boldsymbol{S}}_1, \bar{\boldsymbol{S}}_2, \bar{\boldsymbol{X}})$. By definition, $\forall \boldsymbol{s} \in \mathcal{S}, \bar{\boldsymbol{s}}_1 \in \bar{\mathcal{S}}_1, \bar{\boldsymbol{s}}_2 \in \bar{\mathcal{S}}_2, \bar{\boldsymbol{x}} \in \bar{\mathcal{X}}$, $\mathbb{E}[Y|\boldsymbol{S} = \boldsymbol{s}, \bar{\boldsymbol{S}}_1 = \bar{\boldsymbol{s}}_1] = \mathbb{E}[Y|\boldsymbol{S} = \boldsymbol{s}, \bar{\boldsymbol{S}}_2 = \bar{\boldsymbol{s}}_2] = \mathbb{E}[Y|\boldsymbol{S} = \boldsymbol{s}, \bar{\boldsymbol{S}}_1 =$

$\bar{\boldsymbol{s}}_1, \bar{\boldsymbol{S}}_2 = \bar{\boldsymbol{s}}_2, \bar{\boldsymbol{X}} = \bar{\boldsymbol{x}}]$. Let $\boldsymbol{x} = (\boldsymbol{s}, \bar{\boldsymbol{s}}_1, \bar{\boldsymbol{s}}_2, \bar{\boldsymbol{x}})$. Under Assumption 2.1,

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

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

Now we prove the original theorem.

Proof of Theorem 3.2. 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 $\text{MinStable}_1(Y)$ and $\text{MinStable}_2(Y)$. By definition, $\text{MinStable}_1(Y), \text{MinStable}_2(Y) \in \text{Stable}(Y)$. Under Assumption 2.1, according to Lemma C.2, $\text{MinStable}_1(Y) \cap \text{MinStable}_2(Y) \in \text{Stable}(Y)$. Because $\text{MinStable}_1(Y)$ has no proper subset that is in $\text{Stable}(Y)$, we have $\text{MinStable}_1(Y) \cap \text{MinStable}_2(Y) = \text{MinStable}_1(Y)$. Similarly, $\text{MinStable}_1(Y) \cap \text{MinStable}_2(Y) = \text{MinStable}_2(Y)$, which means $\text{MinStable}_1(Y) = \text{MinStable}_2(Y)$.

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

$$\Omega = \{\boldsymbol{S} \subseteq \boldsymbol{X} \mid \text{MinStable}(Y) \subseteq \boldsymbol{S}\}.$$

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

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

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

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

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

C.3. Proof of Theorem 5.1

We need the following lemma first.

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

Proof of Lemma C.3. $\forall \boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}$,

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

□

Now we prove the original theorem.

Proof of Theorem 5.1. Let $\boldsymbol{X}_{-i}$ denote variables other than X_i and $\mathcal{X}_{-i}$ denote the support of $\boldsymbol{X}_{-i}$.

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

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

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

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

□

C.4. Proof of Theorem 5.2

Proof. Let $\boldsymbol{X}_{-i}$ denote the rest variable except X_i and P^{tr}_{-i} denote the marginal distribution of P^{tr} on $\boldsymbol{X}_{-i}$. Because $X_i \in \text{MinStable}(Y)$, $\mathbb{E}_{P^{\text{tr}}(\boldsymbol{X},Y)}[Y|\boldsymbol{X}]$ 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. $\boldsymbol{X}_{-i}$ are mutually independent under $\tilde{P}_{-i}$, and
2. $g(X_i) \triangleq \mathbb{E}_{\tilde{P}_{-i}(\boldsymbol{X}_{-i})}[\mathbb{E}_{P^{\text{tr}}(\boldsymbol{X},Y)}\left[Y|\boldsymbol{X}_{-i},X_i\right]]$ 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 $(\boldsymbol{X},Y)$ and $\tilde{P}(\boldsymbol{X}_{-i},X_i,Y) = \tilde{P}_{-i}(\boldsymbol{X}_{-i})\tilde{P}_i(X_i)P^{\text{tr}}(Y|X)$. Hence,

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

Let $w(\boldsymbol{X}) = \tilde{P}(\boldsymbol{X})/P^{\text{tr}}(\boldsymbol{X})$. Because $\mathbb{E}_{P^{\text{tr}}(\boldsymbol{X},Y)}[Y|\boldsymbol{X}]$ 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

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

□

C.5. Proof of Theorem 5.3

We observe that

$$\begin{aligned}
\left\|\hat{\boldsymbol{\beta}}_{\hat{w}}-\boldsymbol{\beta}_w\right\|_2^2 &\leq \left(\left\|\hat{\boldsymbol{\beta}}_{\hat{w}}-\boldsymbol{\beta}_{\hat{w}}\right\|_2+\|\boldsymbol{\beta}_{\hat{w}}-\boldsymbol{\beta}_w\|_2\right)^2 \\
&\leq 2\left(\left\|\hat{\boldsymbol{\beta}}_{\hat{w}}-\boldsymbol{\beta}_{\hat{w}}\right\|_2^2+\|\boldsymbol{\beta}_{\hat{w}}-\boldsymbol{\beta}_w\|_2^2\right) \\
&= 2\left(\left\|\Sigma_{\hat{w}}^{-1/2}\Sigma_{\hat{w}}^{1/2}\left(\hat{\boldsymbol{\beta}}_{\hat{w}}-\boldsymbol{\beta}_{\hat{w}}\right)\right\|_2^2+\|\boldsymbol{\beta}_{\hat{w}}-\boldsymbol{\beta}_w\|_2^2\right) \\
&\leq 2\left(\left\|\Sigma_{\hat{w}}^{-1/2}\right\|_2^2\left\|\Sigma_{\hat{w}}^{1/2}\left(\hat{\boldsymbol{\beta}}_{\hat{w}}-\boldsymbol{\beta}_{\hat{w}}\right)\right\|_2^2+\|\boldsymbol{\beta}_{\hat{w}}-\boldsymbol{\beta}_w\|_2^2\right) \\
&= 2\left(\left\|\Sigma_{\hat{w}}^{-1}\right\|_2\left\|\hat{\boldsymbol{\beta}}_{\hat{w}}-\boldsymbol{\beta}_{\hat{w}}\right\|_{\Sigma_{\hat{w}}}^2+\|\boldsymbol{\beta}_{\hat{w}}-\boldsymbol{\beta}_w\|_2^2\right)
\end{aligned}$$

We analyze the upper bounds of the terms in the above equation and the first part of the claim follows from Proposition C.4, Proposition C.5, and Proposition C.6. Furthermore, the second part of the claim is then straightforward from Theorem 5.1 and Theorem 5.2.

C.5.1. Error Caused by WLS from Finite Samples

Proposition C.4. *Suppose Assumption 5.4 (with parameter Λ_w) and Assumption 5.5 (with parameter ϵ) hold. Then* $\left\|\Sigma_{\hat{w}}^{-1}\right\|_2 \leq 1/\left(\Lambda_w-\epsilon\sqrt{\mathbb{E}\left[\|\boldsymbol{X}\|_2^4\right]}\right)$.

Proof. Let $\Delta_w(\boldsymbol{X}) \triangleq \hat{w}(\boldsymbol{X})-w(\boldsymbol{X})$ and $\Delta_\Sigma=\mathbb{E}\left[\Delta_w(\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^T\right]$.

$$\begin{aligned}
&\|\Delta_\Sigma\|_2 \\
=&\sup_{\|c\|_2=1}\|\Delta_\Sigma\cdot c\|_2=\sup_{\|c\|_2=1}\left\|\mathbb{E}\left[\Delta_w(\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^Tc\right]\right\|_2 \\
\leq&\sup_{\|c\|_2=1}\mathbb{E}\left[\Delta_w(\boldsymbol{X})\|\boldsymbol{X}\boldsymbol{X}^Tc\|_2\right] && \text{(triangle inequality of norms)} \\
\leq&\sup_{\|c\|_2=1}\sqrt{\mathbb{E}\left[\Delta_w(\boldsymbol{X})^2\right]\mathbb{E}\left[\|\boldsymbol{X}\boldsymbol{X}^Tc\|_2^2\right]} && \text{(Cauchy–Schwarz inequality)} \\
=&\epsilon\sup_{\|c\|_2=1}\sqrt{\mathbb{E}\left[\|\boldsymbol{X}\boldsymbol{X}^Tc\|_2^2\right]} && \left(\mathbb{E}\left[\Delta_w(\boldsymbol{X})^2\right]=\epsilon\right) \\
\leq&\epsilon\sqrt{\mathbb{E}\left[\sup_{\|c\|_2=1}\|\boldsymbol{X}\boldsymbol{X}^Tc\|_2^2\right]} && \left(\sup\mathbb{E}[\cdot]\leq\mathbb{E}[\sup\cdot]\right) \\
=&\epsilon\sqrt{\mathbb{E}\left[\|\boldsymbol{X}\|_2^4\right]}
\end{aligned}$$

As a result, according to Weyl's theorem (Horn & Johnson, 2012), Assumption 5.4, and Assumption 5.5,

$$\lambda_{\min}\left(\Sigma_{\hat{w}}\right)=\lambda_{\min}\left(\Sigma_w+\Delta_\Sigma\right)\geq\Lambda_w-\|\Delta_\Sigma\|_2\geq\Lambda_w-\epsilon\sqrt{\mathbb{E}\left[\|\boldsymbol{X}\|_2^4\right]}>0.$$

Therefore,

$$\left\|\Sigma_{\hat{w}}^{-1}\right\|_2=\frac{1}{\lambda_{\min}\left(\Sigma_{\hat{w}}\right)}\leq\frac{1}{\Lambda_w-\epsilon\sqrt{\mathbb{E}\left[\|\boldsymbol{X}\|_2^4\right]}}$$

□

Proposition C.5. *Suppose Assumption 5.1 (with parameter B), Assumption 5.2 (with parameter C_w), Assumption 5.3 (with parameter σ), Assumption 5.4 (with parameter Λ_w), Assumption 5.5 (with parameter ϵ), and Assumption 5.6 (with*

parameter $\delta_{\hat{w}}$) hold. Then there exist constants $\rho_{\hat{w}}, b_{\hat{w}}, \sigma_{\hat{w}} > 0$ such that weighting function $\hat{w}$ satisfies Condition D.1 (with parameter $\rho_{\hat{w}}$), Condition D.2 (with parameter $b_{\hat{w}}$), and Condition D.3 (with parameter $\sigma_{\hat{w}}$) and

$$\begin{cases} \rho_{\hat{w}} \leq \dfrac{\sqrt{\delta_{\hat{w}}}B}{\sqrt{d\left(\Lambda_w - \epsilon\sqrt{\mathbb{E}\left[\|\boldsymbol{X}\|_2^4\right]}\right)}}, \\ b_{\hat{w}} \leq \dfrac{\delta_{\hat{w}}C_w B}{\sqrt{d\left(\Lambda_w - \epsilon\sqrt{\mathbb{E}\left[\|\boldsymbol{X}\|_2^4\right]}\right)}}, \\ \sigma_{\hat{w}} \leq \sqrt{\delta_{\hat{w}}}\sigma. \end{cases}$$

Furthermore, pick any $t > \max\{0, 2.6 - \log d\}$, let

$$n \geq \frac{6\delta_{\hat{w}}B^2(\log d + t)}{\Lambda_w - \epsilon\sqrt{\mathbb{E}\left[\|\boldsymbol{X}\|_2^4\right]}}.$$

Then with probability at least $1 - 3e^{-t}$,

$$\begin{aligned} \left\|\hat{\boldsymbol{\beta}}_{\hat{w}} - \boldsymbol{\beta}_{\hat{w}}\right\|_{\Sigma_{\hat{w}}}^2 &\leq \frac{2\delta_{\hat{w}}\sigma^2(d + 2\sqrt{td} + 2t)}{n} + \frac{4\delta_{\hat{w}}B^2\mathbb{E}\left[\hat{w}(\boldsymbol{X})\mathrm{approx}_w(\boldsymbol{X})^2\right]}{n\left(\Lambda_w - \epsilon\sqrt{\mathbb{E}\left[\|\boldsymbol{X}\|_2^4\right]}\right)}\left(1 + \sqrt{8t}\right)^2 + o(1/n) \\ &\leq \frac{2\delta_{\hat{w}}\sigma^2(d + 2\sqrt{td} + 2t)}{n} + \frac{4\delta_{\hat{w}}B^2C_w^2(1+\epsilon)}{n\left(\Lambda_w - \epsilon\sqrt{\mathbb{E}\left[\|\boldsymbol{X}\|_2^4\right]}\right)}\left(1 + \sqrt{8t}\right)^2 + o(1/n). \end{aligned}$$

Proof. Based on Assumption 5.1, Assumption 5.4, Assumption 5.5, and Assumption 5.6, according to Proposition C.4, almost surely,

$$\frac{\sqrt{\hat{w}(\boldsymbol{X})}\left\|\Sigma_{\hat{w}}^{-1/2}\boldsymbol{X}\right\|_2}{\sqrt{d}} \leq \frac{\sqrt{\delta_{\hat{w}}}\left\|\Sigma_{\hat{w}}^{-1/2}\right\|_2\|\boldsymbol{X}\|_2}{\sqrt{d}} \leq \frac{\sqrt{\delta_{\hat{w}}}B}{\sqrt{d\left(\Lambda_w - \epsilon\sqrt{\mathbb{E}\left[\|\boldsymbol{X}\|_2^4\right]}\right)}}$$

Based on Assumption 5.1, Assumption 5.2, Assumption 5.4, Assumption 5.5, and Assumption 5.6, according to Proposition C.4, almost surely,

$$\frac{\hat{w}(\boldsymbol{X})\left\|\Sigma_{\hat{w}}^{-1/2}\mathrm{approx}_w(\boldsymbol{X})\boldsymbol{X}\right\|_2}{\sqrt{d}} \leq \frac{\delta_{\hat{w}}C_w B\left\|\Sigma_{\hat{w}}^{-1/2}\right\|}{\sqrt{d}} \leq \frac{\delta_{\hat{w}}C_w B}{\sqrt{d\left(\Lambda_w - \epsilon\sqrt{\mathbb{E}\left[\|\boldsymbol{X}\|_2^4\right]}\right)}}$$

Based on Assumption 5.3 and Assumption 5.6, almost surely,

$$\forall\eta\in\mathbb{R},\quad \mathbb{E}\left[\exp\left(\eta\sqrt{\hat{w}(\boldsymbol{X})}\mathrm{noise}(\boldsymbol{X})\right)\middle|\boldsymbol{X}\right] \leq \exp\left(\eta^2\hat{w}(\boldsymbol{X})\sigma^2/2\right) \leq \exp\left(\eta^2\left(\sqrt{\delta_{\hat{w}}}\sigma\right)^2/2\right).$$

Because

$$\begin{aligned} \mathbb{E}\left[\hat{w}(\boldsymbol{X})\mathrm{approx}_w(\boldsymbol{X})^2\right] &\leq C_w^2\mathbb{E}\left[|\hat{w}(\boldsymbol{X})|\right] \leq C_w^2\left(\mathbb{E}[w(\boldsymbol{X})] + \mathbb{E}[|w(\boldsymbol{X}) - \hat{w}(\boldsymbol{X})|]\right) \\ &\leq C_w^2\left(1 + \sqrt{\mathbb{E}\left[\left(w(\boldsymbol{X}) - \hat{w}(\boldsymbol{X})\right)^2\right]}\right) = C_w^2(1+\epsilon). \end{aligned}$$

Now the claim follows from Theorem D.1. Theorem D.1 provides the asymptotic property of WLS and we analyze it in detail in Section D. □

C.5.2. Error Caused by Imperfectly Learned Weights

Proposition C.6. *Suppose Assumption 5.4 (with parameter Λ_w) and Assumption 5.5 (with parameter ϵ) hold. Then*

$$\|\boldsymbol{\beta}_{\hat{w}} - \boldsymbol{\beta}_w\|_2 \leq \frac{\epsilon \|\Sigma_w\|_2 \|\boldsymbol{\beta}_w\|_2}{\Lambda_w - \epsilon \sqrt{\mathbb{E}\left[\|\boldsymbol{X}\|_2^4\right]}} \left(\frac{\sqrt{\mathbb{E}\left[\|\boldsymbol{X}\|_2^4\right]}}{\|\Sigma_w\|_2} + \frac{\sqrt{\mathbb{E}\left[\|\boldsymbol{X}Y\|_2^2\right]}}{\|\mathbb{E}[w(\boldsymbol{X})\boldsymbol{X}Y]\|_2} \right)$$

Proof. Let $\Delta_w(\boldsymbol{X}) \triangleq \hat{w}(\boldsymbol{X}) - w(\boldsymbol{X})$ and $\Delta_b \triangleq \mathbb{E}\left[\Delta_w(\boldsymbol{X})\boldsymbol{X}Y\right]$. We can prove that

$$\begin{aligned}
\|\Delta_b\|_2 &= \|\mathbb{E}\left[\Delta w(\boldsymbol{X})\boldsymbol{X}Y\right]\|_2 \\
&\leq \mathbb{E}\left[\Delta w(\boldsymbol{X})\|\boldsymbol{X}Y\|_2\right] && \text{(triangle inequality of norms)} \\
&\leq \sqrt{\mathbb{E}[\Delta w(\boldsymbol{X})^2]\mathbb{E}\left[\|\boldsymbol{X}Y\|_2^2\right]} && \text{(Cauchy–Schwarz inequality)} \\
&= \epsilon\sqrt{\mathbb{E}\left[\|\boldsymbol{X}Y\|_2^2\right]} && \left(\mathbb{E}\left[\Delta w(\boldsymbol{X})^2\right] = \epsilon\right)
\end{aligned}$$

In addition, $\Sigma_w\boldsymbol{\beta}_w = \mathbb{E}[w(\boldsymbol{X})\boldsymbol{X}Y]$ and $(\Sigma_w + \Delta_\Sigma)\boldsymbol{\beta}_{\hat{w}} = \mathbb{E}[w(\boldsymbol{X})\boldsymbol{X}Y] + \Delta_b$. As a result, according to Lemma E.1 and Proposition C.4,

$$\begin{aligned}
\frac{\|\boldsymbol{\beta}_{\hat{w}} - \boldsymbol{\beta}_w\|_2}{\|\boldsymbol{\beta}_w\|_2} &\leq \frac{\|\Sigma_w\|_2\|\Sigma_w^{-1}\|_2}{1 - \|\Sigma_w^{-1}\|_2\|\Delta_\Sigma\|_2} \left(\frac{\|\Delta_\Sigma\|_2}{\|\Sigma_w\|_2} + \frac{\|\Delta_b\|_2}{\|\mathbb{E}[w(\boldsymbol{X})\boldsymbol{X}Y]\|_2} \right) \\
&\leq \frac{\epsilon\|\Sigma_w\|_2}{\Lambda_w - \epsilon\sqrt{\mathbb{E}\left[\|\boldsymbol{X}\|_2^4\right]}} \left(\frac{\sqrt{\mathbb{E}\left[\|\boldsymbol{X}\|_2^4\right]}}{\|\Sigma_w\|_2} + \frac{\sqrt{\mathbb{E}\left[\|\boldsymbol{X}Y\|_2^2\right]}}{\|\mathbb{E}[w(\boldsymbol{X})\boldsymbol{X}Y]\|_2} \right)
\end{aligned}$$

□

C.6. Proof of Proposition A.1

The proof is based on the following lemma.

Lemma C.7 (Intersection Property)**.** *Under Assumption 2.1, let $\boldsymbol{V}_1$, $\boldsymbol{V}_2$, and $\boldsymbol{S}$ be subset of $\boldsymbol{X}$. Then,*

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

The proof of Lemma C.7 can be found in (Pearl, 2014, Section 3.1.2). Now we prove the original theorem.

Proof of Proposition A.1. According to Statnikov et al. (2013), if the distribution P^{tr} satisfies the intersection 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 C.7, we can know that under Assumption 2.1, if $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are Markov blankets of Y, so does $\boldsymbol{S}_1 \cap \boldsymbol{S}_2$. As a result, for any $\boldsymbol{S} \in \mathrm{BL}(Y)$, we have $\boldsymbol{S} \cap \mathrm{BD}(Y) \in \mathrm{BL}(Y)$. Because $\mathrm{BD}(Y)$ is the minimal element in $\mathrm{BL}(Y)$, we have $|\boldsymbol{S} \cap \mathrm{BD}(Y)| \geq |\mathrm{BD}(Y)|$. Hence, $\mathrm{BD}(Y) \subseteq \boldsymbol{S}$.

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

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

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

C.7. Proof of Theorem A.2

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

Therefore, $\forall \boldsymbol{S} \in \mathrm{BL}(Y)$, $\boldsymbol{S} \in \mathrm{Stable}(Y)$. According to Theorem 3.2, $\mathrm{MinStable}(Y) \subseteq \boldsymbol{S}$. In particular, let $\boldsymbol{S} = \mathrm{BD}(Y) \in \mathrm{BL}(Y)$ and we have $\mathrm{MinStable}(Y) \subseteq \mathrm{BD}(Y)$. □

C.8. Proof of Theorem A.3

The proof is based on the following proposition.

Proposition C.8 (Statnikov et al. (2013); Strobl & Visweswaran (2016))**.** *Suppose $\mathbb{M}$ is a performance metric that is maximized only when $P(Y|\boldsymbol{X})$ is estimated accurately and $\mathbb{L}$ is a learning algorithm that can approximate any conditional probability distribution. Suppose $\boldsymbol{S} \subseteq \boldsymbol{X}$ is a subset of variables, then*

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

Now we can prove the original theorem.

Proof of Theorem A.3. We use $\mathrm{BL}^{\mathrm{test}}$ and $\mathrm{BD}^{\mathrm{test}}$ to denote the Markov blankets and Markov boundary in the test distribution. We first prove that $\mathrm{BL}^{\mathrm{test}}(Y) = \mathrm{BL}(Y)$ and $\mathrm{BD}^{\mathrm{test}}(Y) = \mathrm{BD}(Y)$.

Suppose $\boldsymbol{S}$ is a Markov blanket under the training distribution P^{tr}. Let $\boldsymbol{V} = \boldsymbol{X} \backslash \boldsymbol{S}$. Under Assumption 1.1 and Assumption 2.1, $\forall \boldsymbol{v} \in \mathcal{V}, \boldsymbol{s} \in \mathcal{S}, y \in \mathcal{Y}$,

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

Hence,

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

As a result, $\boldsymbol{S}$ is a Markov blanket under P^{te}, which implies $\mathrm{BL}(Y) \subseteq \mathrm{BL}^{\mathrm{test}}(Y)$. With similar calculations, 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 claim follows from Proposition C.8. □

D. Asymptotic Oroperty of WLS

D.1. Main Result

Condition D.1 (Bounded statistical leverage)**.** For a weighting function $w \in \mathcal{W}$, there exists a finite constant $\rho_w \geq 1$, such that, in the training distribution P^{tr}, almost surely,

$$\frac{\sqrt{w(\boldsymbol{X})} \left\| \Sigma_w^{-1/2} \boldsymbol{X} \right\|_2}{\sqrt{d}} \leq \rho_w.$$

Condition D.2 (Bounded approximation error)**.** For a weighting function $w \in \mathcal{W}$, there exists a finite constant $b_w \geq 0$ such that, in the training distribution P^{tr}, almost surely,

$$\frac{w(\boldsymbol{X}) \left\| \Sigma_w^{-1/2} \mathrm{approx}_w(\boldsymbol{X}) \boldsymbol{X} \right\|_2}{\sqrt{d}} \leq b_w.$$

Condition D.3 (Noise)**.** For a weighting function $w \in \mathcal{W}$, there exists a finite constant $\sigma_w \geq 0$ such that, in the training distribution P^{tr}, almost surely,

$$\forall \eta \in \mathbb{R}, \quad \mathbb{E}\left[\exp\left(\eta\sqrt{w(\boldsymbol{X})}\text{noise}(\boldsymbol{X})\right)\middle|\boldsymbol{X}\right] \leq \exp\left(\frac{\eta^2\sigma_w^2}{2}\right).$$

Theorem D.1. *For a weighting function $w \in \mathcal{W}$. Pick any $t > \max\{0, 2.6 - \log d\}$. Suppose w satisfies Condition D.1 (with parameter ρ_w), Condition D.2 (with parameter b_w), and Condition D.3 (with parameter σ_w) and that*

$$n \geq 6\rho_w^2 d(\log d + t).$$

With probability at least $1 - 3e^{-t}$,

$$\left\|\hat{\boldsymbol{\beta}}_w - \boldsymbol{\beta}_w\right\|_{\Sigma_w}^2 \leq \frac{2\sigma_w^2\left(d + 2\sqrt{td} + 2t\right)}{n} + \frac{4\rho_w^2 d \cdot \mathbb{E}[w(\boldsymbol{X})\text{approx}_w(X)^2]}{n}(1 + \sqrt{8t})^2 + o(1/n).$$

Remark D.1. The constant b_w only appears in $o(1/n)$ terms.

D.2. Proof

The main scope of the proof follows Hsu et al. (2014), which provides the asymptotic properties of OLS and ridge regression. We further adapt it to the WLS here. We use $\mathbb{E}[\cdot]$ to denote $\mathbb{E}_{P^{\text{tr}}}[\cdot]$ throughout the section.

Let

$$\bar{\boldsymbol{\beta}}_w \triangleq \hat{\Sigma}_w^{-1}\hat{\mathbb{E}}[w(\boldsymbol{X})\boldsymbol{X}\mathbb{E}[Y|\boldsymbol{X}]].$$

Then

$$\left\|\hat{\boldsymbol{\beta}}_w - \boldsymbol{\beta}_w\right\|_{\Sigma_w}^2 \leq \left(\left\|\bar{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_w\right\|_{\Sigma_w} + \left\|\bar{\boldsymbol{\beta}}_w - \boldsymbol{\beta}_w\right\|_{\Sigma_w}\right)^2 \leq 2\left(\left\|\bar{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_w\right\|_{\Sigma_w}^2 + \left\|\bar{\boldsymbol{\beta}}_w - \boldsymbol{\beta}_w\right\|_{\Sigma_w}^2\right)$$

We analyze the two terms $\left\|\bar{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_w\right\|_{\Sigma_w}^2$ and $\left\|\bar{\boldsymbol{\beta}}_w - \boldsymbol{\beta}_w\right\|_{\Sigma_w}^2$ separately and the result is a straightforward combination of Proposition D.2, Proposition D.4, and Proposition D.5. We first define the following Δ.

$$\Delta \triangleq \Sigma_w^{-1/2}(\hat{\Sigma}_w - \Sigma_w)\Sigma_w^{-1/2},$$

D.2.1. Effect of errors in $\hat{\Sigma}_w$

Proposition D.2 (Spectral norm error in $\hat{\Sigma}_w$)**.** *Suppose w satisfies Condition D.1 (with parameter ρ_w) holds. Pick $t > \max\{0, 2.6 - \log d\}$. With probability at least $1 - e^{-t}$,*

$$\|\Delta\|_2 \leq \sqrt{\frac{4\rho_w^2 d(\log d + t)}{n}} + \frac{2\rho_w^2 d(\log d + t)}{3n}.$$

Proof. First, define $\tilde{\boldsymbol{X}} \triangleq \sqrt{w(\boldsymbol{X})}\Sigma_w^{-1/2}\boldsymbol{X}$ and let

$$\boldsymbol{Z} \triangleq \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T - I = \Sigma_w^{-1/2}\left(w(\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^T - \Sigma_w\right)\Sigma_w^{-1/2}.$$

So $\Delta = \hat{\mathbb{E}}[\boldsymbol{Z}]$. Observe that $\mathbb{E}[\boldsymbol{Z}] = 0$, and

$$\|\boldsymbol{Z}\|_2 = \max\{\lambda_{\max}(\boldsymbol{Z}), \lambda_{\max}(-\boldsymbol{Z})\} \leq \max\{\|\tilde{\boldsymbol{X}}\|_2^2, 1\} \leq \rho_w^2 d.$$

Here the second inequality is based on Condition D.1. Moreover,

$$\mathbb{E}\left[\boldsymbol{Z}^2\right] = \mathbb{E}\left[\left(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\right)^2\right] - I = \mathbb{E}\left[\|\tilde{\boldsymbol{X}}\|_2^2\left(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\right)\right] - I.$$

As a result,

$$\lambda_{\max}\left(\mathbb{E}\left[\boldsymbol{Z}^2\right]\right) \leq \lambda_{\max}\left(\mathbb{E}\left[\|\tilde{\boldsymbol{X}}\|_2^2\left(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\right)\right]\right) \leq \rho_w^2 d \cdot \lambda_{\max}(I) \leq \rho_w^2 d,$$

$$\text{tr}\left(\mathbb{E}\left[\boldsymbol{Z}^2\right]\right) \leq \text{tr}\left(\mathbb{E}\left[\|\tilde{\boldsymbol{X}}\|_2^2\left(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\right)\right]\right) \leq \rho_w^2 d \cdot \text{tr}(I) = \rho_w^2 d^2.$$

The proposition now follows from Lemma E.2. □

Proposition D.3 (Relative spectral norm error in $\hat{\Sigma}_w$ (Hsu et al., 2014)). *If* $\|\Delta\|_2 < 1$, *then*

$$\left\|\Sigma_w^{1/2}\hat{\Sigma}_w^{-1}\Sigma_w^{1/2}\right\|_2 \leq \frac{1}{1-\|\Delta\|_2}.$$

Proof. Observe that,

$$\Sigma_w^{-1/2}\hat{\Sigma}_w\Sigma_w^{-1/2} = \Sigma_w^{-1/2}\left(\Sigma_w + \hat{\Sigma}_w - \Sigma_w\right)\Sigma_w^{-1/2} = I + \Delta,$$

and according to the assumption $\|\Delta\|_2 < 1$ and Weyl's theorem (Horn & Johnson, 2012), we have

$$\lambda_{\min}(I+\Delta) \geq 1 - \|\Delta\|_2 > 0.$$

Therefore,

$$\left\|\Sigma_w^{1/2}\hat{\Sigma}_w^{-1}\Sigma_w^{1/2}\right\|_2 = \lambda_{\max}\left(\left(\Sigma_w^{-1/2}\hat{\Sigma}_w\Sigma_w^{-1/2}\right)^{-1}\right) = \lambda_{\max}\left((I+\Delta)^{-1}\right) = \frac{1}{\lambda_{\min}(I+\Delta)} \leq \frac{1}{1-\|\Delta\|_2}.$$

□

D.2.2. Effect of approximation error

Proposition D.4. *Suppose* w *satisfies Condition D.1 (with parameter* ρ_w*) and Condition D.2 (with parameter* b_w*) hold. Pick any* $t > 0$. *If* $\|\Delta\|_2 < 1$, *then*

$$\|\bar{\boldsymbol{\beta}}_w - \boldsymbol{\beta}_w\|_{\Sigma_w} \leq \frac{1}{1-\|\Delta\|_2}\left\|\hat{\mathbb{E}}[w(\boldsymbol{X})\mathrm{approx}_w(\boldsymbol{X})\boldsymbol{X}]\right\|_{\Sigma_w^{-1}}.$$

Moreover, with probability at least $1 - e^{-t}$,

$$\begin{aligned}\left\|\hat{\mathbb{E}}[w(\boldsymbol{X})\mathrm{approx}_w(\boldsymbol{X})\boldsymbol{X}]\right\|_{\Sigma_w^{-1}} &\leq \sqrt{\frac{\mathbb{E}\left[\left\|\Sigma_w^{-1/2}w(\boldsymbol{X})\mathrm{approx}_w(\boldsymbol{X})\boldsymbol{X}\right\|_2^2\right]}{n}}(1+\sqrt{8t}) + \frac{4b_w t\sqrt{d}}{3n} \\ &\leq \sqrt{\frac{\rho_w^2 d\cdot\mathbb{E}[w(\boldsymbol{X})\mathrm{approx}_w(\boldsymbol{X})^2]}{n}}(1+\sqrt{8t}) + \frac{4b_w t\sqrt{d}}{3n}\end{aligned}$$

Proof. By definition,

$$\begin{aligned}\bar{\boldsymbol{\beta}}_w - \boldsymbol{\beta}_w &= \hat{\Sigma}_w^{-1}\left(\hat{\mathbb{E}}[w(\boldsymbol{X})\boldsymbol{X}\mathbb{E}[Y|\boldsymbol{X}]] - \hat{\Sigma}_w\boldsymbol{\beta}_w\right) \\ &= \Sigma_w^{-1/2}\left(\Sigma_w^{1/2}\hat{\Sigma}_w^{-1}\Sigma_w^{1/2}\right)\Sigma_w^{-1/2}\left(\hat{\mathbb{E}}[w(\boldsymbol{X})\boldsymbol{X}(\langle\boldsymbol{\beta}_w,\boldsymbol{X}\rangle + \mathrm{approx}_w(\boldsymbol{X}))] - \hat{\Sigma}_w\boldsymbol{\beta}_w\right) \\ &= \Sigma_w^{-1/2}\left(\Sigma_w^{1/2}\hat{\Sigma}_w^{-1}\Sigma_w^{1/2}\right)\Sigma_w^{-1/2}\hat{\mathbb{E}}[w(\boldsymbol{X})\mathrm{approx}_w(\boldsymbol{X})\boldsymbol{X}].\end{aligned}$$

Therefore, with the submultiplicative property of the spectral norm,

$$\begin{aligned}\|\bar{\boldsymbol{\beta}}_w - \boldsymbol{\beta}_w\|_{\Sigma_w} &\leq \left\|\Sigma_w^{1/2}\Sigma_w^{-1/2}\right\|_2\left\|\Sigma_w^{1/2}\hat{\Sigma}_w^{-1}\Sigma_w^{1/2}\right\|_2\left\|\hat{\mathbb{E}}[w(\boldsymbol{X})\mathrm{approx}_w(\boldsymbol{X})\boldsymbol{X}]\right\|_{\Sigma_w^{-1}} \\ &\leq \frac{1}{1-\|\Delta\|_2}\left\|\hat{\mathbb{E}}[w(\boldsymbol{X})\mathrm{approx}_w(\boldsymbol{X})\boldsymbol{X}]\right\|_{\Sigma_w^{-1}}.\end{aligned}$$

Here the last inequality is according to Proposition D.3.

Now prove the second part of the claim. Observe that

$$\begin{aligned}\mathbb{E}[w(\boldsymbol{X})\mathrm{approx}_w(\boldsymbol{X})\boldsymbol{X}] &= \mathbb{E}[w(\boldsymbol{X})\boldsymbol{X}(\mathbb{E}[Y|\boldsymbol{X}] - \langle\boldsymbol{\beta}_w,\boldsymbol{X}\rangle)] \\ &= \mathbb{E}\left[w(\boldsymbol{X})\boldsymbol{X}\mathbb{E}[Y|\boldsymbol{X}]\right] - \mathbb{E}[w(\boldsymbol{X})\boldsymbol{X}\langle\boldsymbol{\beta}_w,\boldsymbol{X}\rangle] \\ &= 0.\end{aligned}$$

Therefore,

$$\mathbb{E}\left[\Sigma_w^{-1/2} w(\boldsymbol{X})\mathrm{approx}_w(\boldsymbol{X})\boldsymbol{X}\right] = \Sigma_w^{-1/2}\mathbb{E}[w(\boldsymbol{X})\mathrm{approx}_w(\boldsymbol{X})\boldsymbol{X}] = 0.$$

In addition, according to Condition D.2,

$$\left\|\Sigma_w^{-1/2} w(\boldsymbol{X})\mathrm{approx}_w(\boldsymbol{X})\boldsymbol{X}\right\|_2 \leq b_w\sqrt{d}.$$

Moreover, by Condition D.1

$$\mathbb{E}\left[\left\|\Sigma_w^{-1/2} w(\boldsymbol{X})\mathrm{approx}_w(\boldsymbol{X})\boldsymbol{X}\right\|_2^2\right] = \mathbb{E}\left[w(\boldsymbol{X})\mathrm{approx}_w(\boldsymbol{X})^2\left\|\Sigma_w^{-1/2}\sqrt{w(\boldsymbol{X})}\boldsymbol{X}\right\|_2^2\right] \leq \rho_w^2 d\cdot\mathbb{E}[w(\boldsymbol{X})\mathrm{approx}_w(\boldsymbol{X})^2].$$

The claim now follows from Lemma E.3. □

D.2.3. Effect of noise

Proposition D.5. *Suppose w satisfies Condition D.3 (with parameter σ_w) holds. Pick any $t > 0$. With probability at least $1 - e^{-t}$, either $\|\Delta\|_2 \geq 1$, or*

$$\|\Delta\|_2 < 1 \quad \textit{and} \quad \left\|\bar{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_w\right\|_{\Sigma_w}^2 \leq \frac{1}{1-\|\Delta\|_2}\frac{\sigma_w^2\left(d + 2\sqrt{td} + 2t\right)}{n}.$$

Proof. Observe that

$$\left\|\bar{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_w\right\|_{\Sigma_w}^2 \leq \left\|\Sigma_w^{1/2}\hat{\Sigma}_w^{-1/2}\right\|_2^2\left\|\bar{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_w\right\|_{\hat{\Sigma}_w}^2 = \left\|\Sigma_w^{1/2}\hat{\Sigma}_w^{-1}\Sigma_w^{1/2}\right\|_2\left\|\bar{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_w\right\|_{\hat{\Sigma}_w}^2.$$

According to Proposition D.3, if $\|\Delta\|_2 < 1$, then $\left\|\Sigma_w^{1/2}\hat{\Sigma}_w^{-1}\Sigma_w^{1/2}\right\|_2 \leq 1/(1-\|\Delta\|_2)$.

Let $\boldsymbol{\xi} \triangleq (\sqrt{w(\boldsymbol{x}^{(1)})}\mathrm{noise}(\boldsymbol{x}^{(1)}), \sqrt{w(\boldsymbol{x}^{(2)})}\mathrm{noise}(\boldsymbol{x}^{(2)}), \ldots, \sqrt{w(\boldsymbol{x}^{(n)})}\mathrm{noise}(\boldsymbol{x}^{(n)}))$ be the random vector and $\mathrm{noise}(\boldsymbol{x}^{(i)}) = \left(y^{(i)} - \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}^{(i)}]\right)$. By the definition of $\hat{\boldsymbol{\beta}}_w$ and $\bar{\boldsymbol{\beta}}_w$,

$$\left\|\bar{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_w\right\|_{\hat{\Sigma}_w}^2 = \left\|\hat{\Sigma}_w^{-1/2}\hat{\mathbb{E}}\left[w(\boldsymbol{X})\boldsymbol{X}(\mathbb{E}[Y|\boldsymbol{X}] - Y)\right]\right\|_2^2 = \boldsymbol{\xi}^T\hat{K}\boldsymbol{\xi},$$

where $\hat{K} \in \mathbb{R}^{n\times n}$ is a symmetric matrix whose (i,j)-th entry is

$$\hat{K}_{i,j} \triangleq 1/n^2\left\langle\hat{\Sigma}_w^{-1/2}\sqrt{w(\boldsymbol{x}^{(i)})}\boldsymbol{x}^{(i)}, \hat{\Sigma}_w^{-1/2}\sqrt{w(\boldsymbol{x}^{(j)})}\boldsymbol{x}^{(j)}\right\rangle.$$

According to the proof of Lemma 6 (Hsu et al., 2014), the nonzero eigenvalues of $\hat{K}$ are the same as those of

$$\frac{1}{n}\hat{\mathbb{E}}\left[\left(\hat{\Sigma}_w^{-1/2}\sqrt{w(\boldsymbol{X})}\boldsymbol{X}\right)\left(\hat{\Sigma}_w^{-1/2}\sqrt{w(\boldsymbol{X})}\boldsymbol{X}\right)^T\right] = \frac{1}{n}\hat{\Sigma}_w^{-1/2}\hat{\Sigma}_w\hat{\Sigma}_w^{-1/2} = \frac{1}{n}I_d,$$

where I_d is the identity matrix with dimension d. By Lemma E.4, with probability at least $1 - e^{-t}$ (conditioned on $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)}$),

$$\boldsymbol{\xi}^T\hat{K}\boldsymbol{\xi} \leq \sigma_w^2\left(\mathrm{tr}(\hat{K}) + 2\sqrt{\mathrm{tr}(\hat{K}^2)t} + 2\left\|\hat{K}\right\|_2 t\right) \leq \frac{\sigma_w^2\left(d + 2\sqrt{td} + 2t\right)}{n}.$$

Now the claim follows. □

E. Important lemmas

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

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

Lemma E.2 (Matrix Bernstein bound (Hsu et al., 2012a)). *Let A be a random matrix, and $r > 0$, $v > 0$, and $k > 0$ be such that, almost surely,*

$$\mathbb{E}[A] = 0, \lambda_{max}(A) \leq r, \lambda_{max}\left(\mathbb{E}\left[A^2\right]\right) \leq v, tr\left(\mathbb{E}[A^2]\right) \leq vk.$$

If $A_1, A_2, \ldots, A_n$ are independent copies of A, then for any $t > 0$,

$$\Pr\left[\lambda_{max}\left(\frac{1}{n}\sum_{i=1}^{n} A_i\right) > \sqrt{\frac{2vt}{n}} + \frac{rt}{3n}\right] \leq kt\left(e^t - t - 1\right)^{-1}.$$

If $t > 2.6$, then $t\left(e^t - t - 1\right)^{-1} \leq e^{-t/2}$.

Lemma E.3 (Vector Bernstein bound (Hsu et al., 2012b)). *Let $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)}$ be independent random vectors such that*

$$\sum_{i=1}^{n} \mathbb{E}\left[\left\|\boldsymbol{x}^{(i)}\right\|_2^2\right] \leq v \quad \text{and} \quad \left\|\boldsymbol{x}^{(i)}\right\|_2 \leq r$$

for all $i = 1, 2, \ldots, n$, almost surely. Let $\boldsymbol{s} \triangleq \boldsymbol{x}^{(1)} + \boldsymbol{x}^{(2)} + \cdots + \boldsymbol{x}^{(n)}$. For all $t > 0$,

$$\Pr\left[\|\boldsymbol{s}\|_2 > \sqrt{v}(1 + \sqrt{8t}) + (4/3)rt\right] \leq e^{-t}.$$

Lemma E.4 (Quadratic forms of a sub-Gaussian random vector (Hsu et al., 2012b)). *Let $\boldsymbol{\xi}$ be a random vector taking values from $\mathbb{R}^n$ such that for some $c \geq 0$,*

$$\mathbb{E}\left[\exp(\langle \boldsymbol{u}, \boldsymbol{\xi} \rangle)\right] \leq \exp\left(c\|\boldsymbol{u}\|_2^2/2\right), \quad \forall \boldsymbol{u} \in \mathbb{R}^n.$$

For all symmetric positive semidefinite matrices $K \succeq 0$, and all $t > 0$,

$$\Pr\left[\boldsymbol{\xi}^T K \boldsymbol{\xi} > c\left(tr(K) + 2\sqrt{tr(K^2)t} + 2\|K\|_2 t\right)\right] \leq e^{-t}.$$